Multivariate statistical modelling of the pharmaceutical process of wet granulation and tableting
Westerhuis, Johannes Arnold

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
1997

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 13-12-2018
In multivariate data analysis it is essential that training set systems are selected to provide data with information essential for the construction of calibration models; data should encompass any future situation where the multivariate model is used for predictions. Chapter 2 describes the selection of model compounds that will be used for the multivariate calibration of the process of tablet manufacturing with wet granulation. From a set of potential model compounds physical properties were obtained that, from previous experience, were thought to be important for the wet granulation process. Principal component analysis is used to reduce the dimension of the drug space, to enable a good selection. Eight compounds that are representative for the whole set were selected. In the following chapters these compounds will be used to study the whole process of tablet manufacturing with wet granulation.

Introduction

If a textbook on the art of baking would claim that all types of bread, e.g. loaves of black bread, sponge-cakes and ginger biscuits, should be baked at the same oven temperature and for an equally long time, that book would probably not be a best-seller. Each sort of bread has its own optimal process settings, and they must be compared at their own optimal response values [1].

The process of baking is not very different from the process of granulation. When formulations with new drugs or excipients have to be granulated, it is not likely that the amounts of binder and granulation liquid have to be equal to previous granulations. A comparison of the granulation of various formulations using the same processing conditions would not be fair. Just as the black bread, the cakes and biscuits all have their own optimal oven temperature and baking time, granulations all have different optimal settings of the process variables.

In a process such as baking or granulation we can distinguish two different multivariate spaces, the experimental space and the response space. The experimental space is spanned by the variables that are varied during the experiments. Each point
in the experimental space represents an experiment performed at specific settings of the process variables. In the granulation process the experimental space is spanned with process variables such as the amount of granulation liquid, the granulation time, impeller speed etc. The composition of the formulation (amount of drug, disintegrant and filler binders) is also part of the experimental space. The extremes of this space are set by the limitations of the apparatus, the formulation constraint (amounts of all ingredients sum to 100%) and by what the technician thinks is reasonable.

The response space is spanned by response variables that characterize the product of the experiments. The response variables of wet granulation comprise particle size distribution, flowability and porosity of the granules. Each product can be viewed as a point in the response space. Each point in the experimental space is connected to a point in the response space. The response space may not be continuous; not all combinations of responses are possible. Optimising a process is searching for optimal areas in the response space, and finding areas in the experimental space that are related to these optimal areas.

When a new drug or excipient is used in the formulation, the new experiment does not fit in the original experimental space anymore, because there is no single variable that distinguishes between several drugs. A new experimental space for each new formulation would be needed. This new space is connected to the response space in a different way than the former experimental space. The optimal area in the experimental space might not be the same for the various experimental spaces. If predictions have to be made about the optimal region in the experimental space of a new drug, there must be a relation between the properties of the drug and the optimal region in its experimental space. All drugs can be represented as a point in the drug space. The drug space is spanned by variables that make distinctions between the drugs and are thought to be of interest for the studied process.

In the analysis of multivariate data it is essential that the training set consists of drugs, selected to provide data with essential information for the construction of calibration models, and to represent any future situation where the multivariate model is used for predictions [2]. The model drugs must have a sufficient range of variation in all physical properties considered important in the granulation process. When the model drugs are rather similar, no predictions can be made from the model for a new drug that differs too much from the model drugs. On the other hand, when the drugs are too widely different, a good model may be very difficult to construct. In some papers it seems that the authors have used the "what could be found on the shelf" strategy to select their model-objects. This may lead to highly biased information due to the narrow span of important properties of the objects [1].

Before a selection can be made, all potential model drugs have to be described by descriptors that may be of importance for the studied process. Descriptors are variables that describe the properties of the drugs as good as possible in relation to the specific problem. These descriptors can be physical properties (melting point, particle size), chemical properties (pKa, logP) or spectroscopic data (UV, IR). The descriptors span a specific space of drug properties. The dimensionality of this drug space is equal to the
Multivariate design considerations

number of descriptors used. All potential model drugs can be described as a point in that space. So I drugs give a swarm of I points in the same space. If the drugs are similar, the variation will be very small. Correlation can exist between two or several descriptors. The real dimension of the drug space is lower than the number of descriptors. Not every axis in the drug space describes an independent property of the drugs because this may also have been described by another descriptor. If the selection of model drugs is based only on the descriptors, it may be biased. Model drugs that are rather similar may be chosen because correlated descriptors were expected to describe new variation between the drugs. Principal component analysis (PCA) is used to prevent this biased selection. PCA selects orthogonal directions in the drug space.

Principal component analysis
The goals of PCA can be: simplification of the data, data reduction, outlier selection and classification [3]. The central idea of PCA is to reduce the dimensionality of a data set which consists of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. The orthogonal principal components (PC’s) are linear combinations of the original descriptors. Because of the orthogonality each direction in the new space describes a new source of variation between the drugs.

Assume a set of I objects has been characterized by J descriptors ($X_1,..X_J$). The data can be presented in an I x J data matrix. Each object can be described as a point in the J dimensional space. The first step of PCA is to determine the direction through the cloud of points along which the data show the largest variation. This is the direction of the first principal component (PC). Say this direction is called $h_1$ then:

$$h_1 = p_{11} * X_1 + p_{12} * X_2 + ....... + p_{1J} * X_J$$

The extent to which a descriptor contributes to this first principal component is called the loading ($p_{ij}$) of a descriptor. $p_i$ is the first eigenvector of the $XX'$ matrix. The new direction can be seen as a principal property of the objects. This property cannot be measured directly but it explains most of the variation between the objects.

After projection of all data points on the first PC, it is possible to calculate how well this vector describes the data. The sum of all distances between the original point and its projection is a measure of the variation that is not described by the first PC. A second PC can be calculated to describe a part of the remaining variation. The second PC is chosen in a direction that explains the largest part of the remaining variation, orthogonal to the first PC. This can be repeated until all information in the data set is described by J PC’s. The objects can be displayed in a new J dimensional space with orthogonal axes. Using all PC’s the total variation in the data is described. Usually the first few PC’s describe a large amount of variation, the last L ones only describe minor variation. This means that one can describe by only $K=J-L$ principal components almost as much variation as with J descriptors. The remaining minor variation can be considered as noise around the important properties. It is, therefore, desirable to subtract it from the total variation in the data.
Plots
From the PC analysis two sets of data are obtained: the coordinates of the objects on the PC’s (the scores) and the amount ($p_{jk}$) that descriptors contribute to the PC’s (loadings). A scores plot presents the objects projected on two PC’s, and a loadings plot presents the weights of the descriptors on the PC’s. When descriptors lie close to each other in the loading plot of PC$_1$ against PC$_2$ they are highly correlated provided the total variation explained by the first two PC’s is large enough. Descriptors that have a strong contribution on one PC are far away from the origin and lie close to the PC axis. Descriptors that have high influence on several PC’s are projected in the quadrants of the loading plot. Descriptors that give no specific information about the objects lie close to the origin.

A two dimensional example
Two physical properties ($X_1$ and $X_2$) of 6 objects have been measured and are plotted in Figure 1. The first principal component (PC$_1$) is chosen in the direction that describes most of the variation. The objects are projected on this PC. A second component (PC$_2$) is chosen orthogonal to the first in the direction that explains the remaining variation. The two PC’s are shown in Figure 2 with the projected objects.

An important result is that we now can describe the main systematic variation (say 75%) in the data using fewer variables (only PC$_1$) than the $J$ original ($X_1$ and $X_2$) descriptors. When there are more than three descriptors it is difficult to visualize the data and, therefore, the selection of objects that give a wide spread in most of the important descriptors is not easy. By the use of principal components analysis, the data can be presented in fewer dimensions. Most of the variation can still be explained and selection of objects is much easier. The remaining variation that is not explained by the principal components can be seen as noise. Figure 2 shows the scores and loading

![Figure 1: Principal component analysis in two dimensions.](image)
A B

Figure 2: The scores of the objects (A) and the loadings of the variables (B) on the first two principal components.

plots of the data given in Figure 1.

In the scores plot (Figure 2A) we see that the variation in the PC₁ direction is the largest. In some software programs the scores plot shows equal spread in both directions, even when higher order PC's are shown. The axes have equal length, but are scaled differently. Therefore, the importance of the higher order PC's is visually exaggerated by the plot. It must always be kept in mind the amount of variance that is explained by a given PC. The loading plot shows the weights of the descriptors $X₁$ and $X₂$ on the PC's. Along the positive direction of PC₁ (in Figure 1, left under to upper right) we see that both descriptors $X₁$ and $X₂$ increase but $X₂$ increases more than $X₁$. Therefore, $X₂$ has a higher weight on the first PC (Figure 2B). The second PC shows that $X₁$ has a high positive and $X₂$ a small negative weight. The advantages of PCA become more clear when it is used for larger data sets with more descriptors where the dimension reduction is more obvious than it is for the two descriptor case.

Calculation of principal components

PCA can be computed in several ways. Here the singular value decomposition (SVD) of the $X$ matrix will be shown. The use of SVD for calculating PC’s is well described in literature [3–5]. Carlson visualizes PCA with many pictures [1]. Other methods that can be used are the eigenvector analysis [1,4,5,6] that handles the symmetric ($X'X$) matrix, and the NIPALS (nonlinear iterative partial least squares) algorithm. NIPALS is an iterative method to calculate SVD.

The $I×J$ $X$-matrix can be decomposed according to the singular value decomposition ($I≥J$):

$$X = UDV' = d₁u₁v₁' + d₂u₂v₂' + ... + d_Ju_Jv_J'$$

with $U'U=I_J$ and $V'V=VV'=I_J$, and $D$ diagonal with nonnegative diagonal elements ($d₁...d_J$).
Figure 3: The building up of $X$ with the principal components that consist of the scores $t_1...t_k$ and the loadings $p_1...p_K$. $E$ consists of the residuals.

Singular values) arranged from high to low and $X'X = V'D^2V$. The columns of $U$ and $V'$ are called the left and right-hand singular vectors of $X$ respectively. The relation of SVD with PCA becomes clear when $V'$ equals the loadings $P'$ and $UD$ equals $T$, the scores of $X$, which gives $X = TP'$.

When only $K$ columns of $T$ are used, some variation is not be described: $X = T_kP'_k + E$, where $E$ contains the unexplained variation. In Figure 3 we see that the $X$-matrix is built up from several PC's where each PC is again built from the outer product of a column-vector $t_k$ of the $T$ matrix times a row-vector $p'_k$ of the $P'$ matrix which gives an estimation of the $X$-matrix. All these estimates are mutually independent (because of the orthogonality) and can be added together to produce $X$.

Selection of model drugs from the PCA plots
The selected model drugs should have maximal spread in all their properties. Such a selection is accomplished by choosing objects projected at the borders of the score plot and as far as possible from each other. Some caution, however, must be exercised: one must be careful not to choose outliers. Besides objects at the borders it is wise also to take some random objects. The objects selected from the plot of PC1 against PC2 must also have a good spread in all the other PC score plots to make sure that enough information is caught. When some obvious classes of objects that belong together can be discerned, it is wise to select representatives from those classes [7]. Representatives of classes or typical objects lie close to the middle of the class.

Materials
A set of 42 potential model drugs were selected by considering first their price kg$^{-1}$, because large amounts of drug are used during the experiments. The drugs were also selected to have a broad range in the solubility in the granulating liquid. Some model drugs, that were suspected to give problems in the granulation experiments, were removed from the list. From the final nineteen potential model drugs, descriptors were obtained by measurement and from literature.
It is important to remember that the selected model drugs represent just a set of physical properties. The name of the drug represents a property that says only little about the drug considering its behaviour in the granulation process (e.g. paracetamol 45µm and paracetamol 180µm are two different model drugs that have the same name but a different particle size, and therefore a different compactibility, poured density etc.) For the selection of the model drugs, the following easy to obtain descriptors were selected that were thought to be important for the calibration of the granulation process.

**Solubility:** The solubility of the drugs in the granulation liquid is of main importance [8,9]. When a drug dissolves well in the granulating solution, it cannot take part in the granulation process. A relatively small amount of solution can lead to over wetting. In our case the granulation liquid will be water. The solubility of the drugs in water is given.

**Compactibility:** The compactibility of the drug is especially important for tableting of the granulation. It is a measure for several aspects that involve the binding between particles. The specific surface area is important for the mechanical strength of the tablets. This area is affected by physical properties of the drug such as particle size and shape, but also by fragmentation or plastic deformation occurring during compaction.

**Thickness tablet:** The thickness of the tablets of the pure drugs is influenced by the crystal form and poured density of the drug and the particle size, the particle size distribution and the poured density of the granulation. The thickness must be related to the weight of the tablets.

**Poured density:** The poured density gives in combination with the tapped density an idea about the flowability and the porosity of the starting material. The tapped density is the density of the bulk after 500 taps in a tap apparatus (J. Engelsman, Ludwigshafen a. Rhr).

**Hausner ratio:** The Hausner ratio is defined as the poured density divided by the tapped density. It measures of flowability of the drug. A low Hausner ratio means that the drug has a high flowability.

**Contact angle:** The contact angle is a measure for the wettability of the material. If drugs are easily wetted, granule formation is much faster. The contact angles were measured with the h-e method [10]. The influence of the wettability was already described by Jaiyeoba et al.[11].

**Particle size:** The particle size of the drugs is of main importance. It is a measure for the relative surface of the drug. It affects the speed in which the drug dissolves, the poured density, the compactibility etc. The particle size was not actually measured at first but the substances were classified to have a large or a small particle size. Most of the drugs have a small particle size. Ascorbic acid, paracetamol cryst. and
Table 1 shows the data set used. The solubility of the drugs is given as the logarithm of the parts of water needed to dissolve 1 part of drug. The logarithm is used because the original values do not have a normal distribution. High values stand for low solubility. Tablets (13 mm, 500 mg, 20 kN) were compressed using a hydraulic one punch tablet press (Mooi / ESH). The compactibility is given as the crushing strength of a tablet of the pure drug (mean of 10 measurements). The poured density is given in (g.ml⁻¹) and it is the mean of 5 experiments. The Hausner ratio is determined as the ratio of the volume of 100 g drug before and after 500 taps. The contact angle is measured using the ‘h-e’ method. The cosine of the values is given. Ascorbic acid, nicotinamide and thiamine.HCl were known to have low contact angles [12]. The contact angle of these drugs were set to the lowest available value of 48°.

The open places in the table are caused by missing values. Missing values in the tablet thickness are caused by the fact that it was not possible to produce tablets of some drugs. The table of Lagas [11] with contact angles did not contain all substances used in the experiments. For the calculation of the principal components the column means are used for the missing values.
Results and discussion

The results of the PCA were calculated using the PLS toolbox [13] for Matlab [14]. The first three PC’s explain about 74% of the variation in the data. These three PC’s were used for the selection of the model drugs. Table 2 shows the loadings of the descriptors on the first three PC’s and the percentage of variance explained by each PC.

The method used is rather rough because only few drug descriptors were used to separate between the drugs. Figures 4A-C show score plots of the first three PC’s. Each plot shows the score values of two PC’s. The numbers marked within a square (1,2,3,7,11,13,16,18) are selected as model drugs. The selection was done on sight from all three score plots. Experiments showed that model component number 3 (aluminium oxide; marked with a circle) could not be used in the Gral high-speed mixer. Therefore, another substance far away from object 3 in the multivariate space (19, dicalciumphosphate) was selected. In the first analysis, dicalciumphosphate was not taken into account. The figures show results from the analysis which included calcium phosphate. Aluminium oxide is still presented to show its position in the drug space. According to the score plots it is not outlying object, and its strange behaviour in the granulation process cannot be explained from its position in the drug space.

The first step was the selection of the outer corner points in all plots: 2, 18 and 19 in Figure 4A and 1 and 16 in Figure 4B. Object 15 in Figure 4B is close to objects 2 an 18 and was, therefore, not selected. Object 7 is selected because it is in the middle of each score plot. Objects 11 and 13 were selected to complete the eight model drugs. The selected drugs have a good spread in all PC’s. The first PC gives the spread of the drugs in particle size and poured density. The contact angle dominates the second PC and the variation in the third PC is mainly caused by the solubility of the drugs. The selected drugs have enough variation in these important variables. Three drugs with large particles and zero compactibility were selected (1, 13 and 19), one drug with high contact angle (2) and two drugs with very low solubility (16 and 19). Two drugs (7, 11) have intermediate score values on each PC. The selected substances that will be used in following experiments are: ascorbic acid, salicylic acid, isoniazid, nicotinamide, paracetamol cryst., sulfadimidine, thiamine.HCl and dicalciumphosphate.

Table 2: Loadings of the descriptors on the principal components and the percentage of variance explained by each PC.

<table>
<thead>
<tr>
<th>drug descriptors</th>
<th>PC1 (42%)</th>
<th>PC2 (18%)</th>
<th>PC3 (14%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solubility</td>
<td>0.07</td>
<td>-0.45</td>
<td>0.75</td>
</tr>
<tr>
<td>Compactibility</td>
<td>-0.30</td>
<td>0.13</td>
<td>-0.41</td>
</tr>
<tr>
<td>Thickness tablet</td>
<td>-0.41</td>
<td>-0.36</td>
<td>-0.04</td>
</tr>
<tr>
<td>Poured density</td>
<td>0.52</td>
<td>-0.18</td>
<td>-0.06</td>
</tr>
<tr>
<td>Hausner ratio</td>
<td>-0.44</td>
<td>0.35</td>
<td>0.40</td>
</tr>
<tr>
<td>Cos (θ)</td>
<td>0.24</td>
<td>0.71</td>
<td>0.31</td>
</tr>
<tr>
<td>Size</td>
<td>0.46</td>
<td>-0.01</td>
<td>-0.13</td>
</tr>
</tbody>
</table>
Figure 4: The scores of the compounds on PC1 and PC2 (A) and on PC1 and PC3 (B). The compounds are represented by the corresponding numbers. The selected model drugs are presented in a square. Compound 3, (aluminum oxide) indicated in a circle could not be used in the GRAL granulator.
Conclusion

In the multivariate calibration of the granulation process a calibration model has to be developed that predicts the settings of process variables of the granulation and physical properties of the granulation to produce tablets of new drugs. To develop the calibration model, some model drugs are needed. The selection of the model drugs for experimentation has to be based on information relevant for the specific process. Because of possible interaction effects all properties must be considered simultaneously. Multicollinearity can bias the selection of the model drugs. Principal component analysis simplifies the problem of selection, because it searches for new orthogonal directions in the drug space. The spread in the properties of the model drugs, should guarantee the validity of the calibration model for the granulation process over a large area in the drug space. In this application only few and easy to obtain variables were used to separate between the drugs. In the calibration of the tablet manufacturing process, in the following chapters, more physical properties of the selected drugs will be used to relate to the characteristics of the granulations and tablets.
Chapter 2

References

12. Bolhuis GK, Personal communication.
13. PLS Toolbox for Matlab, Eigenvector Research Inc., Manson WA, USA.
14. Matlab is a registered environment for matrix calculations, The Mathworks Inc., Natick, MA, USA.