Chapter 8

Multiblock partial least squares path modelling for multivariate processes

The process of tablet manufacturing with wet granulation has been described as a two-step process in the Chapters 6 and 7. The process variables can be divided into two groups, belonging to either step of the process. In Chapter 5 the physical granule properties were used as response variables. They were fit to the composition of the powder mixture and the process variables of the granulation step. In the Chapters 6 and 7, the granule properties were used as predictor variables to improve the modelling of the tablet properties. In the present chapter a multiblock partial least squares path model is constructed that incorporates the hierarchical structure of the process. The process variables of the wet granulation step affect both granule and tablet properties, where the process variables of the tableting step only affect the tablet properties. The same path model can be used for prediction of the tablet responses whether the granule properties are unknown (at the start of the process) or whether they have been measured (after the wet granulation step). The effect of introducing the granule properties in the model on the regression coefficients of the composition and process variables is smaller than when standard PLS or multiblock PLS models are used. The prediction of the tablet properties is comparable to the standard PLS models, but the granulation properties are predicted worse.

Introduction

Multiblock data analysis methods have their origin in path analysis and path modelling in the fields of sociology and econometrics. Path analysis was developed as a means for studying the direct and indirect effects of variables, where some variables are viewed as causes and other variables are viewed as effects. In the early days, PLS was described as a least squares path modelling technique to deal with several blocks of data [1]. Path modelling with PLS has been thoroughly described by Lohmöller [2]. The PLS model tries to incorporate the hierarchical structure of the process. In large processes a hierarchical structure exists between measurements at different parts of
the process. Intermediate materials in the middle of the process, are influenced by previous steps, and the physical properties of these intermediates also affect subsequent steps. For example, in a two-step process one can define starting materials, intermediates and the final product. At the start of the process materials are combined and several process variables are set to specific values. The intermediate, which is the product of the first step, has to meet certain specifications. Dependent on physical properties of the intermediate, one can set new process variables to specific values. The first set of process variables influences both the intermediates and the final product, whereas the second set of process variables only influences the final product.

Quality properties of the starting materials, intermediates and final products can be measured. The data sets can be placed in the hierarchical order according to the process. Between the data sets of properties the settings of the process variables, that work at a specific part of the process, can be varied to influence the product. In a two-step process, such as tablet manufacturing by wet granulation, process variables can affect either step. In the first step the starting materials are combined in the granulation bowl. The granulation time or amount of water can be varied to influence the wet granulation process. From a single tablet mixture several granulations are produced. In the second step, the different granulations can be tableted with varying compression force or compression speed. From each granulation several tablets are produced.

In monitoring the two-step wet granulation and tableting process, predictions are to be made for the granule and tablet properties at the start of the process. The influence of the composition and process variables of both steps on the properties has to be shown. When the granule properties have been obtained, they will be added to the model in order to improve the prediction of the tablet properties. In this chapter the multiblock PLS model will be extended to a path model to monitor the two-step process. The basic algorithm as presented by Wangen and Kowalski will be used to build the path model [3]. The performance and properties of the model will be investigated and compared to the standard PLS and multiblock PLS models presented in the Chapters 6 and 7.

The path model

The causal pathway of the model is assumed from left to right, where the left end blocks only predict (predictors) and the right end blocks are only predicted (predictees). Blocks in the middle of the process, interior blocks, are both predictors and predictees.

The whole process of tablet manufacturing with wet granulation is a hierarchical batch process that exists of two steps. Three blocks of physical properties can be distinguished. Figure 1 in Chapter 6 shows the three blocks in the granulation process. The left end block consist of the composition of the tablet mixture. The right end block contains the physical tablet properties and the granule properties are placed in the intermediate block. The two sets of process variables are placed at the stage where
they influence the process.

The main goal of the process is to produce tablets with specified physical properties. The granules must have specific properties necessary to perform the tableting step. The flowability and percentage of fines of the granulations have to meet certain specifications. The granule properties $G$ are, therefore, response values of the granulation step. On the other hand, the tablet properties $Z$ depend on the physical granule properties. The intermediate block has to be used as a predictee and as a predictor block. Two obvious relations exists, the first between the tablet mixture $D$ and the granule properties, and the second between the granule and tablet properties. The third relation between the tablet mixture and the tablet properties is also important. The granule properties do not contain all the information from the first block that is necessary for the modelling of the tablet properties. The process variables of the first step influence both the granule and tablet properties and can be combined with the composition of the powder mixture. The process variables of the second step only influence the tablet properties. Therefore, they are placed in a separate block that only influences the tablet block $Z$. The data sets are placed in the hierarchical way with the causal relation from left to right. The connections are shown with straight arrows. Figure 1 shows the four blocks and their relations in this model.

In 1988, Wangen and Kowalski presented a base algorithm from which an algorithm for every model with any number of blocks and relations could be made [3]. The algorithm presented here is mainly based on theirs. The MBPLS path model for the granulation of mannitol microcrystalline cellulose granulations consists of the four data blocks. The step one process variables are placed within the first block of starting materials. They could also have been placed in a separated block parallel with the first block. The step two process variables have to be placed parallel to the granule

![Diagram of MBPLS path model](image)

**Figure 1**: The placement of the various blocks in the MBPLS path model for the wet granulation and tableting process. Explanation of the blocks is given in the text. $D$ consists of the composition of the tablet mixture and the process variables for step 1. $G$ contains the granule properties, the process variables for the second step are in $P$ and the tablet properties in $Z$. 
Chapter 8

properties block. The step two process variables block is also a left end block. It is only used to make predictions of the tablet properties.

The D block consists of the composition of the tablet mixture and the settings of two process variables that influence the first granulation step. The mixture of microcrystalline cellulose (MCC), mannitol and HPC is presented by two variables, the amount of HPC in the blend (2, 3 or 5%), and the percentage of MCC from the remaining (100–HPC)%. With only two variables the three components can be presented. The two process variables of the first step are the granulation time (3, 5 or 7 minutes) and the amount of water added during granulation. The amount of water is a process variable and not a composition variable because the water is removed from the granulations in a drying step. The quadratic terms of the variables are also included because they were found to influence the tablet responses (e.g. MCC²).

The second left end block P consists of the process variables that affect only the second step of the process, the compression of granules into tablets. These process variables are the amount of moisture in the granulations (3, 4 or 5%), and the compression force (10, 20 or 30 kN). The moisture content of the granulation is in fact a response variable, but it was set to a specific value by additional drying or moistening. For block P, the quadratic effects of the process variables were also included, so for each process variables also the quadratic terms are used in the blocks. Block D has eight columns and block P has four columns.

The interior block G is filled with physical properties of the granulations. These physical properties consist of the particle size distribution (800, 550, 428, 284, 169, 100, 38 µm), the median granule diameter (D₅₀; µm), the flow of the granulation through funnels with orifices of 9, 6, and 4 mm (g.s⁻¹), the poured volume of the granulation (ml.g⁻¹), the tapped volume after 1000 taps (ml.g⁻¹). These granule properties more or less describe a latent variable that characterizes the flowability of the granules.

The right end block Z consists of the tablet properties. In all cases only one tablet property was handled at the time. The MBPLS path model is capable of handling more responses simultaneously, but then the interpretation becomes much more difficult. The crushing strength of the tablets and the disintegration time are the two tablet responses that will be handled separately with the model. The responses were logarithmically transformed because they both have heteroscedastic variance structures.

Block G of the granule properties needs to be weighted before calibration. This is necessary because block D has to predict both blocks G and Z simultaneously. The granule block however is much larger (14 columns) than the tablet block (1 column). When both blocks are autoscaled, the total variance of the blocks equals the number of columns. The granule block would, therefore, be favoured. The weigh factor of block G is defined to be the number by which G is divided. With a weight factor of 2 all granule properties are divided by 2 (after auto scaling). The weighing must not be too strong because the granule properties also are used in predicting the tablet properties. A too high weight factor would diminish this extra prediction power.

For all predictor blocks a t score vector will be calculated. This t vector is a linear
combination of all variables in the specific block. The vector $w$ contains the weights for each variable of the block for the $t$ score. The $t$ score vector will be used as predictor for the next block or blocks. Furthermore, for each predictee block a $u$ score vector is estimated. The $c$ weights vector gives the weight for every variable in the block according to the $u$ score.

**Notation**

Matrices are denoted by bold uppercase characters: $D$. Vectors are always column vectors and denoted with bold lowercase characters: $t_0$, $D'$. Scalars are denoted with normal lowercase characters: $d$. No multiplication signs are used, so $G'u_3$ means the transpose of matrix $G$ times the $u_3$ vector. Predictions of matrices of vectors are indicated with a hat: $\hat{n}_2$. Each block has the same number of objects, the number of variables may be different for each block. The size of each matrix or vector is given.

$I$ number of objects  
$J$ number of variables  
$K$ number of PLS factors  
$D$ block with mixture composition and process variables of step 1 ($I'J_0$)  
$G$ block with physical granule properties ($I'J_0$)  
$P$ block with process variables for step 2 ($I'J_0$)  
$Z$ block with tablet properties ($I'J_0$)  
$T$ temporary super block combining $t$ scores of $D$, $G$ and $P$ ($I'3$)  
$U$ temporary super block combining $u$ scores of $G$ and $Z$ ($I'2$)  
$t_{D,G,P}$ predictor block score of block $D$, $G$, $P$ ($I$)  
$t_r$ predictor super score ($I$)  
$w_{D,G,P}$ predictor block weight of block $D$, $G$, $P$ ($I,J_0,J_0,J_0$)  
$w_T$ predictor super weight for variables in super block $T$ ($3$)  
$u_{Z,G}$ predictee block score of block $Z$, $G$ ($I$)  
$u_r$ predictee super score ($I$)  
$c_{G,Z}$ predictee block score of block $G$, $Z$ ($J_0,J_2$)  
$c_U$ predictee super weight for variables in super block $U$ ($2$)  
$||w_D||$ norm of $w_D$  
$p_{D,G,P}$ predictive block loading of $D$, $G$, $P$ ($J_0,J_0,J_0$)  
$q_{Z,G}$ predictee block loading of $Z$, $G$ ($J_0,J_2$)  
$b_u$ regression coefficient between $t_0$ and $u_j$  
$b_{UG}$ partial regression coefficient of $t_0$ to the column in $U$ belonging to $G$  
$b_T$ regression coefficient between $t_r$ and $u_2$  
$b_{TP}$ partial regression coefficient of the column in $T$ belonging to $P$ to $u_2$  
$E_{D,G,P,Z}$ residuals of $D$, $G$, $P$, $Z$ after subtraction of explained variance ($I'J_0,J_0,J_0,J_2$)  
$r_G$ the fraction of $G$ being a predictor  
$s_G$ the fraction of $G$ being a predictee  
$n_{D,G,P}$ block $D$, $G$, $P$ data for a new object ($J_0,J_0,J_0$)  
$t_{D,P,G}$ predictor score for block $D$, $P$, $G$ for new object
The multiblock PLS path algorithm

The algorithm is divided into a backward phase, where the predictor vectors (t, w) are calculated, and a forward phase for the predictee vectors (u, c). The phases alternate until u₂ converges. The first step is to scale and mean centre each block. Furthermore, the blocks can be weighted according to additional information. For initialisation a t and u vector for each block are selected. These may be the column with maximal variance.

**Backward phase**

In the backward phase, the t scores of the predictor blocks are calculated. Figure 2A shows an arrow scheme for the backward phase. Block P and G predict only the tablet property and can be calculated directly:

\[ w_G = G' u_2; \text{ scale } w_G \text{ to } ||w_G||=1 \]
\[ t_G = G w_G \]
\[ w_P = P' u_2; \text{ scale } w_P \text{ to } ||w_P||=1 \]
\[ t_P = P w_P \]

Both block scores \( t_G \) and \( t_P \) have maximal covariance with \( u_2 \). Block D has to predict both G and Z. To come to a t₀ score that predicts both blocks, a temporary U block is defined, that contains the u scores of all blocks that are predicted by the specific block.

\[ U = [u_G, u_Z] \]

An ordinary PLS2 step is performed between D and U to calculate the t₀ score.

\[ c_U = U' t_D; \text{ scale } c_U \text{ to } ||c_U||=1 \]
\[ u_U = U c_U \]
\[ w_D = D' u_U; \text{ scale } w_D \text{ to } ||w_D||=1 \]
\[ t_D = D w_D \]

**Forward phase**

In the forward phase, the u scores of the predictee blocks are determined. Figure 2 shows an arrow scheme of the forward phase. Only the blocks G and Z are predicted and need a u score. G is only predicted by D, so \( u_G \) can directly be calculated:

\[ c_G = G' t_D; \text{ scale } c_G \text{ to } ||c_G||=1 \]
\[ u_G = G c_G \]
Figure 2: Arrow scheme of the backward (A) and forward phase (B) for the development of the MBPLS path model. In the backward phase u scores are combined in U to determine $t_0$, and in the forward phase t scores are combined into T to calculate $u_z$. 
Block $Z$ is predicted by $D$, $G$ and $P$. Now a temporary $T$ block is introduced consisting of the $t$ scores of these blocks.

$$T = [t_D, t_G, t_P]$$

An ordinary PLS2 step is performed between $Z$ and $T$ to calculate the $u_z$ score.

$$w_T = T' u_z; \text{ scale } w_T \text{ to } ||w_T||=1$$
$$t_T = Tw_T$$
$$c_Z = Z't_T; \text{ scale } c_Z \text{ to } ||c_Z||=1$$
$$u_z = Zc_Z$$

After completing one cycle of backward and forward phase, $u_z$ is tested for convergence within a desired precision (e.g. $10^{-8}$).

**Loadings**

Loadings for predictor blocks ($p$) and predictee blocks ($q$) are calculated. Just as in the multiblock PLS algorithm introduced in Chapter 7, a distinction can be made between block scores and super scores. Super scores appear when two or more blocks are combined to do a prediction. The block scores $t_D$, $t_G$ and $t_P$ are combined to give the super score $t_T$. For the loadings of $G$ and $P$ the super score update method is used. Block $D$ however, also predicts $G$ directly. The block score $t_D$ is used for calculation of the loading and residual of $D$, because the super score $t_T$ is also partly dependent on $t_D$. This part which may also be present in $D$ would be subtracted of $D$ without ever being used to estimate $G$.

$$p_D = D't_D/(t_D't_D)$$
$$p_G = G't_D/(t_D't_D)$$
$$p_P = P't_D/(t_D't_D)$$
$$q_G = G'u_0/(u_0'u_0)$$
$$q_z = Z'u_z/(u_z'u_z)$$

**Path regression coefficients**

Path regression coefficients are calculated for each block involved in prediction.

$$b_U = u_U't_D/(t_D't_D)$$
$$b_{UG} = c_Ub_U/(c_U'c_U)$$
$$b_{UZ} = c_Ub_U/(c_U'c_U)$$
$$b_T = u_Z't_T/(t_T't_T)$$
$$b_{TD} = w_Tb_T/(w_T'w_T)$$
$$b_{TG} = w_Tb_T/(w_T'w_T)$$
$$b_{TP} = w_Tb_T/(w_T'w_T)$$
The regression coefficients $b_U$ and $b_T$ are used for prediction of $G$ and $Z$ respectively. $b_{UG}$ and $b_{TG}$ are used to determine the predictor and predictee part of the block $G$.

**Residuals**

The calculation of residuals for each block depends on the role of the block. For block $D$, a left end block, the block score update method is used for reasons given earlier. For the second left end block $P$, the super block score will be used.

$$
E_D = D - t_D p_D', \\
E_P = P - t_T p_T'.
$$

The residual of the right end block $Z$:

$$
E_Z = Z - \hat{u}_Z c_Z', \text{ where } \hat{u} = b_T t_T
$$

The residuals of interior blocks are calculated according to a weighted average of its role as predictor and predictee. Block $G$ is the only interior block. The predictor and predictee roles of $G$ are determined by the ratio of the regression coefficients that take part in predicting $Z$ from $G$ ($b_{TG}$) and in the prediction of $G$ from $D$ ($b_{UG}$).

The fractional role of $G$ as a predictee block from $D$:

$$
r^2_G = \frac{b^2_{TG}}{(b^2_{UG} + b^2_{TG})}
$$

The fractional role of $G$ as a predictor block to $Z$:

$$
s^2_G = \frac{b^2_{TG}}{(b^2_{UG} + b^2_{TG})} \text{; so } r^2_G + s^2_G = 1
$$

The residual of the interior block $G$:

$$
E_G = G - (s_G t_G p_G' + r_G \hat{u}_G c_G'), \text{ with } \hat{u} = b_U t_D
$$

In the next round for the calculation of the following scores and loadings, blocks $D$, $G$, $P$, and $Z$ are replaced by $E_D$, $E_G$, $E_P$ and $E_Z$ respectively.

The number of factors that will be used in the model can be estimated by validation with a test set or by cross validation. The number of MBPLS factors that gives the lowest prediction error is selected for the final model. However, in the MBPLS model of the granulation process, two blocks are predicted. The user has to decide which prediction is the most important. It is also possible to combine both prediction errors in order to find a compromise for the final model.
Prediction

Prediction with the MBPLS path model depends on the calculation of the appropriate t scores for the various blocks of data. For prediction, data for all the left end blocks of the new objects have to be known. Right end blocks are always predicted by the model and are unknown. If data for the interior blocks are unknown, they can be predicted by blocks at the left of the specific block. If the data for these blocks are known, they can be used to improve prediction of the block at the right of the specific block. For prediction, weights of both predictor and predictee blocks and loadings of the predictor blocks have to be used. Furthermore, the regression coefficient $b_{Tz}$ and $r_0$ and $s_0$ are necessary for prediction.

In the tablet manufacturing process the granulation block is the only interior block. At the start of the process $G$ is unknown. When a granulation step has been carried out, $G$ can be measured. Let $n_0$ and $n_p$ be the values for $D$ and $P$ for the new object. First these new values have to be scaled according the scaling of the training set. For the new data for the left end blocks, new t scores can be calculated:

$$t_0 = n_0 w_0; \quad e_0 = n_0 - t_0 p_0'$$

$$t_p = n_p w_p; \quad e_p = n_p - t_p p_p'$$

The parts explained by the first $t_0$ and $t_p$ scores are subtracted from the new data. In the next round $n_0$ and $n_p$ are replaced by $e_0$ and $e_p$ respectively. The granule properties of the new experiment $\hat{n}_G$ can easily be predicted.

$$\hat{u}_G = b_{DG} t_0; \quad \hat{n}_G = \hat{u}_G c_G$$

Besides the response for the first step, the granule block is also a predictor block for the tablet properties.

$$\hat{t}_G = \hat{n}_G w_G$$

If real data for $n_G$ is known, a real $\hat{t}_0$ can be calculated, and this value can be used instead of the predicted $\hat{t}_0$. Prediction of the right end block of tablet properties $\hat{n}$ is performed by combining all t scores from the blocks that predict $\hat{n}_2$. These scores are combined in the temporary $n_T$. Prediction of $t_T$ is done by summation of all t values in $n_T$ with their corresponding weight in $w_T$.

$$\hat{t}_T = \sum_{1}^{NT} n_{T,1} w_{T,1}$$

where $NT$ is the number of t scores in $n_T$. The new tablet properties become:

$$\hat{u}_2 = b_{T2} t_T; \quad \hat{n}_2 = \hat{u}_2 c_2$$
Results and discussion

Several variations can be implemented in the MBPLS path model. Most important ones are the use of the super score or block score update of the blocks, the calculation of the residual of the interior block and the construction of the super blocks $T$ and $U$. As was indicated in Chapter 7, the block score update method subtracts information from the blocks that is never used for prediction purposes. Therefore, less variation of the response can be modelled. However, by using the super score update method, the block scores of the various factors become dependent of each other. The residual of the interior block can be determined in several manners. Wangen and Kowalski defined a predictor and a predictee part of the interior block, and the residual is calculated according to these parts. The construction of the super blocks $T$ and $U$ may also be changed. According to Wangen, the $T$ block that is used to calculate the $u_z$ score must contain not only the $t$ scores of the predicting blocks, but also the $u$ scores of these blocks. In the present model the left end blocks, $D$ and $P$, would have been favoured in the $t$ super score estimation because for left end blocks the $u$ scores equal the $t$ scores, but $u_0$ and $t_0$ differ. In the present path model only the $t$ scores were used in the $T$ block.

The MBPLS path model has been developed with the data described above and given in Tables 4 and 5 of Chapter 5. The path model can be evaluated for prediction performance of the tablet property in two different manners. In case the granule properties have not yet been measured for the new experiment (at the start of the process) the model can be compared with a PLS1 model with the variables of $D$ and $P$ as the only descriptors. This PLS1 model is shown in Chapter 6 as Model 1. However, in case the granulation step has been carried out and the granule properties have been measured, the path model can be compared with a PLS1 model with variables of $D$, $P$ and $G$ as the descriptors, which has been shown in Chapter 6 as Model 2. For the development of the path model, the granule block $G$ is always required.

For both response variables, the crushing strength (CS) and disintegration time (DT) of the tablets, an optimal weight factor was found for the granule block $G$ in the model development. The optimal weight factor is influenced by the trade off between the amount to which $G$ participates in the prediction of the tablet property, and the magnitude in which $t_0$ is drawn away from the tablet response to the granule properties. When $t_0$ is drawn in the granule direction, prediction of the tablet property gets worse.

Figure 3 shows the minimal PRESS values for both tablet response variables CS and DT for both cases, without and with block $G$ used in prediction. When the granule properties are not used, the minimal PRESS goes to the same level as was earlier determined for the PLS1 model [4], and shown in Table 1, when the weight factor of $G$ increases. For both tablet responses the minimal PRESS values were obtained at three PLS factors. When the weight of $G$ is too high (the weight factor is low), $t_0$ is trying to fit too much of $G$, and $Z$ is somewhat neglected. This results in higher PRESS values.
Figure 3: Minimal PRESS values for the crushing strength (CS) and the disintegration time (DT) without (black lines) and with (dotted lines) block G included in prediction when the weight factor of G is increased during development of the model. PRESS values of the PLS model with and without G included (straight dotted lines) are also given.

for the tablet property in Z. When the weight for G is low (a high weight factor), t₀ is free to model only the tablet response, but the extra prediction power of G has disappeared because of the little information used from G, as a result of the high weight factor. A compromise between these two features gives the best prediction model for the tablet properties.

Predictions of the tablet properties with the MBPLS path model improve when the granule properties have been measured and can be used in the prediction. For both the crushing strength and the disintegration time, the minimal PRESS values have been reached when the weight factor for block G is about 2.5 when G was not included in prediction, and a weight factor of 2 when G was used in prediction. When the weight factor is above 3, PRESS increases again. The optimal weight factor is dependent on the use of G in prediction. A weight factor of 2.5 is used for the final models because this weight factor gives low PRESS values for both cases (G included and G not included).

Table 1 shows the minimal PRESS values for the PLS1 model and the MBPLS path model. The MBPLS path model gives almost equal prediction errors for both response variables as the PLS1 method when G is not included. When G is introduced, minimal PRESS values decrease for both methods.
Table 1: Minimal PRESS values for both the crushing strength (CS) and disintegration time (DT) for the PLS1 model and for the MBPLS path model without and with block \( G \) included in the model. The weight factor was set to 2 and 2.5.

<table>
<thead>
<tr>
<th>Minimal PRESS</th>
<th>CS</th>
<th>DT</th>
<th>CS (G incl.)</th>
<th>DT (G incl.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS1 model</td>
<td>5.2</td>
<td>25.5</td>
<td>2.8</td>
<td>16.2</td>
</tr>
<tr>
<td>MBPLS path model, weight factor 2.5</td>
<td>5.3</td>
<td>25.5</td>
<td>2.8</td>
<td>17.7</td>
</tr>
<tr>
<td>MBPLS path model, weight factor 2</td>
<td>5.4</td>
<td>25.9</td>
<td>2.7</td>
<td>17.5</td>
</tr>
</tbody>
</table>

In the MBPLS path model, the scores of block \( D \) have slightly been drawn towards the granule block. This has no negative influence on the prediction quality of the models for the tablet responses. As a bonus one gets a prediction for the granule properties with the same model. However, the prediction of the granule properties with the MBPLS path model are not as good as the use of a standard PLS2 model for the granule properties with block \( D \) as the single descriptors. Table 2 shows the modelling of the granule properties with the MBPLS path model and the standard PLS2 model.

The standard PLS2 model explains almost twice as much of the variance in block \( G \) as the MBPLS path model. The PRESS value is also much lower for the PLS model. The prediction of the MBPLS path model for block \( G \) is, as can be expected, not as good as a simple PLS2 model for the granule properties because the scores are selected to give an optimal fit for the tablet response variables. Furthermore, the optimal number of latent variables for the model are selected according to the lowest PRESS for the tablet properties.

Table 3 shows the explained variances of all blocks for both response variables. The \( \% \) \( G \) presented, includes both the amount used to predict \( Z \) and the amount \( G \) is predicted by \( D \). The weight factor for block \( G \) is set to 2.5. The first factor describes 77 and 71\% of the crushing strength and the disintegration time respectively. This variation is mainly described by block \( D \), where block \( P \) is the main source of information in the second factor. This not only follows from the explained variation of the blocks in Table 3, but it can also be seen from the weight vector \( w \) in Table 4b.

To examine the model, the modelling of the crushing strength is studied in detail to study the properties of the MBPLS path model. The block \( D \) score \( t_d \) has to fit \( u_j \) as good as possible where \( u_j \) is a linear combination of the \( u_g \) and \( u_z \) scores. The weight \( c_u_j \) gives the weights for block \( G \) and \( Z \) respectively for each factor (f1-f4) in the \( u_j \) score.

Table 2: Prediction of \( G \) with PLS2 and MBPLS path model when the crushing strength (CS) or the disintegration time (DT) is modelled. The weight factor for \( G \) in the path model is set to 2.5.

<table>
<thead>
<tr>
<th>model</th>
<th>% G explained</th>
<th>PRESS</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS2</td>
<td>52.8</td>
<td>69</td>
</tr>
<tr>
<td>MBPLS path (CS, w=2.5)</td>
<td>29.5</td>
<td>93</td>
</tr>
<tr>
<td>MBPLS path (DT, w=2.5)</td>
<td>29.2</td>
<td>94</td>
</tr>
</tbody>
</table>
Table 3: Cumulative explained variances of the blocks D, G, P and the response Z in the MBPLS path model for both the crushing strength (CS) and disintegration time (DT) of the tablets. The weight factor for G was set to 2.5.

<table>
<thead>
<tr>
<th>% explained</th>
<th>%D</th>
<th>%G</th>
<th>%P</th>
<th>%Z</th>
</tr>
</thead>
<tbody>
<tr>
<td># Factor</td>
<td>CS</td>
<td>DT</td>
<td>CS</td>
<td>DT</td>
</tr>
<tr>
<td>1</td>
<td>29</td>
<td>29</td>
<td>17</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>46</td>
<td>19</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>57</td>
<td>58</td>
<td>43</td>
<td>44</td>
</tr>
<tr>
<td>4</td>
<td>69</td>
<td>67</td>
<td>62</td>
<td>63</td>
</tr>
</tbody>
</table>

(Table 4a). In the first two factors D is mainly used to fit the u_z score (-0.99 and -0.98) where only in the third and fourth factor u_G is fitted (0.83 and 0.92). Most of the variation in D is used to model Z instead of G.

The super score t_f, the score to fit Z, is a linear combination of t_D, t_G, t_P and t_T. The super weight w_f gives the weights for the scores for all four factors respectively (Table 4b). The super weight w_f shows just as the percentage explained in Table 3 that the first factor mainly consists of D to fit Z. In the second factor, the information of P is used and in the last two factors the t_f scores are chiefly composed of the information from block G.

The interior block G is mainly used as a predictor, and only slightly as a predictee

Table 4a: The weights for the scores u_G and u_z in c_0 for the four PLS factors f1-f4 for the modelling of the crushing strength.

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
</tr>
</thead>
<tbody>
<tr>
<td>u_G</td>
<td>0.17</td>
<td>-0.21</td>
<td>0.83</td>
<td>0.92</td>
</tr>
<tr>
<td>u_z</td>
<td>-0.99</td>
<td>-0.98</td>
<td>-0.56</td>
<td>-0.39</td>
</tr>
</tbody>
</table>

Table 4b: The weights for the scores t_D, t_G and t_P in w_f for the four PLS factors f1-f4 for the modelling of the crushing strength.

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
</tr>
</thead>
<tbody>
<tr>
<td>t_D</td>
<td>-0.84</td>
<td>-0.35</td>
<td>-0.28</td>
<td>0.33</td>
</tr>
<tr>
<td>t_G</td>
<td>0.40</td>
<td>0.28</td>
<td>0.83</td>
<td>0.94</td>
</tr>
<tr>
<td>t_P</td>
<td>0.38</td>
<td>0.90</td>
<td>0.48</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table 4c: The role of G as a predictor (s_G) and as predictee (r_G) in the modelling of the crushing strength.

<table>
<thead>
<tr>
<th></th>
<th>f1</th>
<th>f2</th>
<th>f3</th>
<th>f4</th>
</tr>
</thead>
<tbody>
<tr>
<td>s_G</td>
<td>0.94</td>
<td>0.97</td>
<td>0.29</td>
<td>0.82</td>
</tr>
<tr>
<td>r_G</td>
<td>0.35</td>
<td>0.26</td>
<td>0.96</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Figure 4: 4A) A scatter plot of the first two super scores $t_{1(1)}$ and $t_{1(2)}$ in the modelling of the crushing strength. Three groups can be distinguished (X, MCC=90%; +, Fup=10 kN; •, other experiments; * belongs to both the X and + groups). 4B and 4C show $u_1$ against the predicting super score for the first and the second factor. 4D shows the crushing strength predicted without (O) and with (●) use of the granule properties. The straight line represents the perfect fit.

block as indicated by $s_3$ and $r_3$ respectively (Table 4c). In the first two factors the predictor parts of $G$ (0.94, 0.97) are much larger than the predictee parts (0.35, 0.26). Only in the third factor, the granule properties are described by the D block.

Figure 4A shows a scatterplot of the first two scores of the super block $T$. Three groups can be distinguished, a wide group at the left (X), a group at the bottom (+) and a third group (●). Object * belongs to the first and the second group. The first group consists of experiments with a composition of 90% MCC. This information comes from block D. Experiments of the second group have compression forces of 10 kN. The compression force is placed in block P. The third group exists of the other experiments. Figures 4B and 4C show the $t_1$ and $u_2$ scores of the first two factors. In 4B, the fit of the super score $t_1$ against $u_2$ is mainly due to the MCC=90% group. Table 2 and $w_f$ already
Figure 5: Regression coefficients for the variables of blocks D(1-8) and P(9-12) for the crushing strength (CS) and the disintegration time (DT) when the weight factor of block G during development of the path model has been increased. Some variables are indicated with their corresponding number: 1=MCC, 2=water, 3=time, 4=HPC, 5=MCC, 9=F, 11=F and 12=motion.

showed that the first super score \( t_1 \) was built primarily of the \( t_0 \) score, and the second super score on \( t_p \). The latter is shown in Figure 4C where the 10kN group is the main predictor of small crushing strengths.

Figure 4D shows the predictions of the crushing strength of the tablets with the MBPLS path model, without (O) and with (●) the granule properties used for prediction. The predictions are comparable to the predictions with the PLS and MBPLS model in Chapter 6 and 7. When \( G \) is used, the predicted values of the tablet responses are closer to the observed ones than without \( G \) used.

Regression coefficients

Figure 5 shows the regression coefficients of the variables in block D and P for the modelling of the crushing strength and disintegration time with the path model when the weight factor of block G increases. Some regression coefficients change when the weight factor increases, but they stabilize at a certain level. For the crushing strength, the coefficients of MCC and MCC\(^2\) (1 and 5) show the largest change. These variables also have a large influence on the granule properties. When \( t_0 \) is forced to fit mainly the tablet properties by increasing the weight factor for \( G \), the regression coefficients of
Figure 6 shows the regression coefficients for the variables of blocks D, P and G for the modelling of crushing strength and disintegration time for four different regression methods: the MBPLS path model, the OLS regression method, the PLS method and the MBPLS_a method according to Wangen. The PLS1 method equals the MBPLS method with super score update of the residuals, and the MBPLS_a method uses the block score update method. The regression coefficients were determined by examining the change in predicted values when the values for the specific variable were increased by 1 (after autoscaling of the data). The effect of the variables of D (MCC, water, time, HPC, MCC², water², time², HPC²), P (compression force F_{up}, moisture, F_{up}² and moisture²) and G (granule properties) is found almost the same for all methods except for the OLS regression method. It is obvious that the OLS models are very different from the several PLS models. For the variables 13-27, which are the variables of the G block, the MBPLS_a method gives slightly deviating coefficients compared to the other two PLS
methods. This is mainly caused by the fact that for the MBPLS\textsubscript{a} method only three PLS factors were used (because three were optimal for this method) and for the other two PLS methods, four factors were used.

The effect of the variables of block D and P on the tablet responses changes when the G block variables are included in the model because of the correlations that exist between D and G. Figure 7 shows the differences of the OLS and PLS coefficients for the variables in blocks D and P for the modelling of CS and DT caused by the introduction of the granule properties in the model. The OLS method, as could be expected, gives the largest differences, because OLS suffers from correlations between the predictor variables. The PLS model, which equals the MBPLS model with super score update, suffers somewhat more of the introduction than the MBPLS\textsubscript{a} model. The MBPLS path model shows the smallest differences in the coefficients. The path model is developed with the granule properties present, even if prediction is done without the granule properties. For the other two PLS methods, the model that predicts Z without G is developed without G. When the granule properties are included, another model is
used that was developed with G present. This may cause the effect of the D and P variables to change more than for the MBPLS path model.

Conclusion

The multiblock PLS path model can be used for the modelling of complex processes with two or more steps such as the wet granulation and tableting process. During development of the model, the granule properties have to be used. For prediction purposes, the granule properties may be used if they have been measured (after the first step), but prediction can also be done when they have not been obtained yet. The prediction properties of the path model for the tablet responses are comparable to the standard PLS methods where all data is combined in 1 block to model the tablet responses. The MBPLS path model is less influenced by introduction of the granule properties in the model than the standard PLS or MBPLS methods. The path model can be used to study the real effect of the process and composition variables and the granule properties on the tablet responses.

The prediction of the granule properties with the MBPLS path model is not as good as with a standard PLS2 model. The information of block D is mainly used to model the tablet responses. For monitoring of the whole wet granulation and tableting process, it seems better to use different models for the monitoring of the granule and tablet properties.

Multiblock pathway models may be used when interior blocks are present in the process, i.e. blocks in the middle of a process that are predicted by previous blocks, and predict subsequent blocks. Just as the standard PLS models, the MBPLS path model gives outlier detection and noise reduction for each block separately. All blocks can be predicted with the same model. Predictions of the right end blocks become better when the interior blocks can be filled with measured values. Furthermore, the path model may provide extra information on the way latent variables work through the process, which may lead to a better understanding of the process.

References

1. Jöreskog KG and Wold H, Systems under indirect observation, North Holland, Amsterdam, 1982, Parts I and II.
3. Wangen LE and Kowalski BR, A multi block partial least squares algorithm for investigating complex
4. Chapter 6, this thesis.