Statistical Analysis of sequential Process Chains based on Errors-in-Variables Models

Dipl.-Stat. Oliver Meyer
and Prof. Dr. Claus Weihs
Chair Computational Statistics
TU Dortmund University
Email: meyer@statistik.tu-dortmund.de
and weihs@statistik.tu-dortmund.de

***Abstract*—A process chain comprises a series of sequential (production) processes, mostly in the area of manufacturing engineering. It describes a consecutive sequence of activities, which together form one single system. Within this system the sub-processes are presumed to influence each other by transferring characteristics. The single process steps of such a system can easily be simulated using regression (or other statistical learning) methods. The main obstacle in simulating entire process chains, however, is to determine how to handle prediction uncertainty in the transferred characteristics. In this paper, we will describe how using Error-in-Variable models instead of ordinary regression models can solve this problem. We especially focus on the question how uncertainty (measured by variance) develops through the process chain and its influences on the results along the process chain. We will also discuss how the presented methods can be applied to the field of process control. At this point, our research is mainly limited to polynomial regression, but the basic principals can be applied to other statistical learning techniques, including classification and time series as well.**

I. Introduction

Industrial (production) processes usually do not consist of one single production step but of several sequential sub-processes. In common process analysis these sub-processes are treated to be independent of one the other, even though they interact in the way that the results of earlier sub-processes are often critical for the performance of later ones. A more suitable approach to describe this kind of processes was introduced under the term process chains [1]. A process chain by definition describes a consecutive sequence of activities (sub-processes), which together are seen as one single system. Within this system the sub-processes are presumed to influence each other by transferring characteristics or features from one to another. Throughout the process chain inputs are sequentially transformed from one state into another. Figure 1 illustrates an example of a process chain from the field of automotive battery production. It shows that the production indeed consists of several consecutive operations which are conducted mainly independent from each another. They only influence each other through the characteristics of the manufactured "intermediate product" that is passed forward to the next sub-process. This is in fact a very simple example of a process chain. In more sophisticated applications some of the sub-processes could, for example, be conducted on parallel production lines without being directly connected. Such a system would still qualify as one process chain as long as all these production lines merge into a single sub-process at some point of the production chain.

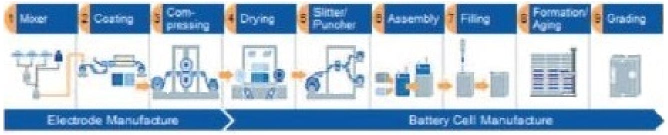

Fig. 1. Process Chain of Automotive Battery Production

In the following section, we will talk about simulating complete process chains using statistical learning techniques, focusing on the question how to handle the connection between the sub-processes appropriately. This will lead us to the introduction of the principals of Error-in-Variable models for polynomial regression in section III. Since there are two different types of Error-in-Variable models, depending on the nature of the observed error, we will argue which of them suits the presented problem in a better manner. In section IV, we will extend our presentation to the case of errors in the variables in polynomial regression models. Furthermore, we will discuss how these kinds of models can be implemented to simulate process chains in section V. In section VI we will suggest how the insight we presented in the previous sections can be used to enhanced process control of process chain. At last, we will give an outlook over future developments in section VII. Before, we summarize our results in section VIII.

II. Simulation of Process Chains

It is easy to see that the single sub-processes of a process chain can be simulated using regression, or depending on the type of sub-process other statistical learning methods. Two different types of factors serve as independent variables for these models: The parameters of the sub-process itself, like machine settings, and the characteristics of the product at the beginning of the sub-process. Its characteristics after the sub-process would then be the target variables. For every target variable a unique model would have to be trained. In the structure of the simulated process chains these models would be interpreted as parallel sub-processes based on the same independent variables. For reasons of simplicity we will assume that we are only interested in one single characteristic of the intermediate product after each sub-process from this point on, so every sub-process can be represented by one single model. With y_i and $x_{i,j}$ being the target and independent variables, $\beta_{i,j}$ being the regression parameters, and ϵ_i being the residuals of model i, a process chain consisting of four sub-processes mapped by linear regression models could look as follows:

LM1: $y_1 = \beta_{1,0} + \sum_{i=1}^{k1} x_{1,i} \cdot \beta_{1,i} + \epsilon_1$

LM2: $y_2 = \beta_{2,0} + \sum_{i=1}^{k_2} x_{2,i} \cdot \beta_{2,i} + \mathbf{y_1} \cdot \beta_{2,(k_2+1)} + \epsilon_2$

LM3: $y_3 = \beta_{3,0} + \sum_{i=1}^{k_3} x_{3,i} \cdot \beta_{3,i} + \mathbf{y_2} \cdot \beta_{3,(k_3+1)} + \epsilon_3$

LM4: $y_4 = \beta_{4,0} + \sum_{i=1}^{k_4} x_{4,i} \cdot \beta_{4,i} + \mathbf{y_3} \cdot \beta_{4,(k_4+1)} + \epsilon_4$

An easy and straight forward way to predict the results at the end of a complete process chain would be to link these models together with respect to the structure of the original process chain. This would mean that the predicted results of earlier models are used as independent variables in later ones. Since these values are subject to prediction uncertainty all predictions starting from the second sub-process would be made based on faulty values in at least one independent variable. Although the predictions made by every single model are unbiased, this can still lead to problems in both prediction precision and accuracy for the whole process chain, as we will show later in this paper. Luckily, the theory of Errors-in-Variables regression addresses exactly this kind of problems. In the following sections, we will introduce these kinds of regression models and discuss how they can be applied to simulate process chains more accurately. Since up to now this kind of regression was mainly used to simulate single processes we will handle the question of how the resulting models influence each other throughout a process chain in sections V and VI.

III. Error-in-Variables Models

Errors-in-Variables models deal with statistical learning problems where the values of the independent variables are subject to errors. This means that the true value x of at least one independent variable is unknown. Instead, only a faulty value x^* which differs from x by an unbiased (and in most cases normal distributed) error term u is known or can be observed. For this reason, the value x^* is called the observed value of the independent variable x. In the situation described in section II, we use predicted values as independent variables and these prediction ($\hat{y}$) differ from the true values (y), only by the prediction error ϵ. So in this case $\hat{y}$ is the observed value of y.

In 1950, J.Berkson [2] proved that for linear regression the effect of errors in the independent variables on the models in the observed variables differs highly with respect to the nature of the relationship between the two values x and x^*. Actually, two different kinds of relationships were described which lead to different solutions. Before we proceed, we first have to decide which of the two cases fits our problem best. To do so, we will take a look at both cases for the simple linear regression model. The following model represents the true relationship between the independent and the target variable.

$$y = \beta_0 + x \cdot \beta_1 + \epsilon$$

A. Classical Measurement Error in Simple Linear Regression

In classical Errors-in-Variables models the errors in the variables are the result of measurement uncertainty. In these cases the observed variable x^* equals the true variable x plus an independent measurement error term u with $\mathrm{E}(u) = 0$ and $\mathrm{Var}(u) = \sigma_u^2$, leading to $x^* = x + u$. For example, if the ambient temperature has to be measured. In this case, u and x are independent from each other while u and x^* are not. If we assume the stochastic variable from which the true value x is drawn to be $X \sim N(\mu_x, \sigma_x^2)$, the standard least sum of squares (LSS) estimation for the Error-in-Variables model (although called the model in the observed variables) looks like this:

$$\mathrm{E}(y|x^*) = \beta_0^* + \mathrm{E}(x|x^*) \cdot \beta_1^* = \beta_0^* + x^* \cdot \beta_1^*$$

with

$$\beta_1^* = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} \cdot \beta_1$$

$$\beta_0^* = \beta_0 + \left(1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\right) \cdot \beta_1 \cdot \mu_x$$

And the prediction variance becomes:

$$\mathrm{Var}(y|x^*) = \sigma_\epsilon^2 + \beta_1^2 \cdot \sigma_u^2 \cdot \frac{\sigma_x^2}{\sigma_u^2 + \sigma_x^2}$$

The adjusted model in the observed variable is still a linear model. Its prediction error has grown, which is no surprise since we added uncertainty to the independent variable. But the regression coefficients have changed too. And both changes do not only depend on the variance of the measurement error but also on the variance of the true independent variable X. If we take a closer look, we can see that the values of the regression coefficients change with respect to the ratio of the variance of the true variable X, $\mathrm{Var}(X) = \sigma_x^2$ and the variance of the observed variable $\mathrm{Var}(X + u) = \mathrm{Var}(X) + \mathrm{Var}(u) = \sigma_x^2 + \sigma_u^2$.

B. *Berkson Error in Simple Linear Regression*

In the historically earlier Berkson error case, the true variable x equals the observed variable x^* plus an independent error term u with $\mathrm{E}(u) = 0$ and $\mathrm{Var}(u) = \sigma_u^2$, changing their relationship to $x^* + u = x$. On first sight, this does not seem to differ from the measurement error case. However, it results in a change of dependency between the error and the variables. This case is relevant, if the value of the observed variable can be controlled by an operator. For example, if the ambient temperature can be adjusted by the operator using a thermostat. The temperature entered in the thermostat would then be the observed value x^*. In this case, u and x^* are independent from each other while u and x are not. In this kind of situation, the distribution of the true variable X has no influence on the LSS estimation of the Errors-in-Variables model, which looks like this:

$$\mathrm{E}(y|x^*) = \beta_0 + x^* \cdot \beta_1$$

with a prediction variance of:

$$\mathrm{Var}(y|x^*) = \sigma_\epsilon^2 + \beta_1^2 \cdot \sigma_u^2$$

Thus, the model based on the observed variable is the same as the original one. But still, adding uncertainty to the independent variables leads to a worse prediction precision.

The effect of a measurement error and a Berkson error on a simple linear regression model are illustrated in figure 2.

In both scatterplots the original regression line was estimated using the same data sample (red). Then an amount of uncertainty (meassured by variance) has been added to the independent variable (blue). For the first graphic the uncertainty has been added in form of a classical meassurement error and for the second in form of a berkson type error.

Proofs for both the effect of measurement and Berkson errors on a simple linear regression model can for example be found in [2] and [3].

IV. Error-in-Variables models in Process Chains

When we take a look at the relationship between the predicted and the true values in a simulated process chain, we see that the true values y are in fact the sum of the predicted values $\hat{y}$ and the error term ϵ. Since every predicted value is the result of a linear combination of independent variables ($\hat{y} = \beta_0 + \sum_{i=1}^{k} x_i \cdot \beta_i$), the operator can control it by controlling the independent variables. This leads to the conclusion that the errors created by linking the single models are in fact Berkson type errors. For that reason, in this section we will discuss the effect of such Berkson errors on the simulation of a process chain by expanding the considered models from linear to polynomial regression.

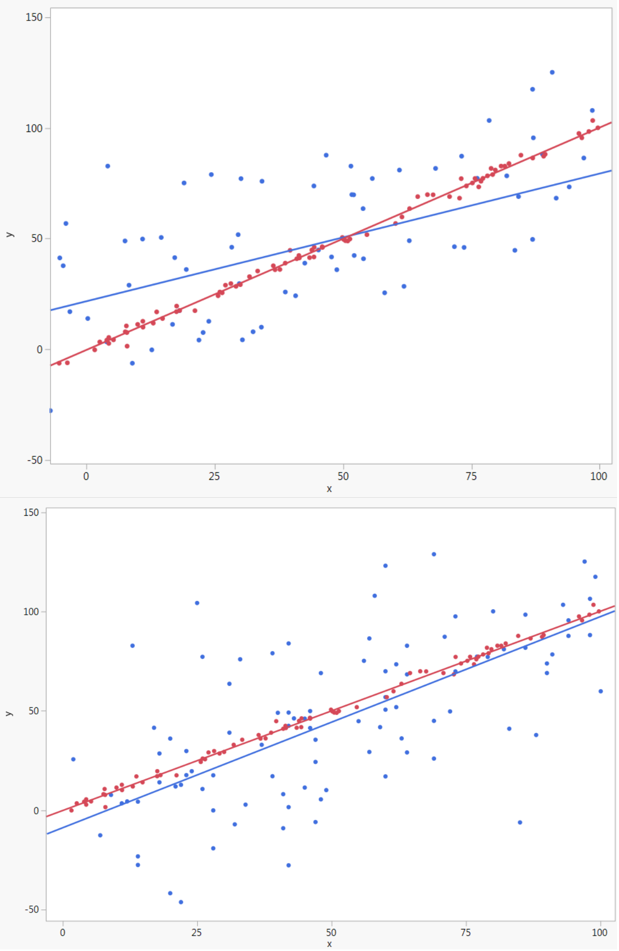

Fig. 2. Effects of different kinds of errors in the independent variable (blue) on a simple linear regression model (red). (top: measurement error, bottom: Berkson-error)

We have said in section II, that we will focus on sequential process chains where only one characteristic is transferred from one sub-process to the next and by that only one independent variable in every regression model will be subject to an error. Yet, we will discuss the theory of Errors-in-Variables models in a more general form allowing different types of errors in more than one independent variable.

A. *Berkson Error in multiple linear Regression Models*

It has been proven in [4], that in case of Berkson errors in a multiple linear regression model with $k > 1$ independent variables the effect is similar to that in a simple regression model. The Errors-in-Variables model stays the same as the one in the true variables and the prediction variance increases. Assume the original model to take the form:

$$y = \beta_0 + \sum_{i=1}^{k} x_i \cdot \beta_i + \epsilon$$

If the first $l \leq k$ variables $x_1, \ldots, x_l$ were subject to Berkson style errors $u_1, \ldots, u_l$ and we would define $x^* = (x_1^*, \ldots, x_l^*, x_{l+1}, \ldots, x_k)$, the observed model and its prediction variance would lead to:

$$\mathrm{E}(y|x^*) = \beta_0 + \sum_{i=1}^{l} x_i^* \cdot \beta_i + \sum_{i=l+1}^{k} x_i \cdot \beta_i$$

$$\mathrm{Var}(y|x^*) = \sigma_\epsilon^2 + \sigma_u^2 = \sigma_\epsilon^2 + \sum_{i=1}^{l} \beta_i^2 \cdot \sigma_{u_i}^2$$

B. Berkson Error in multiple linear Regression Models with Interactions between two variables

Let us now consider a situation where the linear regression model includes interaction effects between some of the variables:

$$y = \beta_0 + \sum_{i=1}^{k} x_i \cdot \beta_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} x_i \cdot x_j \cdot \beta_{i,j} + \epsilon$$

Using the notations above, the conditional expected value and variance become (for a proof, see [4]):

$$\begin{aligned}\mathrm{E}(y|x^*) = \beta_0 &+ \sum_{i=1}^{l} x_i^* \cdot \beta_i + \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} x_i^* \cdot x_j^* \cdot \beta_{i,j} \\ &+ \sum_{i=1}^{l} \sum_{j=i+l}^{k} x_i^* \cdot x_j \cdot \beta_{i,j} \\ &+ \sum_{i=l+1}^{k} x_i \cdot \beta_i + \sum_{i=l+1}^{k-1} \sum_{j=i+1}^{k} x_i \cdot x_j \cdot \beta_{i,j}\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(y|x^*) = \sigma_\epsilon^2 &+ \sum_{i=1}^{l} \beta_i^2 \sigma_{u_i}^2 + \sum_{i=1}^{l} \sum_{j \neq i}^{k} \beta_{i,j}^2 x_j^2 \sigma_{u_i}^2 \\ &+ \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} \beta_{i,j} \sigma_{u_i}^2 \sigma_{u_j}^2\end{aligned}$$

Thus, the model in the true variables would still produce unbiased predictions using the observed values. The prediction variance in the observed model, on the other hand, does not only grow but now depends on the true values x_i. This means that the variance is no longer homoscedastic and the LSS estimators are no longer efficient in the observed model.

C. Berkson Error in Quadratic Regression Models with Interactions between two variables

If we add quadratic terms to the model above, another problem occurs in the Errors-in-Variables model. The new true quadratic model takes the form:

$$y = \beta_0 + \sum_{i=1}^{k} x_i \cdot \beta_i + \sum_{i=1}^{k} \sum_{j=i}^{k} x_i \cdot x_j \cdot \beta_{i,j} + \epsilon$$

Now, following the results from [4] the expected value and prediction variance of the model in the observed variables become:

$$\begin{aligned}\mathrm{E}(y|x^*) = \beta_0^* &+ \sum_{i=1}^{l} x_i^* \cdot \beta_i + \sum_{i=1}^{l} \sum_{j=i}^{l} x_i^* \cdot x_j^* \cdot \beta_{i,j} \\ &+ \sum_{i=1}^{l} \sum_{j=i+l}^{k} x_i^* \cdot x_j \cdot \beta_{i,j} \\ &+ \sum_{i=l+1}^{k} x_i \cdot \beta_i + \sum_{i=l+1}^{k} \sum_{j=i}^{k} x_i \cdot x_j \cdot \beta_{i,j}\end{aligned}$$

with

$$\beta_0^* = \beta_0 + \sum_{i=1}^{k} \beta_{i,i} \cdot u_i^2$$

$$\begin{aligned}\mathrm{Var}(y|x^*) = \sigma_\epsilon^2 &+ \sum_{i=1}^{l} \beta_i^2 \sigma_{u_i}^2 + \sum_{i=1}^{l} \sum_{j \neq i}^{k} \beta_{i,j}^2 x_j^2 \sigma_{u_i}^2 \\ &+ \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} \beta_{i,j} \sigma_{u_i}^2 \sigma_{u_j}^2 \\ &+ \sum_{i=1}^{k} \beta_{i,i}^2 \sigma_{u_i^2}^2\end{aligned}$$

with

$$\sigma_{u_i^2}^2 = \mathrm{Var}(u_i^2)$$

Thus, at this point using the true model for prediction based on the observed values would lead to biased estimations. Moreover, using LSS estimation to calculate a model in the observed values would lead to inefficient estimators for the coefficients.

D. Higher dimensional polynomials and efficient estimators

If the sub-processes are simulated using higher dimensional polynomials, the effects described above are getting stronger and the bias becomes a polynomial function of the true values x. We have shown that training regression models in the true variables for all the sub-processes and executing them in the correct order does not provide accurate estimations for a process chain.

However, with respect to the variances of the errors in the variables, the true models can be used to calculate unbiased regression models for the observed variables. In the next section, we will discuss how this can be used for the statistical analysis of process chains.

V. Unbiased Estimation in Process Chains

Before we start to describe how to use the results displayed in the last section, we need to take a closer look on the situation and the data basis we are dealing with.

In most applications of Errors-in-Variables regression, only the observed values are known and the original data and the variances of the errors in the variables are unknown. In this situation, it is usually the aim to identify the model in the true unobservable variables. In the case of process chain analysis we are facing a slightly different task. We actually know the true values of the variables and we want to construct models in the observed variables. Since the observed values are the results of estimations done based on the true values, we also know the observed values. Since we are interested in accurate predictions over the whole process chain, we could use the observed values to calculate regression models. This means that we would train the model for the first sub-process based on the true variables (since there is no error in the variables here). Then, we would predict the values of the target variable for our training data and use these predicted values as substitutes of the true values of the corresponding "independent" variable in the training data set for the simulation of the next sub-process and so on. Now we would be able to predict unbiased results for the whole process chain. We will refer to this as the *practical approach.*

Another way to construct models in the observed variables would be what we call the *functional approach.* From the last section, we know that the expected value of y conditioned on x^* is actually a function of y, x, and u.

$$\mathrm{E}(y|x^*) = \mathrm{E}(y|(x+u) := g(y, x, u)$$

More precisely, it is a function of the expected value of y conditioned on x and the variance of the error σ_u^2.

$$g(y, x, u) = g(\mathrm{E}(y|x), \sigma_u^2)$$

This means that we can calculate the models in the observed variables from the models in the true variables if we know the variances of the errors in the variables.

Although both approaches seem to be sufficient to simulate a process chain, the latter one holds some significant advantages over the practical approach. First of all, we have learned that the errors of the model in the observed variables are heteroscedastic. Because of that, the LSS estimators of the regression coefficients are not efficient, and more complicated ways to predict them have to be considered.

A method to receive more efficient estimators for the observed model are the so-called weighted least squares (WLS) estimators. A way to construct these WLS estimators for Errors-in-Variable regression models has been presented in [5].

By using the true models and adjusting them, we simply avoid this problem. Another reason to choose this method, is that if there is a change in one of the models because a of a change in its corresponding sub-process, the effect on the other models (later in the process chain) can been calculated without the need to train new models. This leaves us with the problem that we still need to know the variance of the errors in the variables at every point of the process chain to calculate the models in the observed variables from the models in the true variables. A solution for this problem will be presented in the next section.

VI. Process Control in Process Chains

In this section, we want to present several ideas on how to use error-in-variables models to improve statistical process control (SPC) and process optimization of sequential process chains. This includes a method to gain a better understanding of the effect of variance in the results of sub-processes and using this knowledge to optimize the results of the whole process chain. We will also propose a new way to predict a process capability index based on the simulated process chain. In this context it is crucial to realize that the prediction variance of the models trained to simulate the sub-processes of a process chain is an estimator for the production variance of the said sub-processes, since the predicted value itself is an estimator for the mean of the production given a fixed set of independent variables.

A. Predictions within the Process Chain

Until now, we have only discussed a way to predict the final results of a process chain from its start. In practical application, we might also be interested in predicting the final result of an intermediate product from a later sub-process. Imagine a process chain containing ten sequential sub-processes. As explained before, one characteristic of the result of every sub-process is transferred to the next one. Now, we start production of one single product. If we know the values of all external input variables, we are able to predict the results of the final product using the regression models calculated following the instructions in the previous section. Let us further imagine that we let our product pass the first two sub-process of the process chain. Thus, we know the true value of the characteristic transferred to the third sub-process. If we now want to predict the final result of that product, the models we calculated earlier are not sufficient to do so. Since we do not have to predict the results of the first two sub-processes, the variance in the prediction of the rest of the process chain has changed.

The best example for this, is the prediction of the third sub-process. Since we know the results of the second sub-process there is no error in the variables of the third model at all. Meaning that we should actually use the model in the true variables to predict this step. As we have seen in section IV, the prediction variance of the true model is different from the variance of the model in the predicted variables. This again changes the error in the variables of the next model and so on.

And it gets even worse: The prediction variance of the models in the observed variables depends on the value of the true variable itself (heteroscedasticity). This means that the following models also depend on that value. So the set of models used to simulate the process chain is only valid for a fixed set of external variables. Luckily, the observed values are unbiased estimators for the true values ($x^* + u = x$ and $\mathrm{E}(u) = 0$) and we are able to calculate new models for every situation, as long as we know the prediction variance of each regression model in the observed variables.

This leads us to another matter of importance: In all our considerations so far, we have assumed that we know the structure of the errors in the variables of all the models. In fact, we only know the prediction variances of the models in the true variables (named σ_ϵ^2 in section IV) because those were the models we trained. However, we can show that the theoretical prediction variance of the models in the observed variables can be calculated from those in the true variables. Let us take another look at the process chain described above and remind ourselves that we use the true model to predict the first sub-process (since there is no error in the variables). Now, the prediction variance of this model is the variance of the error in the variable in the next model, because the predicted value of the target variable of the first model is used as an independent variable in the next one. As we saw in section IV, the prediction variance of a model in the observed variables depends on the prediction variable of the true model and the variance of the errors in the variables. Generally speaking, the prediction variance of any model in the observed variables is a function of the prediction variances of all the true models up to that moment. We call this the variance development within the process chain.

$$\begin{aligned}\mathrm{Var}(y_i|x_i^*) &:= \tilde{f}_i(u_i) + \mathrm{Var}(y_i|x_i)\\ &= \tilde{f}_i(\mathrm{Var}(y_1|x_1), \ldots, \mathrm{Var}(y_{i-1}|x_{i-1})) + \mathrm{Var}(y_i|x_i)\\ &:= f_i(\mathrm{Var}(y_1|x_1), \ldots, \mathrm{Var}(y_i|x_i))\end{aligned}$$

with $i = 1, \ldots, 10$ and y_i being the target variable, and x_i being the vector of independent variables of model i.

For the process chain consisting of four sub-processes mapped by linear regression models that we have presented in section II the variance development would look like this:

$$\begin{aligned}\mathrm{Var}(y_1|x_1) &= \sigma_{\epsilon_1}^2 := \sigma_{u_2}^2\\ \mathrm{Var}(y_2|(x_2, \hat{y}_1)) &= \sigma_{\epsilon_2}^2 + \beta_{2,2}^2 \cdot \sigma_{u_2}^2\\ &= \sigma_{\epsilon_2}^2 + \beta_{2,2}^2 \cdot \sigma_{\epsilon_1}^2 := \sigma_{u_3}^2\\ \mathrm{Var}(y_3|(x_3, \hat{y}_2)) &= \sigma_{\epsilon_3}^2 + \beta_{3,2}^2 \cdot \sigma_{u_3}^2\\ &= \sigma_{\epsilon_3}^2 + \beta_{3,2}^2 \cdot (\sigma_{\epsilon_2}^2 + \beta_{2,2}^2 \cdot \sigma_{\epsilon_1}^2) := \sigma_{u_4}^2\\ \mathrm{Var}(y_4|(x_4, \hat{y}_3)) &= \sigma_{\epsilon_4}^2 + \beta_{4,2}^2 \cdot \sigma_{u_4}^2\\ &= \sigma_{\epsilon_4}^2 + \beta_{4,2}^2 \cdot \sigma_{\epsilon_3}^2 + \beta_{4,2}^2 \cdot \beta_{3,2}^2 \cdot \sigma_{\epsilon_2}^2\\ &\quad + \beta_{4,2}^2 \cdot \beta_{3,2}^2 \cdot \beta_{2,2}^2 \cdot \sigma_{\epsilon_1}^2\end{aligned}$$

A detailed discussion about the development of variance in simulated process chains based on linear regression models can be found in [6].

The insight about this variance development can also be used to help to optimize the process chain. If the functions f_i are known, we can quantify the effect of every sub-process on the variance of the whole process chain and use this information to help to identify which of the sub-processes should be optimized.

B. Instant Process Capability

In SPC process capability indexes are used to determine the ability of a process to produce within given specification limits. While there already are several different approaches to quantify process capability [7], we do not suggest a new form of process capability index but a way to apply them to the special circumstances when dealing with process chains.

The most common version of process capability indexes is the C_{pk}-Index.

$$C_{pk} = min\left[\frac{USL - \hat{\mu}}{3 \cdot \hat{\sigma}}, \frac{\hat{\mu} - LSL}{3 \cdot \hat{\sigma}}\right]$$

with USL and LSL being the upper and lower specification limit, $\hat{\mu}$ the estimated mean of the production, and $\hat{\sigma}$ its estimated standard deviation. The standard way to predict process capability indexes of a process chain would be to collect data at the end of the chain and predict the mean and the standard deviation. The problem with this approach is that the predicted value actually does not represent the capability of the process chain at the time the data was raised. This can be illustrated by an example: One of the most common problems in process control is a change of variance in the process. Let us take another look at the process chain containing ten sub-processes described above and imagine a change of the production variance in the first sub-process. This change will not become visible in the process capability for as long as it takes the product to pass the whole process chain.

So we suggest to use the regression models trained to simulate the process chain to predict the mean and standard deviation of the product at the end of the chain. Because the capability index based on these estimations takes into account the state of all the sub-processes without any temporal delay we suggest the term *Instant Process Capability Index* for this method. This method is mostly intended for situations like the one described above when there is a change in one of the sub-processes. It could help to assess if such a change is crucial to the results of the whole process chain or not. This information could help making the decision to stop the production to adjust said sub-process immediately or to keep producing.

VII. Further Development

At this point, our research in the field of statistical process chain analysis is still in an early stage. For that reason, we would like to give a short outlook on some of the work we are planning to address ourselves or that we deem interesting for further research in general.

A. Application and Implementation

In association with the Center for Solar Energy and Hydrogen Research Baden-Württemberg (ZSW) we are given the possibility to apply the theoretical results we have presented in this paper to a real live process chain from the field of battery cell production. As part of a national research network with the goal to enhance the production of lithium-ion batteries, we hope not only to implement our own research but also to gain from the expertise of our partners.

As part of our research we plan to publish an R [8] software package containing methods for statistical analysis of process chains in 2017.

B. Method Development

The methods described in this paper have been limited to process chains that can be mapped by regression models. For the future it would make sense to extend our view to classification and time series models. The effect of Berkson case errors in the independent variables for logistic regression has already been discussed in [9].

C. Measurement Errors

As described in section IV the Berkson case error produced by linking processes is not the only kind of error in the variables. Uncertainty created by measurement errors is a common problem in many kinds of production processes. Since the effect of this kind of errors is similar to those described in this paper it makes sense to include them in the analysis of process chains.

VIII. Summary

In this paper, we have discussed how production chains consisting of several sequential sub-processes can be simulated using statistical learning methods.

The main obstacle in doing so is the fact that the single sub-processes influence each other by transferring characteristics. In the simulation of the process chain the values of these characteristics can only be determined with respect to prediction errors. In the situation we are facing, we are able to train models for every sub-process based on the true values of the transferred characteristics. However, if we want to predict results for the whole process chain we only have predicted values for the characteristics. This led us to the theory of Errors-in-Variables regression, or more precisely to the special case of Berkson type errors in the independent variables. This method deals with situations in which the observed values of an independent variable differ from their true values by an error term u (in our case the prediction error). We showed that using the models in the true variables and making predictions based on their observed values can lead to biased results and other problems. We further showed, that those models can be adjusted with respect to the variances of the errors in the independent variables to produce unbiased predictions and presented a way to predict said variances from the prediction variance of the original models. This does not only allow us to produce unbiased predictions over fractions of the process chain, but also gives important insight about the influence of single sub-processes on the result at the end of the process chain. Then we suggested to use the models that simulate the process chain to predict its process capability index. The main advantage of this method is, that the effect of a change in the quality of any sub-process on the process capability of the process chain can be predicted nearly without temporal delay. We finished with some ideas for future developments in the field of statistical analysis of process chains.

References

[1] H.-W. Zoch, “From single production step to entire process chain: the global approach of distortion engineering,” *Materialwissenschaft und Werkstofftechnik*, vol. 37, 2006.

[2] J. Berkson, “Are there two regressions?” *Journal of the American Statistical Association*, vol. 45, no. 250, pp. 164–180, 1950. [Online]. Available: http://www.jstor.org/stable/2280676

[3] W. Fuller, *Measurement Error Models.* Wiley-Interscience, 1987.

[4] G. E. P. Box, “The effects of errors in the factor levels and experimental design,” *Technometrics*, vol. 5, no. 2, pp. 247–262, 1963. [Online]. Available: http://www.jstor.org/stable/1266066

[5] L. Huwang and Y. H. S. Huang, “On errors-in-variables in polynomial regression-berkson case,” *Statistica Sinica*, vol. 10, no. 3, pp. 923–936, 2000. [Online]. Available: http://www.jstor.org/stable/24306755

[6] O. Meyer, J. Keller, and C. Weihs, “Uncertainty in process chains,” *Proceedings of the ESREL 2016*, 2016, accepted.

[7] Y. Wooluru, D. Swamy, and P. Nagesh, “The process capability analysis - a tool for process performance measures and metrics - a case study,” *International Journal for Quality Research*, vol. 8, no. 3, 2014.

[8] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: http://www.R-project.org/

[9] D. Burr-Doss, “The process capability analysis - a tool for process performance measures and metrics - a case study,” Division of Biostatistics, Stanford University, Tech. Rep. 104, 2014.