Chemometrics and Intelligent Laboratory Systems 149 (2015) 66–72

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Assumption free modeling and monitoring of batch processes☆

Frank Westad [1], Lars Gidskehaug [1], Brad Swarbrick [1], Geir Rune Flåten [1]

CAMO Software AS, Postboks 1662 Vika, 0120 Oslo, Norway

ARTICLE INFO

Article history:
Received 7 January 2015
Received in revised form 31 July 2015
Accepted 27 August 2015
Available online 5 September 2015

Keywords:
Batch processing
Batch modeling
Multivariate modeling
Batch analysis
Process analysis
Batch monitoring
Quality monitoring
Relative time modeling

ABSTRACT

Modeling strategies currently in use for the monitoring of batch processes where multivariate data are available have some limitations, particularly for batches where the true starting or end point are not the same on an absolute time scale, or the batch progression varies among batches. In this paper, a method capturing these differences and allowing modeling and monitoring of batches in relative time is proposed. Using scores from principal component analysis (PCA) models as a feature space the new methodology is better able to handle the challenges usually experienced in batch analysis. The feasibility of the relative time approach is demonstrated using data from a chemical synthesis and a pharmaceutical drying process.

1. Introduction

Batch processes are widely used in many industries. Typically, raw materials are combined in a suitable batch vessel before chemical, physical, or biological transformation takes place, resulting in an end product. In many cases, the control of the batch process is recipe driven, and the operations are not adjusted to accommodate raw material variation, changes in uncontrollable factors, and other changing circumstances. The best possible end product quality is achieved by adapting batch operations according to any detectable changes during processing, thus providing a control mechanism to drive a product toward its desired state. Optimal run settings and the ability to control them within a design space leads to reduced rework and rejects and improved end product quality, which has the major benefit of saving industry money and resources and more importantly, increased consumer trust in the product name.

Several solutions already exist for batch monitoring and control [1–3]. The starting point for all of them is a data set $\underline{\mathbf{X}}$ ($m \times n \times k$) describing normal operation conditions (NOC) of known batches of high quality, where n variables are measured at m sampling times for each of the k batches. The data are structured in two-dimensional matrices along time points $m \times (n^*k)$ or variables $(m^*k) \times n$ unless a direct three-way modeling approach such as PARAFAC [4] is applied. A model describing NOC is developed and new batches can be projected on the model to provide real-time quality information.

The existing batch modeling approaches assume equal lengths of batches, i.e. the batch is expected to start at the same chemical or biological time t_0 and has the same number of time points for all batches. This leads to problems during model building if the NOC data set is uneven, and ultimately during monitoring if new batches do not meet these criteria. Numerous approaches to handle uneven batch lengths exist, including replacing time with a maturity index [2], dynamic time warping (DTW) [5], time linear expanding/compressing [6] etc. Complications can occur in all of these methods if the first measurement does not coincide with the true t_0, i.e. the new batch(es) does not start at the same chemical/biological state. The PARAFAC approach models the data as a true three-way model which has a possible advantage that the time is modeled as a separate dimension and not connected to either samples or variables as in the unfolding case. However, the challenges with unequal batch length and chemical time still need to be addressed. Also, the monitoring phase requires dynamic recalculating of models up to the current point of time [4].

Another challenge experienced in batch monitoring relates to batches where there are phase changes, i.e. the underlying dynamics in the batch changes. This can occur in processes such as multiple stage chemical reactions where a reactant is added during operation to initiate a subsequent reaction. An equivalent situation may occur in physical processes where the underlying mechanism changes, such as in a pharmaceutical wet mass drying process where there is a state change in the material when all free water is evaporated and bound

☆ Selected papers presented at the 3rd European Conference on Process Analytics and Control Technology, 6–9 May 2014, Barcelona, Spain.

E-mail address: grf@camo.com (G.R. Flåten).
[1] Tel.: +47 22 39 63 00; fax +47 22 39 63 22.

http://dx.doi.org/10.1016/j.chemolab.2015.08.022
0169-7439/

化学计量学与智能实验室系统 149 (2015) 66–72

内容列表可在ScienceDirect获取

化学计量学与智能实验室系统

期刊主页：www.elsevier.com/locate/chemolab

无预设建模与监控批量过程☆

Frank Westad [1], Lars Gidskehaug [1], Brad Swarbrick [1], Geir Rune Flåten [1]

CAMO Software AS, Postboks 1662 Vika, 0120 Oslo, Norway

摘要 文章信息

文章历史：初稿收到日期：2015年1月7日
修订稿收到日期：2015年7月31日 接受日期：2015年8月27日
在线发布日期：2015年9月5日

关键词：批量处理 批量建模 多元建模 批量分析 过程分析 批量监控 质量监控 相对时间建模

目前用于监控具有多元数据的批量过程的建模策略存在一些局限性，特别是对于那些在绝对时间尺度上起始点或结束点不一致或各批次进程不同的情况。本文提出了一种能够捕捉这些差异并在相对时间内进行建模和监控的方法。利用主成分分析（PCA）模型的得分作为特征空间，新的方法能够更好地处理批量分析中通常遇到的挑战。使用来自化学合成和制药干燥过程的数据证明了相对时间方法的可行性。

1. 引言

批量过程广泛应用于许多工业领域。通常，原材料在一个合适的批处理容器中混合后，再进行化学、物理或生物转化，从而得到最终产品。在很多情况下，批处理过程的控制是由配方驱动的，操作不会调整以适应原材料变化、不可控因素的变化和其他不断变化的情况。通过根据过程中可检测到的变化调整批处理操作，可以获得最佳的最终产品质量，从而提供一种控制机制，使产品向其理想状态发展。最优运行设置及其在设计空间内的可控性有助于减少返工和废品，提高最终产品质量，这主要带来了节省工业资金和资源的重大好处，更重要的是增强了消费者对产品品牌的信任。

针对批次监控和控制已经存在多种解决方案[1–3]。它们的起点都是一个数据集X（m ×n ×k），用来描述高质量已知批次的正常操作条件（NOC），其中对于k个批次中的每一个，都在m个采样时间点测量n个变量。除非采用三维建模方法，否则数据通常按二维矩阵结构排列，即时间点×(变量*k) 或 (时间点*k)×变量。

应用了诸如PARAFAC[4]这样的直接三维建模方法。建立了描述NOC的模型，新批次可以投影到该模型上以提供实时质量信息。

现有的批次建模方法假设所有批次长度相等，即批次预期在相同的化学或生物时间t0开始，并且所有批次具有相同数量的时间点。如果NOC数据集不均匀，则会导致建模过程中出现问题，最终在监控新批次不符合这些标准时也会出现问题。处理不等长批次的方法有很多，包括用成熟度指数[2]替代时间、动态时间规整（DTW）[5]、时间线性扩展/压缩[6]等。但如果第一次测量不与真实t0重合，即新批次不在相同的化学/生物状态下开始，则这些方法中的每一种都可能出现复杂情况。PARAFAC方法将数据建模为真正的三维模型，这可能具有的优势在于时间被建模为一个单独的维度，而不是像展开情况下那样与样本或变量相连。然而，仍需解决不等长批次和化学时间的挑战。此外，在监控阶段需要根据当前时间点动态重新计算模型[4]。

批次监控中遇到的另一个挑战涉及存在相变的批次，即批次中的底层动力学发生变化。这可能出现在多阶段化学反应等过程中，其中在操作期间加入反应物以启动后续反应。类似情况也可能出现在物理过程中，当底层机制发生变化时，例如在制药湿料干燥过程中，当所有自由水蒸发完毕并且结合水状态发生变化时。

☆选登于2014年5月6日至9日在西班牙巴塞罗那举行的第三届欧洲过程分析与控制技术会议论文。

电子邮件地址：grf@camo.com (G.R. Flåten)。1
电话：+47 22 39 63 00；传真 +47 22 39 63 22。

http://dx.doi.org/10.1016/j.chemolab.2015.08.022
0169-7439/

water of crystallization starts evaporating. There are approaches for handling such challenges, e.g. multiple local models [5], which can work fine if the phase change occurs at fixed times or a trigger can be used to switch models.

When monitoring batches with phase transitions, varying residence times or reaction rates within the phases may occur. This is often seen in biological batch processes, where bacteria metabolize reactants into products. The modeling solutions applied to the varying residence time are often the same as those used for phase changes, such as multiple local models. However, the uneven progression among batches means that visualizing the trajectory for a batch compared to NOC is not very meaningful, and the presentation is not representative of the underlying batch dynamics. Plotting scores from a multivariate model versus a time axis imposes a one-dimensional assumption on the batch modeling which may not necessarily be correct in relation to the chemical/biological state of the material.

During monitoring of new batches, another challenge experienced by models for data unfolded along the batch dimension is the handling of missing values [7]. At any time during a batch, only the current and previous measurements are available. Measurements at future time points of the batch are missing, and any missing values cannot be directly compared with the established model for NOC. The typical solution for this is to use what is known as lagged multivariate models (LMV), which in effect is a number of models for the different time points [8]. However, this seems to be a work-around solution, although the computational time is often no issue with today's computers for such a complex modeling strategy.

All the above-mentioned approaches effectively try to capture the batch trajectories as described in the multivariate space, making the assumption that time is an attribute of the trajectory. However, [absolute] time is not a necessary attribute of the trajectory [9]. Nevertheless, any process is of course operated along the time dimension. A good modeling approach requires a flexible but true synchronization between trajectory [process state] and time [from a chemical/biological state]. This is a similar approach to the desired state as defined in the quality by design (QbD) paradigm, i.e. processes should not be run for a particular defined time interval, they should be run until the product reaches its desired state (provided the time to reach the end point is not excessive). Thus, a time-independent approach to batch modeling also meets the criteria of the QbD approach.

In this paper, an improved batch modeling approach accommodating uneven batch lengths, unknown true t_0, phase changes, and uneven residence times is proposed. This is achieved by a true multivariate, feature-based approach that does not make any assumptions about the synchronization and duration of batches. Instead, the so-called relative time is estimated by the method itself. Relative time is here used in a broad sense for any transient process including non-linear behavior, and it is often found to correspond with the underlying chemical, biological, or physical changes during the process. The presented method is analogous to existing methods [1–4] in that a calibration set of batches with relevant measurements are used to establish a model representative of NOC. The $m \times n \times k$ calibration set $\underline{\mathbf{X}}$ has data for k batches with m samples in each batch for n variables (where the value of m is not necessarily constant across batches).

The calibration set $\underline{\mathbf{X}}$ is unfolded so that all the observations for each batch are represented as rows and the unfolded data matrix thus has the dimension $(k^*m) \times n$. A principal component analysis (PCA) model for the unfolded data set is calculated and validated. In the resulting score space, a grid is optimized to capture the features of the batch trajectories. A new PCA model is developed based on the samples within the feature grid only, and true trajectories, relative time, and relevant deviations are calculated. It is important at this stage to ensure that the process signature of the trajectories is consistent between batches in order to develop a robust and representative grid. This is highly important and it has been noted as a major flaw in a straight mathematical fitting of data, i.e. if the process trajectories are not visibly overlaying to a high degree, then the process itself is not consistent, therefore a batch model is not representative of tight manufacturing controls.

Monitoring of new batches is done by projection onto the feature grid, which is now based on sound scientific modeling. The feature grid describes NOCs and since it is feature based, any time shifts, phase changes, or rate variations are taken into account by the model. Deviations from the NOC are identified using established diagnostics tools, including a modified Hotelling's T^2 chart that is dynamic over the process trajectory and also a dynamic limit for the F-residual distance to model.

The theory section outlines the relative time approach and the proposed diagnostics plots are discussed in further detail. In the subsequent section, two short examples where the method is applied are shown. The discussion section suggests where the proposed method will be suitable as well as providing a review of the underlying assumptions.

2. Theory

Data collected over time for a number of batches and variables can be represented as a three-way data structure, $\underline{\mathbf{X}}$, of dimensions $(m \times n \times k)$. The indices correspond to m time points for which the n variables are measured for k batches. However, due to various batch lengths and other uncontrollable variations, the sample number within the individual batches may not pertain to the same relative state of the process in terms of the underlying chemistry between batches. Therefore, a more appropriate term for the "time" dimension is "sample number." An example of this is the case where the sampling rate is varied between the batches and thus the sample number does not reflect the batches' development over time, even if the relative states were the same. The situation where the sample number does not reflect the same state is more a rule than an exception for chemical and biological systems.

Some scenarios in addition to the above-mentioned sampling-rate situation are:

a. The batches do not start in the same state, e.g. the material processed has various moisture content at the start of the process.
b. The batches' final sampling points do not reflect the same state.
c. The batch progression is not the same and is in most processes non-linear, i.e. the batch trajectory does not evolve in equidistance steps in relative time.

The solution for handling all the above cases without extensively distorting the relative time by preprocessing is to model the batch trajectory from a common start to a common end state in chemical/biological time.

The approach proposed here comprises the following steps as presented below in pseudo-code:

1. Unfold the $(m \times n \times k)$ data matrix $\underline{\mathbf{X}}$ to the $(k^*m) \times n$ data matrix.
2. Preprocess, center, and scale the data as required.
3. Perform PCA for all batches in one model while cross-validating across batch $\mathbf{X} = \mathbf{t}_a\mathbf{p}_a^t + \mathbf{E}$, a $= 1..A_{optimal}$.
4. Use a grid-search algorithm in the multivariate (score) space to find the common relative start and end points in so-called "grid elements." Various grid resolutions are applied to model the trajectory in the best way. The current criterion is to select the resolution that gives the most grid elements. An alternative would be the highest percentage of grids from the total number of grids. In principle, the grid search can be extended to 3D ("grid cubes") and higher dimensions.
5. Calculate the mean for all samples and the means for individual batches for these grid elements. If not all samples are found inside the grids, recalculate the PCA model.
6. Interpolate to a common batch trajectory for the overall mean with the desired resolution, e.g. in percent of relative time. Linear or

结晶水开始蒸发。有一些方法可以应对这种挑战，例如多个局部模型[5]，只要相变发生在固定时间点或可以使用触发器切换模型，这种方法就能很好地工作。

在监控存在相变的批次时，可能会出现相内停留时间或反应速率的变化。这在生物批次过程中常见，例如细菌将反应物代谢为产物的情况。应用于不同停留时间的建模方案通常与用于相变的相同，如多个局部模型。然而，批次之间的进展不均导致将某个批次的轨迹与NOC进行可视化对比并不十分有意义，且这种呈现方式并不能代表底层的批次动态。将多变量模型的得分与时间轴绘制在一起强加了一维假设于批次建模之上，这可能并不一定与材料的化学/生物状态相符。

在监控新批次时，另一个由沿批次维度展开的数据模型遇到的挑战是缺失值的处理[7]。在批次的任何时间点，只有当前和之前的测量值可用。批次未来时间点的测量值缺失，且这些缺失值无法直接与NOC的已建立模型进行比较。对此的典型解决方案是使用所谓的滞后多变量模型（LMV），实际上是对不同时间点建立多个模型[8]。然而，这似乎是一种权宜之计，尽管就复杂的建模策略而言，如今计算机的计算时间通常不成问题。

上述所有方法都试图有效地在多变量空间中描述批次轨迹，并假设时间是轨迹的一个属性。然而，[绝对]时间并不是轨迹的必要属性[9]。尽管如此，任何过程当然都是沿着时间维度进行的。一个好的建模方法需要在轨迹[过程状态]和时间[从化学/生物状态角度]之间灵活但真实的同步。这与质量源于设计（QbD）范式中定义的理想状态方法类似，即不应按照特定定义的时间间隔运行过程，而应运行到产品达到其期望状态为止（前提是到达终点所需时间不过长）。因此，独立于时间的批次建模方法也符合QbD方法的标准。

本文提出了一种改进的批次建模方法，能够处理不等长批次、未知的真实t0、相位变化和不一致停留时间。这是通过一种真正的多变量、基于特征的方法实现的，该方法不对批次同步性和持续时间做任何假设。相反，所谓的相对时间由方法本身估计。这里的相对时间广义地用于任何瞬态过程，包括非线性行为，通常发现它与过程中发生的潜在化学、生物或物理变化相对应。所提出的方法类似于现有方法[1–4]，因为它们都是利用包含相关测量值的批次校准集来建立代表NOC的模型。m ×n ×k的校准集X包含k个批次的数据，每个批次有m个样本，对应n个变量（其中m的值在各批次间不一定恒定）。

校准集X被展开，使得每个批次的所有观测值都表示为行，因此展开后的数据矩阵具有维度(k*m) ×n。计算并验证展开数据集的主成分分析（PCA）模型。在所得得分空间中，优化网格以捕捉批次轨迹的特征。仅基于特征网格内的样本开发新的PCA模型，并计算真实轨迹、相对时间和相关偏差。在此阶段确保批次轨迹的过程特征在不同批次之间保持一致非常重要，以开发稳健且有代表性的网格。这一点至关重要，已有研究指出，直接进行数学拟合存在重大缺陷，即如果过程轨迹在视觉上没有高度重叠的话。

如果一致性程度不足，则过程本身不一致，因此批次模型不能代表严格的生产控制。

新批次的监控是通过投影到特征网格上完成的，而此网格现在基于科学建模。特征网格描述了正常操作条件（NOC），由于它是基于特征的，因此模型会考虑任何时间偏移、相位变化或速率变化。使用已建立的诊断工具识别与NOC的偏差，包括一个动态Hotelling T2控制图，它随过程轨迹变化，以及F残差距离模型的动态限值。

理论部分概述了相对时间方法，并进一步详细讨论了所提出的诊断图。在后续章节中，展示了该方法应用的两个简短示例。讨论部分建议了所提出方法适用的场景，并回顾了其基本假设。

2. 理论

对于一定数量的批次和变量随时间收集的数据可以表示为一个三维数据结构X，其维度为(m ×n ×k)。索引对应于m个时间点，在这些时间点上测量n个变量，适用于k个批次。然而，由于各个批次的长度不同以及其他不可控的变化，单个批次内的样本数量可能不对应于各批次间潜在化学反应意义上的相同相对状态。因此，更合适的术语是样本编号。例如，当批次间的采样率变化时，即使相对状态相同，样本编号也不能反映批次随时间的发展。对于化学和生物系统来说，样本编号不能反映相同状态的情况更为常见。

除了上述采样率情况之外，还有一些其他场景：

a. 批次不在相同的状态下开始，例如加工材料在过程开始时具有不同的含水量。
b. 批次的最终采样点不反映相同的状态。c. 批次进展并不一致，在大多数过程中都是
非线性，即批次轨迹在相对时间下并不是等距发展的。

解决上述所有情况而不通过预处理严重扭曲相对时间的方法是从化学/生物时间中的公共起点到公共终点建模批次轨迹。

本文提出的解决方案包括以下步骤，如下伪代码所示：

1. 将(m ×n ×k)数据矩阵X展开为(k*m) ×n数据矩阵。
2. 根据需要对数据进行预处理、中心化和缩放。
3. 对所有批次的数据进行一次PCA建模，同时跨批次进行交叉验证 X = tapat + E，a = 1..Aoptimal。
4. 在多变量（得分）空间中使用网格搜索算法，找到所谓的网格单元中的公共相对起点和终点。应用各种网格分辨率以最佳方式建模轨迹。当前标准是选择产生最多网格单元的分辨率。另一种可能是从全部网格中选取最高比例的网格。原则上，网格搜索可以扩展到三维（网格立方体）及更高维度。
5. 计算所有样本的均值以及这些网格单元内各批次的均值。如果并非所有样本都在网格内，则重新计算PCA模型。
6. 插值到整体平均值的通用批次轨迹，达到所需的分辨率，例如相对时间的百分比。根据需要可采用线性或

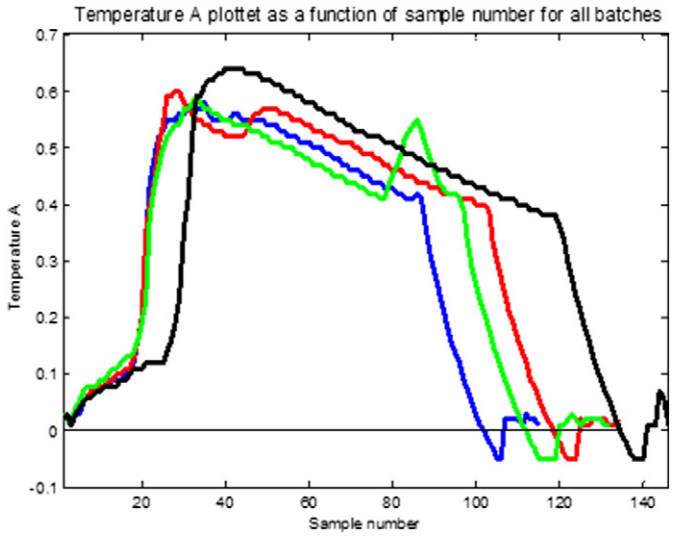

Fig. 1. Variable (Temperature A) plotted vs. sample number.

spline interpolation may be applied depending on the nature of the process.

7. Project the mean for the individual batches onto this trajectory and estimate the relative time, distance within the model space, as well as the residual distance. The projection is orthogonal to the line between two adjacent points on the trajectory and gives the relative time as the ratio of actual point and the total number of points on the trajectory. For simplicity, it is scaled to 0–100.
8. From these projected distances estimate the standard deviation around the common trajectory. The actual points to be plotted as lower or upper limits are always following the direction of the trajectory. This is to ensure that the limits are not crossing when plotting connection lines as shown in Figs. 5, 7, and 10 below.
9. The residual distance for each grid element is also estimated and is thereafter interpolated to yield a residual distance limit for each relative time. The residuals for individual objects are actually the same as the normal residuals in the PCA model but are plotted for the relative time and its actual limit.
10. The individual samples' scores are projected onto the trajectory for estimation of the relative time, distance to the trajectory, and distance to the model.

An alternative would be to perform PCA on the individual batches with individual centering and scaling. However, due to rotational ambiguity, this would require reflection and/or rotation of the models to a common basis before estimating the average trajectory. Also, the objective is mainly to model the differences between the batches and therefore centering and scaling are performed column-wise for the variables. If the process changes character for the whole duration, it might be split into phases to account for various correlation structures. See also section Discussion.

The model is a representation of the dynamics of the batch, in terms of the mean trajectory, so the individual batch mean values are projected onto the nearest point on the trajectory. The average of these distances is defined as the "distance to the model trajectory" and can be displayed at various significance levels. The orthogonal distance to the trajectory model for the objects for a new batch can be compared to the distance limit for monitoring purposes. Given that $t_{new} = x_{new}P$, the distance can be expressed as

$$D_{Trajectory} = \sqrt{\sum_{a=1}^{Aopt}\left(t_{new,a} - t_{new,a\perp}t_{Trajectory,a}\right)^2} \quad (1)$$

Where

$D_{Trajectory}$ is the orthogonal distance from the new score to its projected position on the trajectory
t_{new} is the new score
$t_{new} \perp t_{Trajectory}$ is the projected position on the trajectory

The monitoring phase involves the following steps:

1. Preprocess, center, and scale the new observations
2. Estimate the new scores as $t_{new,a} = \boldsymbol{x}_{new}\boldsymbol{p}_a$ for components 1:$A_{optimal}$
3. Project these scores onto the trajectory for estimation of the relative time, distance to the trajectory, and distance to the model.

Note that this approach estimates batch progression or maturity directly in terms of the relative time.

3. Experimental and results

Two examples will be shown to illustrate the challenges presented with time-dependent processes, both in the modeling phase as well as for monitoring. The examples concern two unit operations from the chemical and pharmaceutical production environments, one with a small number of process variables and one with instrumental data from NIR spectroscopy.

3.1. Example 1: A chemical reaction

The data were taken from a chemical synthesis which, due to a confidentiality agreement, could not be described in detail. The variables were temperature measured at two positions (A and B) and pressure. This example, although only with three variables, illustrates very well the challenges in batch modeling.

For simplicity, the samples were pre-screened to describe the reaction itself and not the initial phase in the synthesis. Nevertheless, although the samples have the same relative starting point, the development over time is quite different. Fig. 1 shows the Temperature A variable as a function of the sample number. The immediate interpretation is that the batches are different, but the 2D score plot from PCA calculated for the three variables (Fig. 2) reveals that the batches follow the same trajectory although not with the same relative rate. Batch 4 has more observations as seen in Fig. 1, which gives the effect of a very condensed cloud of points on the left-hand side in Fig. 2. Batch 3 reveals that the pressure is increasing and thereafter decreasing, yielding the reversed evolution around score value 1 (PC1). The corresponding correlation loading plot [10] revealed that all variables

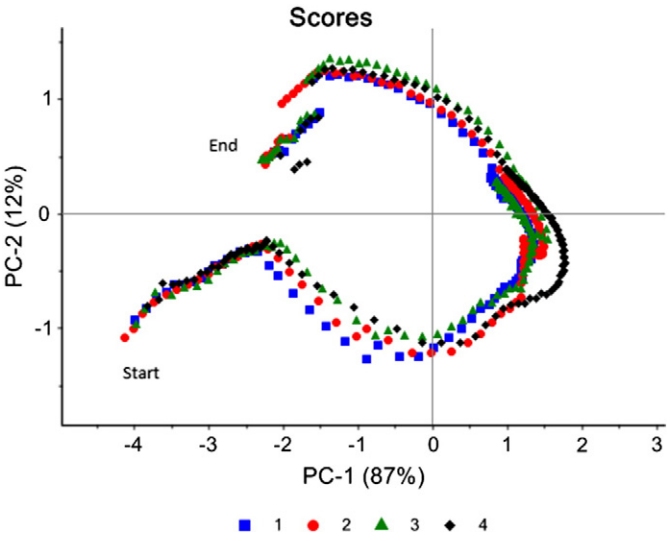

Fig. 2. Score plot of PC1 vs. PC2 with batches shown as different symbols.

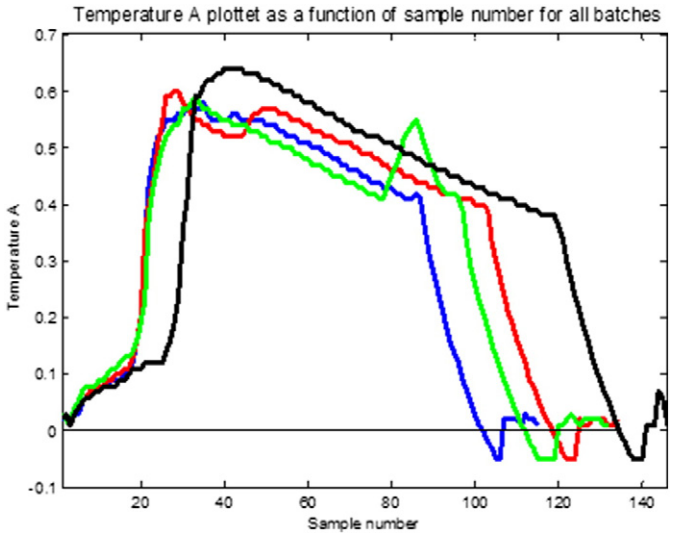

图1. 变量（温度A）与样本数量的关系图。

根据工艺特性，可能应用样条插值。

7. 将各个批次的均值投影到此轨迹上，并估计相对时间、模型空间内的距离以及残差距离。投影垂直于轨迹上两个相邻点之间的连线，并将相对时间表示为实际点与轨迹上总点数的比例。为了简便起见，将其缩放至0-100。

8. 根据这些投影距离估计围绕公共轨迹的标准偏差。要绘制的实际点始终沿着轨迹的方向，这是为了确保在绘制连接线时不出现交叉，如下图5、7和10所示。

9. 每个网格单元的残差距离也被估计，并通过插值得到每个相对时间的残差距离限值。个体对象的残差实际上与PCA模型中的正常残差相同，但在相对时间和其实际限值下进行绘图。

10. 个体样本的得分被投影到轨迹上以估计相对时间、到轨迹的距离以及到模型的距离。

另一种方法是对各个批次分别进行中心化和缩放后的PCA分析。然而，由于旋转模糊性，这需要将模型反射和/或旋转到一个共同基础上，然后再估计平均轨迹。此外，主要目标主要是建模批次之间的差异，因此对变量执行列方向的中心化和缩放。如果整个过程持续时间内发生了性质变化，可能会将其划分为多个阶段以考虑不同的相关结构。另请参见讨论部分。

该模型代表了批次动态的均值轨迹。个体批次的均值被投影到轨迹上的最近点。这些距离的平均值定义为到模型轨迹的距离，并可在不同显著性水平上显示。新批次对象相对于轨迹模型的正交距离可以与监控目的的距离限值进行比较。假设 tnew = xnewP，那么距离可以表示为

$$\text{DTrajectory} = \sqrt{\sum_{a=1}^{\text{XAopt}} \left(t_{new,a} - t_{new,a\perp t_{Trajectory,a}}\right)^2} \quad (1)$$

其中

DTrajectory 是新得分与其在轨迹上的投影位置之间的正交距离

tnew 是新的得分
tnew ⊥tTrajectory 是在轨迹上的投影位置 跟踪阶段包括
以下步骤：

1. 对新观测值进行预处理、中心化和缩放
2. 将新得分估计为 tnew,a = xnewpa ，其中a取值为1到Aoptimal
3. 将这些得分投影到轨迹上以估计相对时间、到轨迹的距离以及到模型的距离。

请注意，这种方法直接以相对时间的形式估算批次进展或成熟度。

3. 实验与结果

将展示两个示例来说明时间依赖性过程中所面临的挑战，不仅在建模阶段，还包括监测方面。这些示例涉及化学和制药生产环境中的两个单元操作，一个包含少量过程变量，另一个则包含来自NIR光谱的仪器数据。

3.1. 示例1：化学反应

数据来自一项化学合成过程，由于保密协议的原因，无法详细描述。变量包括两个位置（A和B）测得的温度和压力。尽管这个例子仅有三个变量，但它很好地说明了批次建模中的挑战。

为了简化，对样本进行了预筛选，以描述反应本身而非合成初始阶段的情况。尽管如此，虽然样本具有相同的相对起点，但随时间的发展却大不相同。图1显示了样本数量的函数下的温度A变量。直观的解读是这些批次不同，但从三个变量计算出的PCA二维得分图（图2）揭示了这些批次遵循相同的轨迹，尽管它们的相对速率并不相同。如图1所示，第4批次有更多的观测值，这导致在图2左侧形成了非常密集的点云。第3批次显示压力先是增加然后减少，在得分值1（PC1）附近出现了反向演化。相应的相关载荷图[10]显示了所有变量

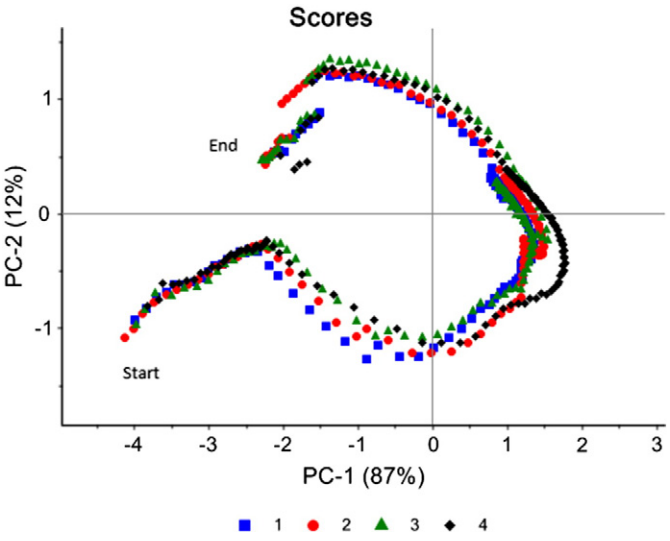

图2. PC1与PC2的得分图，不同批次以不同符号表示。

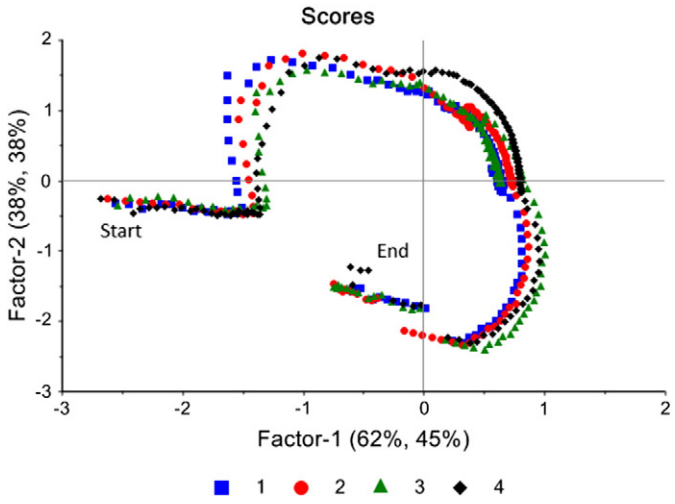

Fig. 3. Score plot from the PLS model for maturity with samples representing common start and end points.

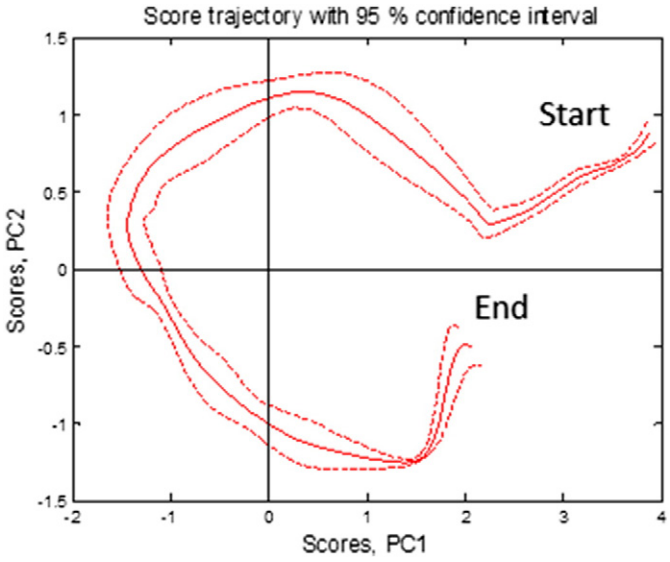

Fig. 5. The estimated trajectory with a 95% confidence interval.

contribute with positive values along PC1 (87% explained variance). Although the second PC explains only 12% of the variance, this is still systematic variation, thus the optimal model rank is two.

One way to model the batch data is to make use of a so-called maturity index where a response variable is constructed from the sample number for the various batches. Each sample is given a maturity index based on the length of individual batches and used as the response variable. However, this approach for modeling historical and predicting new batches does not solve the problem of non-linear behavior or differences in the starting point in relative time. As an illustration of this, a PLS regression model with a maturity index was calculated. A subset of the samples was selected so that they represent the same start and end point in chemical sense, as seen in the score plot (Fig. 3). Since the number of samples was not the same for all batches (from 147 to 158 observations), a relative maturity index from 1 to 100 was calculated as the average step based on the number of observations for each batch. Nevertheless, because of the highly non-linear behavior in the score space, this approach gave a poor explained variance for maturity as the response variable.

The data were then modeled with the new relative time procedure. Fig. 4 shows how the algorithm finds the grids as the basis for the estimation of the batch trajectory. The score space has been split in sections for PC1 and PC2, giving grids as depicted as rectangles. All samples found inside the individual grid elements serve as the basis for estimation of the mean values and the standard deviation. Interpolation is performed with the mean values of the scores inside the grids to produce a continuous trajectory with the desired resolution, e.g 100 points along the trajectory. Depending on the application, a linear or non-linear interpolation must be defined, Fig. 5. Note that the reaction is highly non-linear in terms of sample number versus relative time as e.g. the congested region in the left part of the plot shows. In fact, it seems that the reaction for a short period of time is actually reversed for batch 3 (in green).

Once the batch trajectory has been determined, all observations for the historical data can be represented in relative time irrespectively of the number of observations, thus avoiding the need to regulate sample frequency.

Fig. 6 shows the variable Temperature A in relative time as estimated from the score trajectory in the 2-dimensional model. It shows that the evolution of the batches now corresponds to the chemical time similar to the 2D score plot when plotted as a 1D control chart.

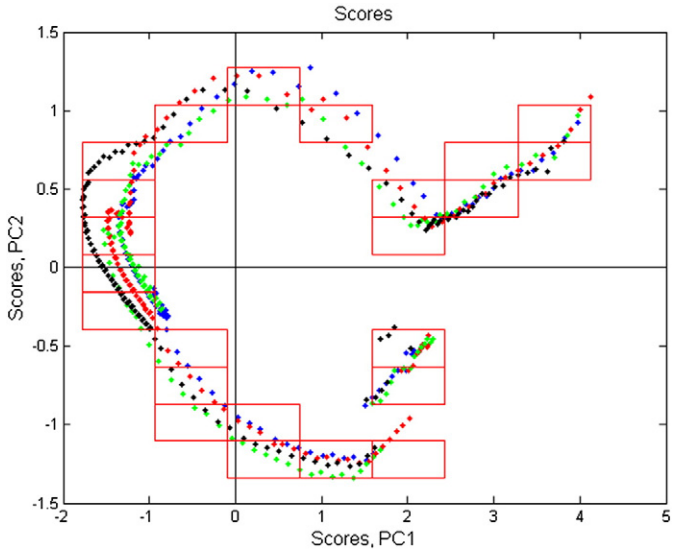

Fig. 4. Score plot for PC1 vs. PC2 with the grids depicted.

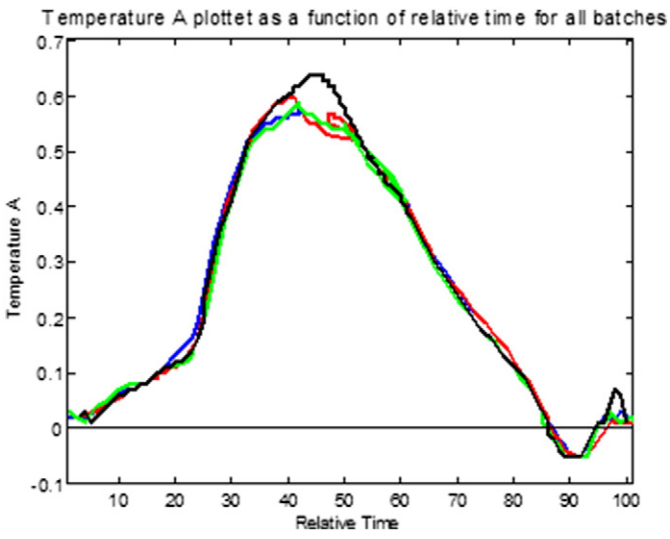

Fig. 6. The variable Temperature A in relative time.

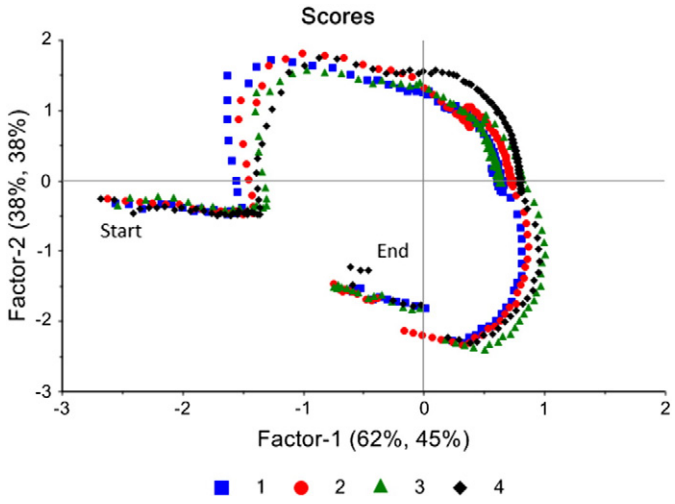

图3. 来自PLS成熟度模型的得分图，其中样本代表共同的起始点和终点。

在PC1上贡献正值（解释方差为87%）。尽管第二个主成分仅解释了12%的方差，但这仍然是系统性的变化，因此最优模型秩为2。

一种对批次数据建模的方法是使用所谓的成熟度指数，其中响应变量由不同批次的样本编号构建。每个样本根据各自批次的长度赋予一个成熟度指数，并用作响应变量。然而，这种方法在建模历史数据和预测新批次时并不能解决非线性行为或相对时间起点差异的问题。为了说明这一点，我们计算了一个带有成熟度指数的PLS回归模型。选择了一部分样本，使其在化学意义上代表相同的起始点和终点，如得分图（图3）所示。由于并非所有批次的样本数量都相同（从147到158个观测值），因此基于每个批次的观测数计算了一个从1到100的相对成熟度指数作为平均步长。然而，由于得分空间中的高度非线性行为，这种方法对于成熟度作为响应变量的解释方差较差。

随后使用新的相对时间程序对数据进行了建模。图4显示了算法如何找到网格作为基础

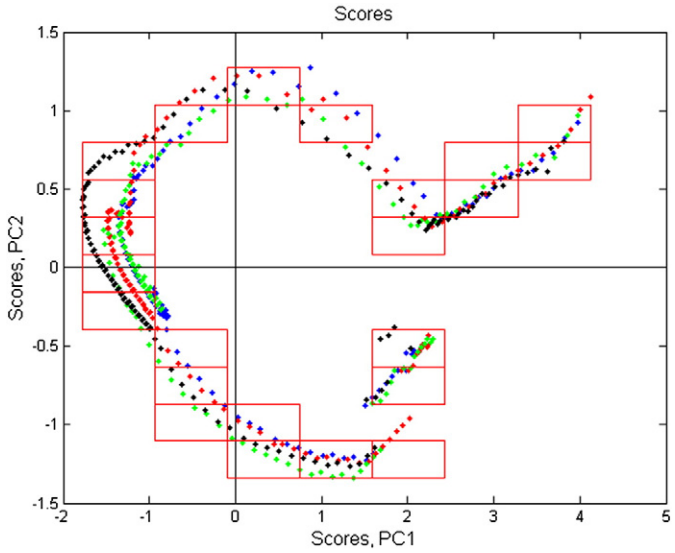

图4。PC1与PC2得分图，包含网格线。

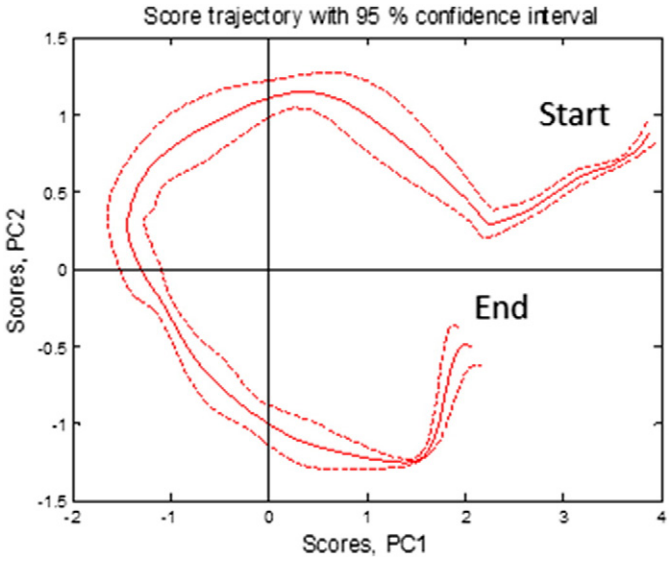

图5. 带有95%置信区间的估计轨迹。

用于估计批次轨迹。PC1和PC2的得分空间被划分为若干区域，形成如矩形所示的网格。位于各个网格单元内的所有样本作为估计平均值和标准偏差的基础。通过对网格内得分的平均值进行插值，生成具有所需分辨率（例如轨迹上的100个点）的连续轨迹。根据应用需求，必须定义线性或非线性插值方法，见图5。请注意，就样本数量与相对时间的关系而言，反应具有高度非线性特征，如图中左侧密集区域所示。事实上，看起来第3批（绿色）的反应在短时间内实际上是逆转的。

一旦确定了批次轨迹，无论历史数据中的观测数如何，都可以以相对时间表示所有观测，从而避免了调节采样频率的需要。

图6显示了在2维模型中根据得分轨迹估计的相对时间下的变量温度A。它表明，现在批次的演变对应于化学时间，类似于绘制为1D控制图时的2D得分图。

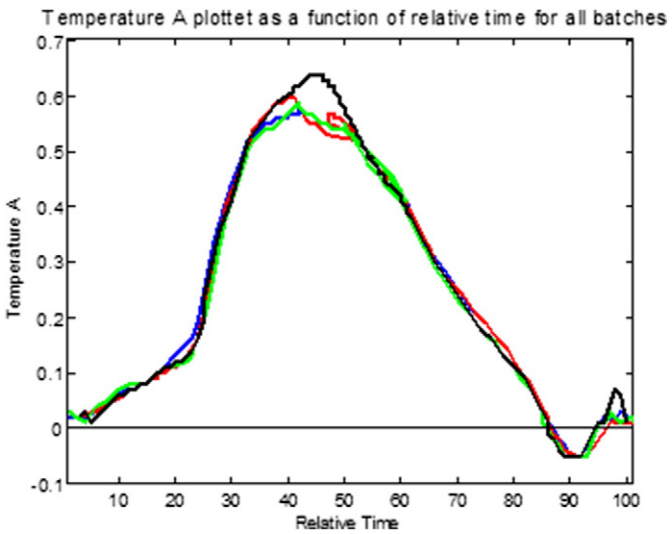

图6：相对时间中的温度A变量。

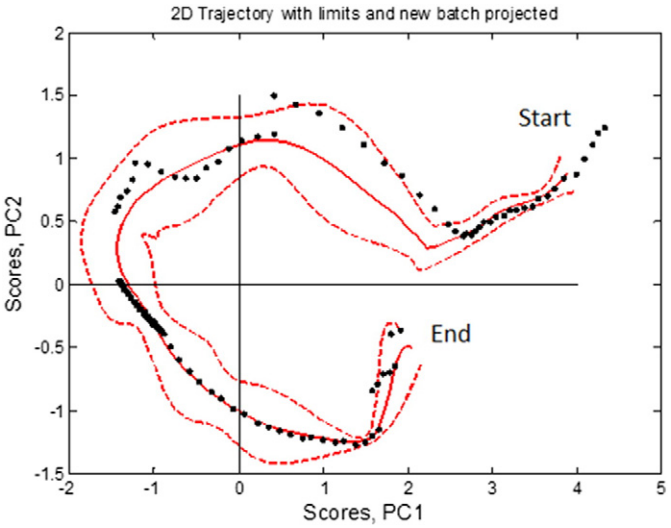

Fig. 7. Projection of a new batch on the chemical reaction batch trajectory model.

Projection of new batches is now straightforward in that new scores are estimated from the loadings and then projected onto the established batch trajectory. Fig. 7 shows how a new batch evolves over time (from right to left). Even if the batch starts outside the trajectory starting point because the chemical state is different, i.e. this batch has not progressed to the common starting point, this poses no problem in terms of visualizing in relative time. This is also the case in one dimension, the only impact is that the new observations may start in negative relative time. This is further discussed below for the fluid bed dryer data of example 2.

3.2. Example 2: Near-infrared (NIR) spectra of a fluid bed dryer operation

NIR spectra were collected during the fluid bed drying of a granular intermediate in a pharmaceutical process. The spectra consist of a total of 1093 wavelengths in the range 9090–4484 cm^{-1} (1100–2230 nm) at 4 cm^{-1} resolution. Due to the noisy nature of Fourier transform (FT) NIR (particularly at such high resolution), the spectra were smoothed with an 11-point moving average before applying a standard normal variate (SNV) to reduce the systematic baseline effects in the spectra. The spectra after SNV are shown in Fig. 8. A PCA model was calculated, showing that the two first PCs accounted for 97 and 1% of the variance, respectively. The validation variance indicated two PCs to be optimal which was confirmed by visual inspection of the 2D score plot.

Fig. 9 shows the historical batches with the grid elements found, and Fig. 10 shows the estimated trajectory with a 95% confidence limit and a new batch projected onto the model (black dots). When the batches are plotted as "sample number" in a 1-dimensional score plot (Fig 11) the initial interpretation is that

1. The historical batches are quite different.
2. The new batch is different from the historical batches.

However, the 2-dimensional score plots in Figs. 9 and 10 reveal that the batches do follow a common trajectory in relative time. Accordingly a line plot of the scores vs sample number gives an incorrect representation of how the batches evolve as the starting point for monitoring of the new batch cannot be assumed to represent the common starting point t_0.

The 2-dimensional score plot in Fig. 9 shows that the batches start in different positions because the starting material has various levels of moisture. This means that the common starting point is found where all batches follow the same trajectory, and the end point granulate was found by confirmation against a loss on drying reference analysis. Thus, the process was stopped when the moisture content presumably was at the correct level. Fig. 9 indicates that some batches were "overdried" as the end points in the relative time differ to some extent.

The scores for the new batch at the start of the process (lower left in Fig. 10) hardly change for the first 23 samples, but this has no impact on the progress in relative time. On the other hand, plotting these scores as sample number in Fig. 11 gives the impression that the new batch right from the start is out of control. The same is the case for the third historical batch (in green),

Fig. 10 shows the new batch starts outside of the common trajectory due to higher moisture than the common starting point for the historical batches. This situation can already be seen in Fig. 9, where some batches have their first observations in the lower left corner. However, as the granulate is dried the new batch follows the trajectory. Also notice the gap from one observation to the next as the batch enters the common trajectory. The reason for this may be heterogeneity of the granulate. Nevertheless, this does not pose any problem in monitoring the progress of the process.

4. Discussion

The results from the two examples presented above show that the new approach is able to model the batch progression in relative time,

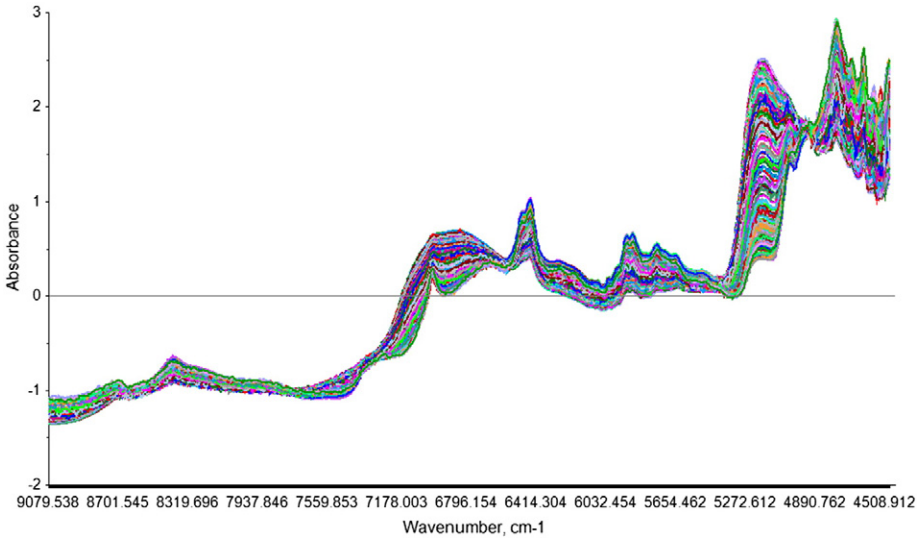

Fig. 8. NIR spectra after SNV.

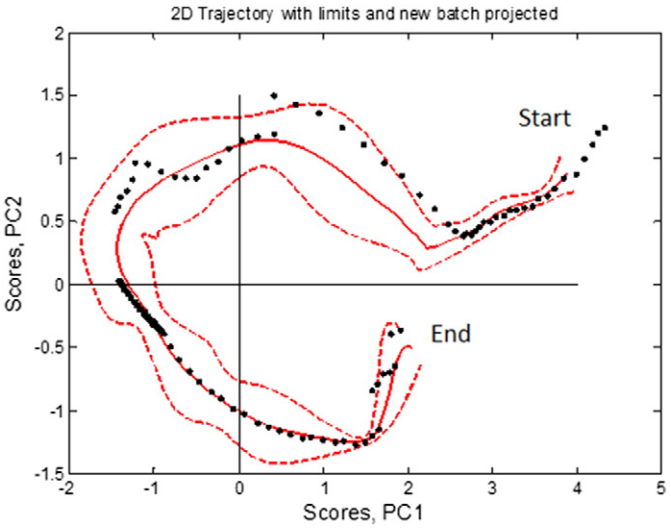

图7：新批次在化学反应批次轨迹模型上的投影。

新批次的投影现在变得非常简单，新的得分是从载荷中估算出来的，然后投影到已建立的批次轨迹上。图7显示了一个新批次随时间推移（从右到左）的演变情况。即使该批次由于化学状态不同（即此批次尚未进展到通用起始点）而从轨迹起点外开始，这在相对时间的可视化方面也不会产生问题。这在一维情况下同样适用，唯一的区别是新观测值可能从负的相对时间开始。这将在下面针对示例2中的流化床干燥数据进一步讨论。

3.2 示例2：流化床干燥操作的近红外（NIR）光谱

在制药过程中，对一种颗粒中间体进行流化床干燥时收集了近红外光谱。光谱共包含1093个波长，范围为9090–4484 cm−1（1100–2230 nm），分辨率为4 cm−1。由于傅里叶变换（FT）近红外光谱（特别是在如此高分辨率下）本身具有噪声特性，因此在应用标准正态变量变换（SNV）以减少光谱中的系统基线效应之前，使用11点移动平均对光谱进行了平滑处理。SNV处理后的光谱见图8。计算出的PCA模型显示，前两个主成分分别占方差的97%和1%。

分别地。验证方差表明两个主成分最优，这一点通过二维得分图的目视检查得到了确认。

图9显示了带有网格元素的历史批次，图10显示了带有95%置信限的估计轨迹以及新批次在模型上的投影（黑色点）。当将批次作为样本编号在一维得分图中绘制（图11）时，初步解读是...

1. 历史批次之间差异较大。
2. 新批次与历史批次不同。

然而，图9和图10中的二维得分图表明，批次确实在相对时间下遵循一条共同轨迹。因此，将得分与样本编号作图无法正确表示批次演变过程，因为不能假设新批次的监控起始点代表通用起始点t0。

图9中的二维得分图显示，由于起始材料含有不同水平的水分，各批次的起始位置各不相同。这意味着，所有批次遵循同一轨迹的位置才是通用起始点，而终点颗粒则是通过对干燥失重参考分析确认得到的。因此，当水分含量大致达到正确水平时，过程被停止。图9表明，一些批次过度干燥了，因为在相对时间下它们的终点略有差异。

新批次在过程开始时的得分（图10左下角）在前23个样本中几乎没有变化，但这对相对时间下的进展没有影响。另一方面，在图11中以样本编号绘制这些得分则给人一种印象，即新批次一开始就失控了。第三个历史批次（绿色）也是如此。

图10显示，由于水分含量高于历史批次的通用起始点，新批次的起点位于共同轨迹之外。这种情况在图9中已经可见，其中一些批次的首次观测位于左下角。然而，随着颗粒干燥，新批次沿轨迹前进。还要注意当批次进入共同轨迹时，从一次观测到下一次观测之间出现了间隙。这可能是由于颗粒的不均匀性造成的。不过，这在监控过程进展时不会造成任何问题。

4. 讨论

上述两个示例的结果表明，新方法能够在相对时间中建模批次进程。

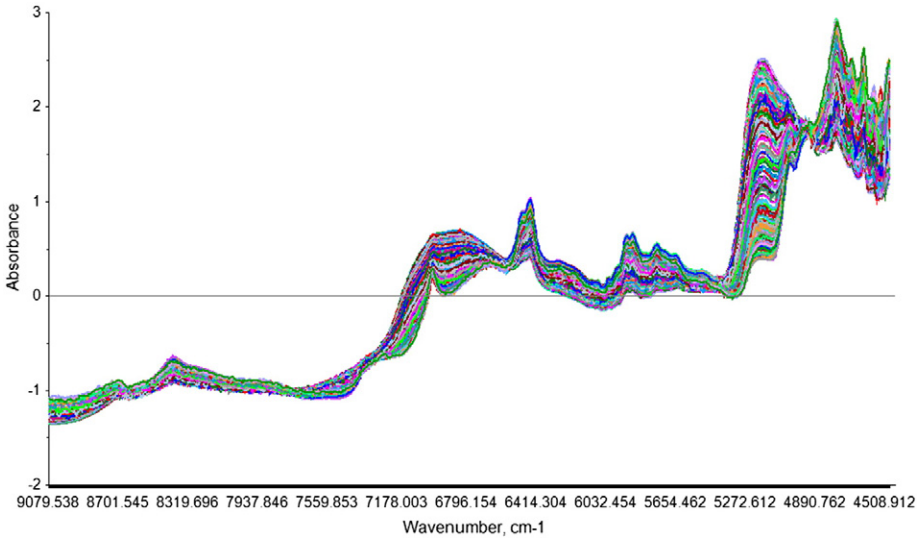

图8：SNV处理后的近红外光谱。

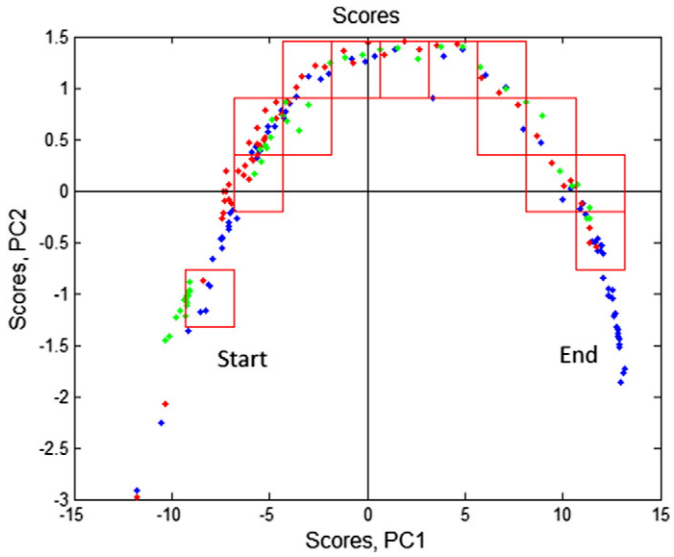

Fig. 9. Score plot of PC1 vs. PC2 for three historical batches with the grid elements shown.

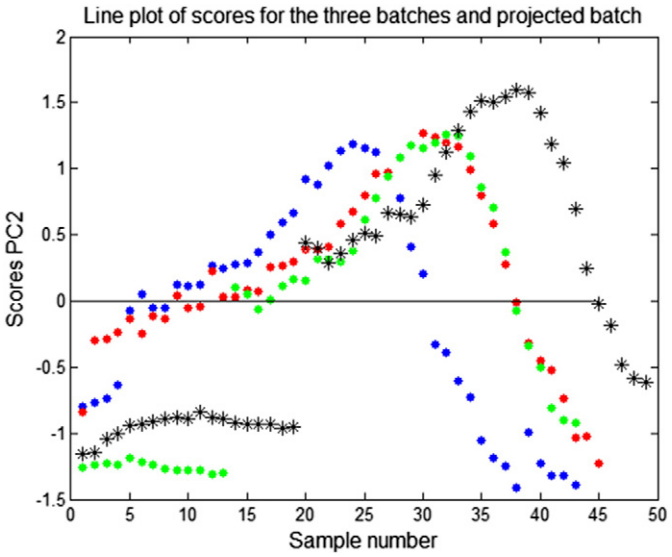

Fig. 11. Line plot of scores for PC2, historical batches as dots and the new batch as stars.

that better represents the batch progress compared to existing modeling approaches. The cross-validation over batches gives a conservative estimate of the explained validation variance as a way to decide on the optimal number of PCs. Figs. 7 and 10 depict the trajectory in 2D score plots; however, confidence intervals are also estimated for scores for individual PCs. If interpretation of the validation and visual assessment of the model concludes that a 1-dimensional model is optimal, the monitoring can still be performed in relative time.

The underlying assumption of the proposed approach is that the batch trajectories can be captured by a PCA model. From an assessment of the literature and from experience, this is true for most batch applications (primarily due to the highly correlated nature of consecutive points in batch models) as the data reflect the transition in the batch. If no feature space model can be found, the batch transition is not reflected in the data and it is doubtful that any of the existing batch modeling approaches would be successful.

In batch processes, there is typically a strong gradient from start to end. It can be imagined that the process is reversing, i.e. that the batch is revisiting an earlier feature state. As demonstrated in example 1, moderate reversing is handled by the relative time method. For more extreme reversing situations, adaptations of the approach might be required but that is outside the scope of this work.

In this paper, a static grid was used in the feature space to capture the systematic evolution of the batches. The static grid provides a robust and true representation of the presented examples. However, in more complex situations where e.g. heteroscedastic noise is present, a more flexible grid-search approach may be required.

Some batch processes, e.g. fermentation, are characterized by various phases transitions. In these situations, the correlation between the variables and the variance between the batches may not be constant for the whole duration.

Hierarchical models may prove useful in this case, where underlying classification models can be used to detect phase changes in an unsupervised manner and define a modeling and monitoring strategy based on joining numerous batch models together. In this situation, based on the location of the new time point in feature grid space, quantitative (or qualitative) models can be run in parallel where local refined models can be applied to assess the process, depending on the exact chemical/biological state of the material.

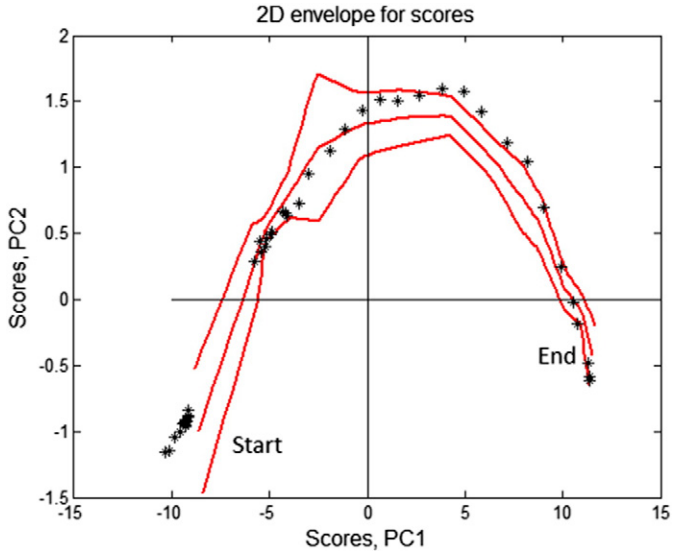

Fig. 10. Score plot of PC1 vs. PC2 with estimated trajectory from samples inside the grid elements, 95% confidence limit and a new batch projected.

5. Conclusions

The novel approach to batch modeling proposed in this paper models historical as well as monitors new batches directly in relative time. The common start and end points for the batches are found by a grid-search method. This eliminates the need for subjectively finding the samples that describe the trajectory in a common way, and thus, there is no need for time warping which may not be the correct remedy to handle various batch lengths and varying progression of the process in the first place.

Confidence intervals are estimated for the 2D and 1D score trajectories and can be extended to models with three (or higher) principal components (PCs). Monitoring of new batches is furthermore independent of the sampling rate allowing for changes in the sampling frequency between the NOC data set and the new batches.

The dynamic distance relative to the trajectory is estimated in a similar way as for Hotelling's T^2 statistic for PCA. This also extends to any model dimension. Dynamic Q-residual or F-residual critical limits are estimated along the trajectory for added diagnostic capability.

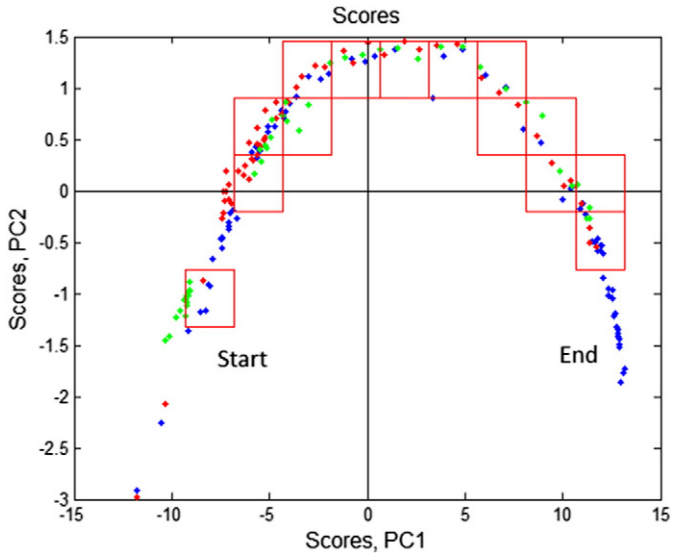

图9：三个历史批次的PC1与PC2得分图，并显示网格元素。

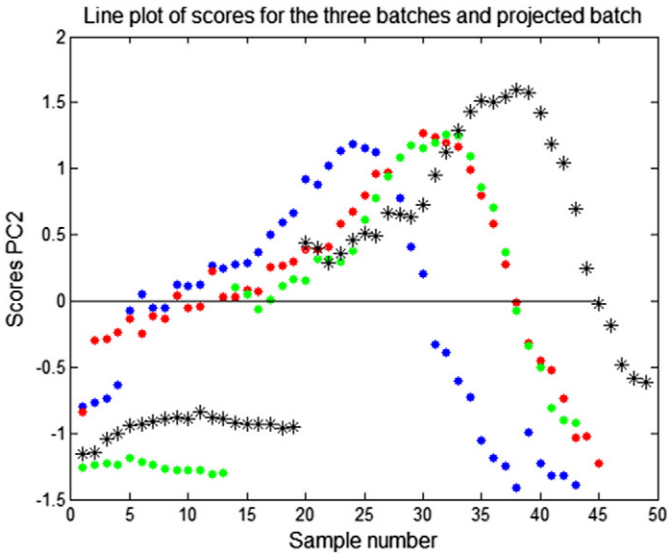

图11：PC2得分的线图，历史批次用圆点表示，新批次用星号表示。

相对于现有建模方法更能代表批次进展的方式。批次间的交叉验证给出了验证方差解释的一个保守估计，以此决定最佳主成分数目。图7和图10展示了二维得分图中的轨迹；然而，也对每个主成分的得分估计了置信区间。如果对验证结果和模型的可视化评估得出一维模型最优的结论，则仍可在相对时间下进行监控。

所提出方法的基本假设是，批次轨迹可以通过一个PCA模型来捕捉。根据文献评估和经验，对于大多数批次应用来说这是成立的（主要由于批次模型中连续点之间高度相关的特性），因为数据反映了批次中的过渡状态。如果没有找到特征空间模型，则数据没有反映批次过渡，此时任何现有的批次建模方法的成功都值得怀疑。

在批次过程中，通常从开始到结束存在强烈的梯度。可以想象的是，过程可能发生逆转，即批次出现倒退的情况。

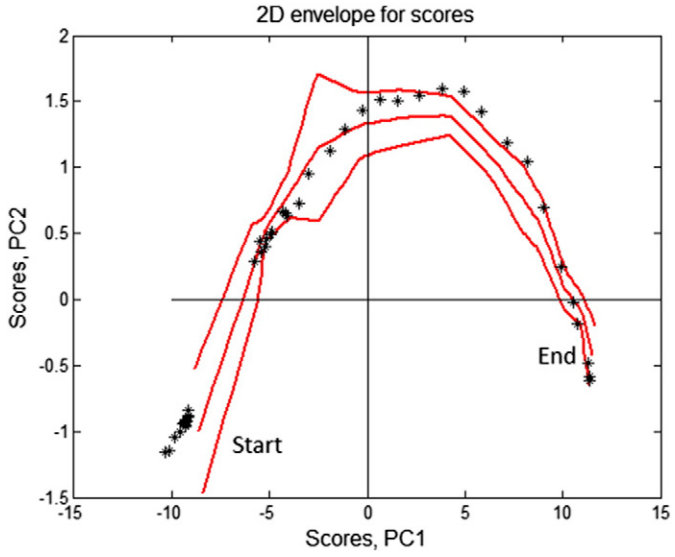

图10：PC1与PC2的得分图，网格元素内样本的估计轨迹、95%置信限及新批次的投影。

正在重新访问早期的特征状态。如示例1所示，适度的反向运动可以通过相对时间方法处理。对于更为极端的反向情况，可能需要对该方法进行调整，但这超出了本工作的范围。

在本文中，特征空间中使用了一个静态网格来捕捉批次的系统演化。静态网格提供了对示例的真实且稳健的表示。然而，在更复杂的情况下，例如存在异方差噪声时，可能需要更灵活的网格搜索方法。

某些批次过程，如发酵，其特点是存在各种阶段转换。在这种情况下，变量之间的相关性和批次之间的方差在整个过程中可能不是恒定的。

在这种情况下，分层模型可能会很有用，其中底层分类模型可用于无监督检测阶段变化，并基于多个批次模型的联合定义建模和监控策略。在这种情况下，根据新时间点在特征网格空间中的位置，可以并行运行定量（或定性）模型，具体取决于材料的确切化学/生物状态，可应用局部优化模型来评估工艺。

5. 结论

本文提出的一种新型批次建模方法直接在相对时间内对历史数据以及新批次进行建模和监控。通过网格搜索法找到批次的共同起点和终点。这消除了主观寻找能够以通用方式描述轨迹样本的需求，因此，不需要进行时间扭曲处理，因为时间扭曲可能并非处理不同长度批次和工艺进展变化的正确解决方案。

为二维和一维得分轨迹估算了置信区间，并可扩展至包含三个（或更高）主成分（PCs）的模型。新批次的监控还独立于采样率，从而允许NOC数据集与新批次之间采样频率的变化。

轨迹相对动态距离的估计方式与PCA中Hotelling T^2统计量类似，这也扩展到了任何模型维度。沿着轨迹估算动态Q残差或F残差的临界限值，以增强诊断能力。

The individual variables can be presented as a line plot in relative time facilitating real-time monitoring for end users with no knowledge about multivariate methods. Overall, this approach represents an objective way of modeling complex systems such as biological fermentations where starting material attributes are highly variable by nature. The application of processing variables to such materials results in different manners that process can proceed. Eliminating time from the analysis (in a modeling sense) allows chemical/biological probing of the process in the state that the material exists in the process, without implying any need to fit the model to a certain time scale in terms of a given sample number.

References

[1] S. Wold, P. Geladi, K. Esbensen, J. Ohman, J. Chemom. 1 (1987) 41–56.
[2] P. Nomikos, J.F. MacGregor, AICHE J. 40 (1994) 1361–1375.
[3] J. Camacho, J. Pico, A. Ferrer, Anal. Chim. Acta 642 (2009) 59–68.
[4] X. Meng, A.J. Morris, E.B. Martin, J. Chemom. 17 (2003) 65–85.
[5] A. Kassidas, J. MacGregor, P. Taylor, AICHE J. 44 (1998) 864–875.
[6] C. Ündey, S. Ertunç, A. Çinar, Ind. Eng. Chem. Res. 42 (2003) 4645–4658.
[7] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor, Chemom. Intell. Lab. Syst. 80 (2006) 1–12.
[8] J. Camacho, J. Pico, A. Ferrer, J. Chemom. 22 (2008) 299–308.
[9] A. Bogomolov, Chemom. Intell. Lab. Syst. 108 (2011) 49–63.
[10] H. Martens, M. Martens, Multivar. Anal. Qual. Wiley & Sons, 2001.

各个变量可以以相对时间的线图形式呈现，便于最终用户在不了解多元方法的情况下进行实时监控。总体而言，这种方法代表了一种客观建模复杂系统的方式，例如生物发酵过程，其中原料属性本质上具有高度变异性。对这类材料应用工艺变量会以不同的方式推进过程。从分析中消除时间因素（在建模意义上）允许以材料存在于工艺中的状态对工艺进行化学/生物探测，而无需将模型拟合到特定的时间尺度上。

参考文献

[1] S. Wold, P. Geladi, K. Esbensen, J. Ohman，《化学计量学杂志》，1卷（1987年），第41–56页。
[2] P. Nomikos, J.F. MacGregor，《美国化学工程师学会志》，40卷（1994年），第1361–1375页。
[3] J. Camacho, J. Pico, A. Ferrer，《分析化学学报》，642卷（2009年），第59–68页。
[4] X. Meng, A.J. Morris, E.B. Martin，《化学计量学杂志》，17卷（2003年），第65–85页。
[5] A. Kassidas, J. MacGregor, P. Taylor，《美国化学工程师学会志》，44卷（1998年），第864–875页。
[6] C. Ündey, S. Ertunç, A. Çinar，《工业与工程化学研究》，42卷（2003年），第4645–4658页。
[7] P.R.C. Nelson, P.A. Taylor, J.F. MacGregor，《化学计量学智能实验室系统》，80卷（2006年），第1–12页。
[8] J. Camacho, J. Pico, A. Ferrer，《化学计量学杂志》，22卷（2008年），第299–308页。
[9] A. Bogomolov，《化学计量学智能实验室系统》，108卷（2011年），第49–63页。
[10] H. Martens, M. Martens，《多变量分析质量》。Wiley & Sons，2001年。