Spectroscopic model calibration in biomanufacturing using just-in-time learning

Aditya Tulsyan[a,*], Hamid Khodabandehlou[b], Tony Wang[b], Gregg Schorner[c], Myra Coufal[a], Cenk Undey[b]

[a] *Digital Integration & Predictive Technologies, Amgen Inc., 360 Binney Street, Cambridge MA 02141, USA.*
[b] *Digital Integration & Predictive Technologies, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA.*
[c] *Digital Integration & Predictive Technologies, Amgen Inc., 40 Technology Way, West Greenwich RI 02817, USA.*

Abstract

Spectroscopic instruments play an instrumental role in the implementation of the U.S. Food and Drug Administration (FDA) as outlined in process analytical technology (PAT) guidance for biopharmaceutical manufacturing. Industrial spectroscopic calibration models are typically developed in an offline setting using traditional regression-based methods, such as partial least squares (PLS) and principal component regression (PCR). Apart from the limiting performances of these offline models under time-varying operating conditions, these methods require access to large historical data, which are seldom available in biopharmaceutical manufacturing. In this paper, we propose a novel just-in-time learning (JITL) platform for automatic real-time model calibration and maintenance using routine campaign data. The proposed framework uses Bayesian non-parametric Gaussian processes (GPs) as calibration models. A GP model not only exhibits superior performance over a PLS or PCR model across different operating conditions but also provides credibility intervals around model predictions. The efficacy of the proposed method is illustrated on a real-time calibration problem for a biopharmaceutical process.

Keywords: Biopharmaceutical manufacturing, Raman spectroscopy, real-time monitoring, machine-learning.

1. Introduction

The Raman spectroscopy is a modern PAT tool widely used in biomanufacturing. As an optical method, Raman enables non-destructive analysis of chemical composition and molecular structure. Applications of Raman in the polymer, biomanufacturing, and biomedical analysis have surged in the past three decades as laser sampling, and detector technology has improved. Raman spectroscopy is now a practical analysis technique inside and outside the laboratory.

Raman models in biomanufacturing are nontrivial to calibrate as biopharmaceutical processes operate under many stringent constraints and tighter regulations. The current state-of-the-art for Raman model calibration in the biopharmaceutical industry is first to run multiple campaign trials to generate relevant data to correlate the Raman spectra to the analytical measurements. These trials are not only expensive to campaign but also time-consuming, as each campaign may last anywhere between two to three weeks in laboratory settings. Further, to ensure that a lab-scale bioreactor (usually with a working volume of $1-3$ L) maintains a healthy mass of viable cells, only limited samples are available for the analytical instruments. It is not uncommon to have only one or two measurements available each day from the in-line or offline analytical tools. Further, once a calibration model is built, it is common to observe the performance of the model degrade with each campaign. This is because, biomanufacturing processes often undergo a different kind of changes, such as the recipe change, raw material variability, and process-drifts that may cause gradual degradation of the calibration model. Calibration models based on PLS or PCR are worst affected, as the performance of these methods is prone to degrade under such dynamic operations. It is therefore imperative to update the calibration model periodically to ensure that the model reflects on the current process operations and can sustain good prediction performance.

To address some of the limitations mentioned above of the current best industrial practices, this paper proposes a JITL platform for building Raman calibration models for biopharmaceutical applications. JITL is a novel nonlinear modeling platform, that is based on local-modeling and database sampling technology. Different from traditional methods, JITL assumes that all available observations are stored in a central database, and models are dynamically built in real-time upon query, using the most relevant data from the database. This allows to approximate complicated process dynamics using simple local models. Under the JILT framework, a library may contain spectral data not only for a single product operating under different operating conditions, but also data for various products, and under different media conditions. In other words, it is possible to recycle spectral data from across different product lines and from across different operating conditions. This significantly reduces the time required to calibrate Raman models, especially for the pipeline drugs, with limited or no past production history. For local modeling, we propose to use Gaussian Process (GP) models [1], which are powerful statistical machine-learning models that can efficiently capture complex nonlinear process dynamics and readily adapt to any process changes. In contrast to a PLS or PCR model, a GP model is a non-parametric method and offers far more flexibility in capturing the complex correlations between the spectra and the analytical measurements. The proposed method in-

*Corresponding author. Tel.:+1 617-3141-608.
Email addresses: atulsyan@amgen.com (Aditya Tulsyan), hkhodaba@amgen.com (Hamid Khodabandehlou), tonyw@amgen.com (Tony Wang), schorner@amgen.com (Gregg Schorner), mcoufal@amgen.com (Myra Coufal), cundey@amgen.com (Cenk Undey)

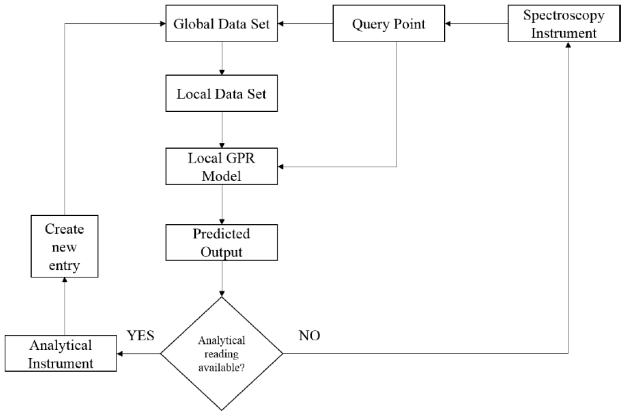

Figure 1: A flowchart for the proposed JITL-based Raman model.

troduces a paradigm shift in the way biopharmaceutical calibration models are built and maintained, which to the best of authors' knowledge have not been done before.

2. Just-in-Time Learning (JITL) Framework

JITL is a novel nonlinear modeling platform, that is based on local-modeling and database sampling technology [2]. Different from traditional methods, JITL assumes that all available observations are stored in a central database, and models are dynamically built in real-time upon query, using the most relevant data from the database. In general, there are three critical steps in JITL. First, when a query sample arrives, samples that are most 'similar' to the query sample are selected from the database as training samples. Second, a local regression model is built using the training samples. Finally, the local model is used for predicting the output of the query sample. A JITL model only has a local validity around the query point as the model is discarded immediately after use, just to be rebuilt again around the next query point. To ensure that the library stays updated, if an analytical measurement is available at any time, as a part of routine offline or in-line sampling schedule, the library is updated by adding the spectrum and the corresponding analytical measurement to the library. This ensures that the current process information is a part of the library.

The critical factor for the success of the JITL technique is to select relevant samples properly, which is based on certain similarity or dissimilarity measurements. Hence, several sample selection methods have been developed in recent years, like the Euclidean or Mahalanobis distance-based, angle-based, and correlation-based similarity indices. Mathematically, given a query point $\mathbf{a}^\star \in \mathbb{R}^{n_a}$, and a central library $\mathcal{L}_t \equiv \{b_i, \mathbf{a}_i\}_{i=1}^{L_t}$ containing $L_t \in \mathbb{N}$ input-output pairs, we are interested in selecting a local training set $\mathcal{D}_t \equiv \{b_j, \mathbf{a}_j\}_{j=1}^{D_t}$ at time $t \in \mathbb{N}$ containing $D_t \in \mathbb{N}$ samples, where $D_t << L_t$. It is assumed that $\mathcal{L}_t$ is dynamic, and may include different entries during a campaign. Now there are numerous ways to construct $\mathcal{D}_t$ from $\mathcal{L}_t$. For the JITL framework proposed in this paper, we select $\mathcal{D}_t$ based on Euclidean distance between the spectra in set $\mathcal{L}_t$. This is motivated by the fact that GPs in the proposed JITL framework is also based on the Euclidean distance. A schematic of the proposed JITL-based Raman model is given in Figure 1 and the algorithm is formally outlined in Algorithm 1. As

Algorithm 1 JITL-based Raman model

1: **Input:** Library $\mathcal{L}_t$, query point $\mathbf{a}^\star$
2: **Output:** Model predictions
3: **for** $t = 1$ to T **do**
4: Set $\mathcal{U}_t \leftarrow \mathcal{L}_t$ and $\mathcal{D}_t \leftarrow \{\emptyset\}$
5: **for** $j = 1$ to D_t **do**
6: $k_\star \in \arg\min_{i \in \mathcal{U}_t} ||\mathbf{a}_i - \mathbf{a}^\star||_2$
7: $\mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{b_{k_\star}, \mathbf{a}_{k_\star}\}$
8: $\mathcal{U}_t \leftarrow \mathcal{L}_t \setminus \mathcal{D}_t$
9: **if** $\mathcal{D}_t \cap \{b_{L_t}, \mathbf{a}_{L_t}\} = \{\emptyset\}$ **then**
10: $\mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{b_{L_t}, \mathbf{a}_{L_t}\}$
11: **end if**
12: **end for**
13: Train GP with $\mathcal{D}_t$ and compute model predictions
14: **if** $b^{\star\star}$ is available **then**
15: $\mathcal{L}_{t+1} \leftarrow \mathcal{L}_t \cup \{b^{\star\star}, \mathbf{a}^\star\}$
16: **end if**
17: **end for**

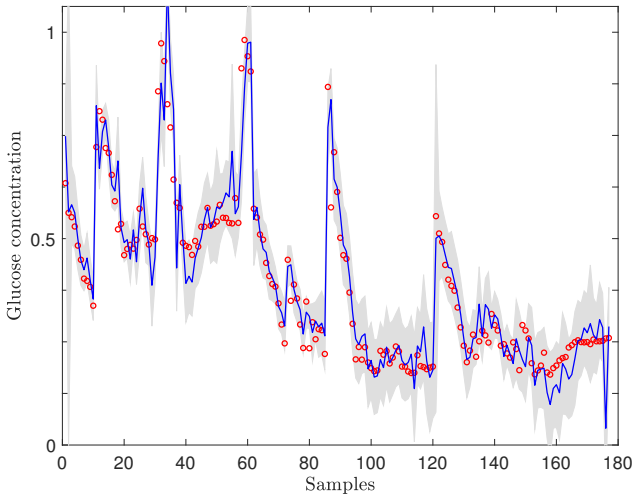

Figure 2: Comparison of the proposed JITL glucose predictions (solid-blue curve) and the true glucose concentrations measured using the analytical instrument. 95% credibility interval around the model predictions are represented by the grey shaded area.

noted earlier, to ensure that the library $\mathcal{L}_t$ remains current, if at anytime during the experiment, if the data point $\{b^{\star\star}, \mathbf{a}^\star\}$ is made available, where $b^{\star\star}$ is the analytical measurement corresponding to the spectral scan $\mathbf{a}^\star$, then the library is updated to include the new entry. This step is captured in Steps 14–15 in Algorithm 1. To ensure that the local model quickly adapts to any abrupt process changes or a new product, we always include the last available measurement from the current experiment in the training set. This is given in Steps 9–11 in Algorithm 1. An industrial case study is discussed next.

3. Industrial Case study

We demonstrate the efficacy of Algorithm 1 in calibrating Raman models in biomanufacturing. A Chinese hamster ovary (CHO) monoclonal antibody (mAb) secreting cell-line derived

from the host was used for experiments. The cell line was stably transfected with a proprietary DNA vector to express the relevant mAb. All media and feeds used in the experiment are proprietary solutions. The experiments were carried out in a laboratory fed-batch cell-culture bioreactor.

Controlling glucose concentration in cell-culture bioreactors through optimized feeding strategies is highly essential, and is critical for increasing cell-growth, and hence productivity as well as product quality consistency. The glucose concentrations are currently analyzed as frequently as every 6 hour using an in-line automated sampling system – BioProfile Flex Analyzer (Nova Biomedical, Waltham, MA). To calibrate a Raman model, in parallel to the analytical instrument, a stainless-steel immersion probe was also used to collect and transfer cell-culture Raman signal through a fiber optic probe to a RXN2 Raman Analyzer (Kaiser Optical Systems Inc., Ann Arbor, MI). A laser with 785 nm excitation wavelength was used, with an approximate power of 200 mW at the probe tip. The Raman spectra were acquired by implementing cosmic-ray removal and dark-spectrum subtraction with an exposure time of 10 seconds, adding 75 scans consecutively to result in a collection time slightly above 750 seconds (considering overhead instrument time). The interval for collecting Raman spectra was approximately 15 minutes.

The cell line used in the experiment did not have previous campaign history. No historical data is available for the cell-line to calibrate the Raman glucose model. In such a scenario, the standard model calibration practice is to continue running multiple campaign trials to collect relevant and sufficient data. Unfortunately, a campaign lasts for about two weeks in a laboratory setting and incurs significant material cost alone. Running additional campaign trials in biopharmaceutical manufacturing is not only labor and cost-intensive, but it also causes delays in getting the drug to the patients. To mitigate this, in the absence of any historical data, we build a library for the JITL platform using cell-culture data from our past experiments. In this work, we take data from two other experiments, that were previously campaigned for two different cell lines and under different media profiles and operating conditions. We use this library as a starting point to implement the proposed JITL framework.

The original library contains 469 data points, and for local modeling, we choose 100 training samples. The size of the training set is fixed throughout the campaign; however, the library size is dynamic, as new measurements from the BioProfile Flex Analyzer are included as they become available. The training set also contains the last recorded entry in the library from the current experiment. This is to ensure that the Raman model quickly adapts to the new cell-line (see Algorithm 1).

Figure 2 compares model predictions from the proposed JITL platform to the BioProfile Flex Analyzer measurements. 95% credibility intervals around the predictions are also shown. Overall, the model predictions are in close agreement with the analytical measurements. Observe that despite having no past campaign history, the model can quickly adapt to the new cell-line, without significant delays. The total root-mean-square error (RMSE) with Algorithm 1 is 13.83.

Figure 3 shows a scatter plot comparing the performances of the model with static and dynamic libraries. A static library is not updated based on the new measurements sampled during the experiment. The benefits of maintaining a dynamic library are evident. The total RMSE with a static library is 68.86,

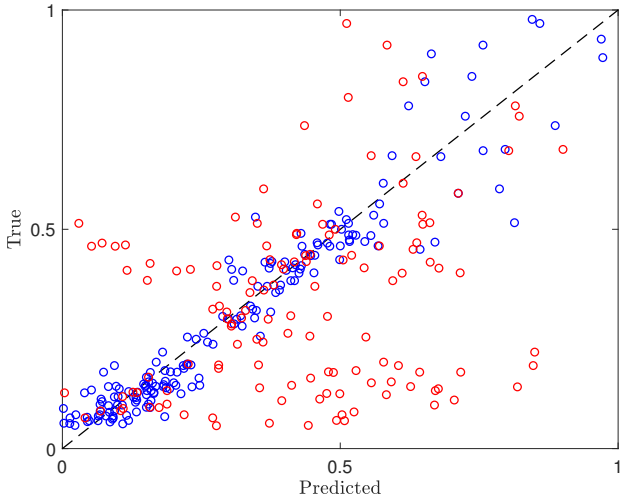

Figure 3: Predicted vs. true glucose concentrations with the proposed method using dynamic and static libraries.

which is significantly higher than that for the dynamic library.

4. Conclusions

In this paper, we proposed a novel just-in-time Gaussian process modeling framework to quickly calibrate spectroscopic instruments in biopharmaceutical manufacturing using routine operating data. The developed method is an adaptive learning algorithm that allows for automatic real-time model calibration and model maintenance for different cell-lines across different operating conditions. Another distinct advantage of the proposed method over traditional methods is that it yields credibility intervals around model predictions, which can be used to design robust control and monitoring strategies. The proposed adaptive framework presents a paradigm shift in the way model calibrations are typically performed in biopharmaceutical manufacturing; and has a potential to significantly reduce material and labor costs associated with such routine calibrations.

References

[1] T. Chen, J. Morris, and E. Martin, "Gaussian process regression for multivariate spectroscopic calibration," *Chemometrics and Intelligent Laboratory Systems*, vol. 87, no. 1, pp. 59–71, 2007.

[2] C. Cheng and M.-S. Chiu, "A new data-based methodology for nonlinear process modeling," *Chemical Engineering Science*, vol. 59, no. 13, pp. 2801–2810, 2004.