European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

Bioprocesses and other production processes with multi-stability for method testing and analysis

Teemu Vesterinen and Risto Ritala
Tampere University of Technology, Institute of Measurement and Information Technology
P.O. Box 692, FIN -33101 Tampere, FINLAND
teemu.vesterinen@tut.fi

Abstract

This paper describes a simulator of a simple but stochastic and bistable bioprocess. Our emphasis is to generate test data for our novel analysis methods of transient and non-stationary systems. We show through simulations how the stochastic effects cause transitions between the stable fixed points of the deterministic model. We illustrate modeling the data with non-normal distributions and demonstrate how transients are analyzed through dynamic simulation of probability density functions.

Keywords: simulation, nonlinear, stochastic, bistable, bioprocess, analysis

Introduction

All the data analysis methods for process data assume stationarity. Most of the methods assume that there is only one operational point at which both the linear and non-linear disturbances are analyzed. However, there exist strongly non-linear multistable stochastic systems that can not be linearized. Over short time periods such systems appear unimodal, but they have long transient time constants to multi-modality.

Our goal is to develop methods to analyze such data and to estimate the time constants so that the analyst can assess whether the data is about stationary behavior or has long transients. A longer term goal is to expand the methods to cover truly non-stationary systems. To develop the methods we need data both from real systems and simulations.

The probability distribution of multistable process data shows two or more peaks rather in contradiction with the unimodal normal distribution. For the most part this is due to different operational states linked through abrupt operator actions on set points, but multimodality indicates also sudden spontaneous or smoothly driven transitions.

This study presents a bioprocess simulator, which generates strongly nonlinear, bistable and stochastic test data for method development and testing. Both driven and spontaneous transitions can be tested. Simulator properties are well known and can be easily varied, thus serving as a good basis for method development.

This paper presents the rationale of our data analysis method development. We first briefly summarize the stability analysis of multidimensional systems. Secondly, for motivation, we describe a stochastic bioprocess simulator with bistability. Next, the data analysis methods [Latva-Käyrä and Ritala, 2004] are illustrated with the simulator data. Finally, the methods are also tested with the real life paper mill data sets.

Non-linear stochastic systems

The dynamics of a system describes deterministic causal effect of the present on the future state. This causal relation is presented with a differential equation

$$\frac{d\bar{x}}{dt} = F(\bar{x}) \tag{1}$$

A fixed point x^* of a system is solution to $F(x^*) = 0$. The stability of the fixed points is analyzed by inspecting the eigenvalues of the Jacobian matrix at the fixed points [Sprott, 2003]. Fixed points may be stable, saddle or unstable.

For a one-dimensional system the Jacobian is a scalar, and thus positive for unstable fixed points, and negative for stable fixed points. The neighboring fixed points must be of different type. In two dimensions the fixed point structure has more options. Furthermore, a limit cycle [Khalil, 1996]. i.e. a non-driven periodic behavior may arise.

In three and higher dimensions the possibilities of asymptotic behavior are even more diverse, including the seemingly random behavior of deterministic chaos.

When a system is under stochastic effects and has several stable attracting asymptotic solutions, the system state jumps randomly between the attractor basins of the solutions. Depending on the size of the attractor basin and the intensity of the stochastic effects, the transients of probability density function (pdf) dynamics may be rather long. We analyze systems having at least two stable fixed points and stochastic effects strong enough to induce transitions between the fixed points over the studied time period.

Production systems and their physico-chemical phenomena are subject to stochasticity. Hence when analyzing sampled data from such systems with discrete-time, continuous-state models, a natural single signal model is the stochastic difference equation

$$x_{n+1} = x_n + g(x_n) + \varepsilon_n \tag{2}$$

where ε_n is white noise process having zero mean and variance of σ_ε^2.The simplest model experiencing bistability and hence our prototype model for bistability is

$$g(x) = ax + bx^3 + h \tag{3}$$

We are developing methods to identify data models for multistable systems and assessing how and to what extent transients affect the identification.

Bioprocess simulator

Chemostat is a biological reactor into which micro-organisms and substrate are fed. Chemostat is a potentially multistable system if the substrate at high concentrations is toxic for micro-organisms. Then an increase of in fed substrate turns a linear behavior with quite normally distributed data, into a strongly nonlinear one with complex state distributions. Chemostat gives also an insight to real life bioprocess systems, in particular biological water treatment.

We implemented an ideally stirred Chemostat [Smith, Waltman, 1995] simulator of water treatment known to have bistability [Dramé et al, 2003]. The model has two input streams: substrate S and biomass of microbes B. The simulator is described by Eq. 4.

$$\begin{cases} \dfrac{d}{dt}c_S = \dfrac{Q_{in}}{V}(c_{Sin} - c_S) - \mu(c_S)c_B \\ \dfrac{d}{dt}c_B = \dfrac{Q_{in}}{V}(c_{Bin} - c_B) + \mu(c_S)c_B \\ \dfrac{d}{dt}V = Q_{in} - Q_{out} \end{cases} \tag{4}$$

where Q_{in} and Q_{out} indicate flows into and out of the tank, c_Xs indicate the stream and tank concentrations, and V is the volume of the tank. At low substrate concentration biomass grows through consuming the substrate, but when substrate concentration gets high enough, it becomes toxic for biomass. Toxic effect is given by Eq. 5.

$$\mu(c_S) = \frac{\mu_0 c_S}{\tilde{K}^{-1}c_s^2 + c_S + K} \tag{5}$$

where μ_0 is the maximal growth rate without toxic effect. $\tilde{K}$ *is* the toxic effect parameter, and μ_0/K the growth rate at small substrate concentrations. This is the simplest form of growth rate leading to bistability through toxic effects.

The volume V is a variable triggering bistability in the deterministic system: volume affects the fixed points of dynamics, (c_S^*, c_B^*). Figure 1 shows how the tank volume and the concentration of substrate flowing into the tank affect the system fixed points under steady state conditions. For small tank volumes the fixed point is unique independently on the incoming substrate concentration. Somewhat below 200 units of volume the single fixed point bifurcates into two stable and one unstable fixed point so that for higher volumes there is region of incoming substrate concentration where two stable conditions coexist. For practical operation of a bioreactor, it is preferable to have both stable operational state and high conversion of substrate into biomass. Below tank volume of 200 units, bioreactor is quite stable, but conversion of incoming substrate to biomass is rather poor. When increasing the tank volume, system can convert better, but a fluctuation in volume may drive the system to poor conversion fixed point. It takes a long time or active undesired manipulation to reach the high-conversion state again.

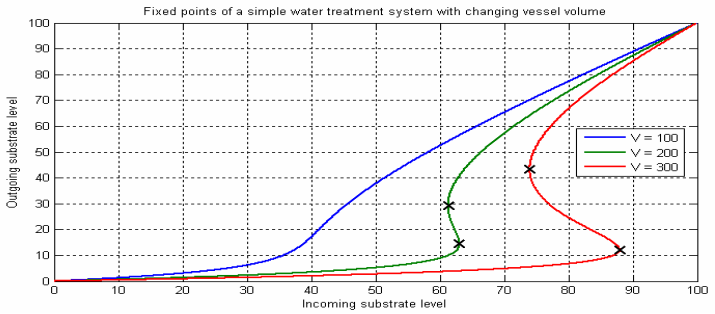

Figure 1. Volume changing the fixed point structure.

We have introduced stochasticity in the model both through random variations in concentration of flows into the tank and those in flow rate into and/or out of the tank as described. Figure 2 shows an example of the simulation results. The nominal point is

$Q_{in} = Q_{out} = 100$, V = 300, $c_{Bin} = 10$ and $c_{Sin} = 65$. Stochasticity has been introduced in c_{Sin}, which is uniformly distributed random value between [35 95] and changing after each 2000 s periods. The growth rate parameters are $\mu_0 = 0.74$, $\tilde{K} = 15$ and $K = 9.28$.

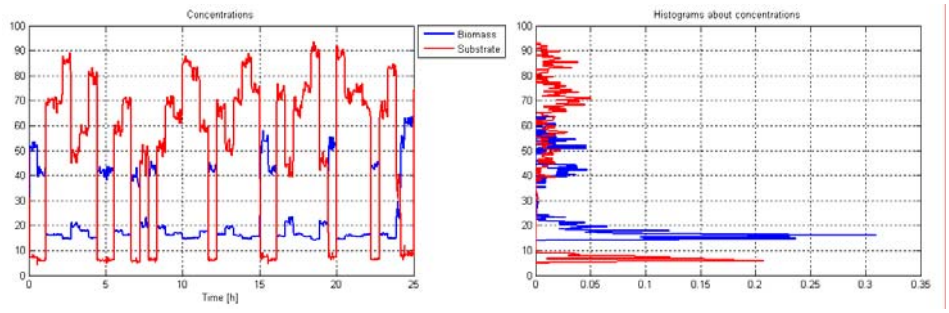

Figure 2. Substrate and biomass concentrations after the tank.

We generated test data for method development with this simulated model under stochastic disturbance. The data proved to be valuable for method development.

Analysis methods and simulation of pdf dynamics

By inspecting the histogram in Fig. 2, it is clear that the simplest distributions to fit in to that kind of data set are a bimodal exponent-of-polynomial distribution (to be referred to also as the Two-Hump Distribution, THD) and the Gaussian Mixture Model (GMM) [Nabney, 2001]. THD reads as [Vesterinen and Ritala, 2004]

$$f(x) = \Sigma^{-1} M(\lambda,\eta)^{-1} \exp\left[\frac{\lambda(x-\mu)^2}{\Sigma^2} - \frac{(x-\mu)^4}{\Sigma^4} + \frac{\eta(x-\mu)}{\Sigma}\right] \qquad (6a)$$

$$M(\lambda,\eta) = \int_{-\infty}^{\infty} \exp(\lambda s^2 - s^4 + \eta s) ds$$

The THD assumes that the system is stationary and bistable. Thus the system is autonomous and all transitions are spontaneous, without an action of external deterministic driving force.

The THD parameters μ, λ, η, Σ are estimated with Maximum Likelihood (ML) method [Vesterinen and Ritala, 2004]. This results in solving a nonlinear optimization problem

$$\min_{\mu_2,\Sigma_2,\lambda,\eta} \left[-\log(\Sigma_2) - \log(M(\lambda,\eta)) + \lambda \frac{1+\mu_2^2}{\Sigma_2^2} - \frac{M_4 - 4\mu_2 M_3 + 6\mu_2^2 + \mu_2^4}{\Sigma_2^4} - \eta \frac{\mu_2}{\Sigma_2} \right] \qquad (6b)$$

where M_n is the nth moment of x, estimated from data, and μ_2 and Σ_2 are the parameters of distribution (6a) for variable which is scaled to zero mean and variance 1. Note that μ_2 is not zero, as (for nonzero η), μ is not the mean of distribution (6a).

This is straightforward to solve numerically. However, as the likelihood function includes up to fourth order moments, it is very sensitive to the outliers.

In the case of GMM the basic idea is that the system has a number of states, each of which described with normal distribution, and the system is driven externally between the states. In GMM the PDF is a linear combination of basis functions [Nabney, 2001]:

$$f(x) = \sum_{j=1}^{M} P_j * (2\pi\sigma_j^2)^{-1/2} \exp\left[-\frac{(x-\mu_j)^2}{2\sigma_j^2} \right] \tag{7}$$

$$\sum_{j=1}^{M} P_j = 1, \quad 0 \le P_j \le 1$$

where P_j is the weight of the j[th] Gaussian probability distribution. In our analysis we used the expectation maximization (EM) algorithm of [Figuereido and Jain, 2002]. Results of the fitted distributions are shown on figure 3.

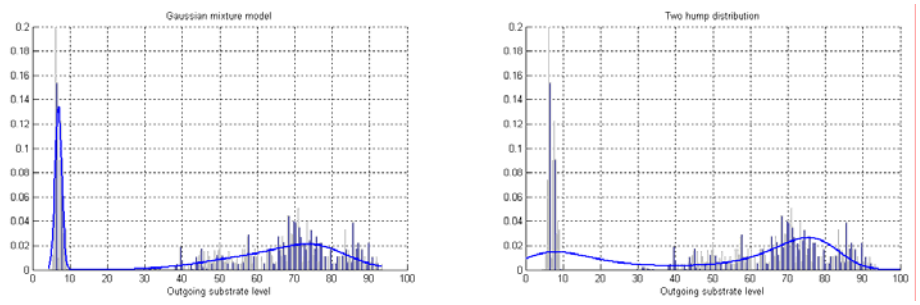

Figure 3. GMM and THD approximations of the data set.

The bioprocess differs considerably from the prototype bistability model Eq. 3 and thus THD with only three polynomial terms is rather poor. GMM gives better results. By using more detailed bistability model, THD would improve, but the resulting ML parameter estimation problem would be hard to solve.

The great advantage of THD is that it is the asymptotic distribution of dynamics, Eqs. 2-3, and yields parameters for Eq. 3. The corresponding distribution dynamics is:

$$f(x;n) = \int_{-\infty}^{\infty} \left[(2\pi\sigma_\varepsilon^2)^{-1/2} \exp\left[-\frac{(x-x'-g(x'))^2}{2\sigma_\varepsilon^2} \right] * f(x', n-1) \right] dx' \tag{8}$$

Distribution dynamics is determined with asymptotic THD parameters of *g(x)*. Distribution dynamics can be analyzed in discretized state space either through eigenvalues [Latva-Käyrä and Ritala, 2004] or by direct simulation. Figure 4 shows a simulation example based on THD model of Fig. 3 at five instants (distribution appropriately normalized). Simulation was started with state distribution *N(-1,0.05)* corresponding to the bioprocess being in the desired state. Integrating the peak around 1.7 at each instant gives probability that the transition to undesired state has occurred.

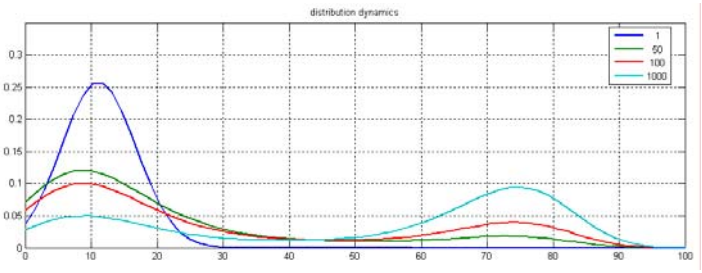

Figure 4. Distribution dynamics after1, 50, 100, and 1000 steps.

The leading eigenvalue of operator (8) gives the time constant for the transition [Risken, 1996], [Latva-Käyrä and Ritala, 2004], in accordance with the simulated result.

Paper mill case study

The distribution estimation methods were also applied to real process data, which obviously is the goal of our development work. With real data we gain good insight also to the robustness of our new methods.

We applied our methods on data from two industrial production processes. The data included hundreds of channels, all of which were analyzed. We analyzed the data with normal distribution, THD and GMM and compared the results with Q-Q plots [Seber, 1984]. Most of the data is best described with GMM which suggests the multimodality to be driven by abrupt operator actions. However, in some cases the THD provides a quite good description of the data. Then also the distribution dynamics can be analyzed and the slowest time constants of spontaneous transitions determined.

The outliers are the main problem with real process data. In particular THD involving high order moments is sensitive to outliers.

Discussion

In this article we have described a bistable stochastic bioprocess simulator for method development and testing. This model has been useful and given information about the developed data analysis methods. With the data analysis methods we have linked the stochastic differential based simulation to pdf dynamics simulation. Aim is to use these methods on real processes and therefore the simulator must be developed to more realistic direction. This could contain outliers and missing values in resulting data. We are continuing this towards more detailed models for both the bioprocess simulator, and for more advanced analysis models in which g(x), Eq. (3) is a higher order polynomial or a multilayer perceptron network.

References

Dramé A.K., Harmand J., A. Rapaport, Lobry C., 2003, ”Multiple Steady State Profiles in Interconnected Biological Systems”, Mathmod Vienna, 352-359.

Figuereido M.A.T., Jain A.K., 2002, ”Unsupervised Learning of Finite Mixture Models”, IEEE Transactions on Pattern. Analysis and Machine Intelligence, Vol 24, (3), 381-396.

Khalil H.K., 1996, ”Nonlinear Systems”, Prentice-Hall, Inc, 1-56.

Latva-Käyrä K., Ritala R., 2004, ”system identification scheme for stochastic bistable process”, ISITA 2004 - International Symposium on Information Theory and its Applications, 733-736.

Nabney I.T., 2001, “Netlab: algorithms for Pattern Recognition” 2nd edition, Springer, 79-116.

Risken H., 1996, “The Fokker-Planck equation: Methods of Solution and Applications”, Springer, 96 -132.

Seber G.A.F., 1984, “Multivariate observations”, John Wiley & Sons, Inc, pp. 141-155.

Smith H.L., Waltman P., 1995, ”The Theory of the Chemostat”, Cambridge University Press, 1-77.

Sprott J.C., 2003, ”Chaos and Time-Series Analysis”, Oxford University Press, 72-103.

Vesterinen T., Ritala R., 2004, “Data analysis of stochastic bistable systems: applications to biological water treatment plant model and paper mill data”, Control Systems 2004 Conference, 245-249.