Stochastic Parameter Estimation Using PDF Shaping

George Papadopoulos and Martin Brown
The Control Systems Centre
University of Manchester

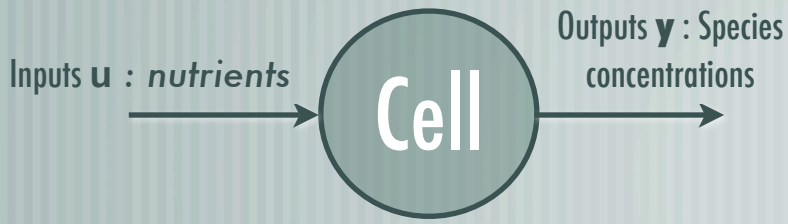
Overview

1. This presentation is structured as follows:
    1. Main Scope: Construct a parameter estimation framework based on the stochastic properties of the biochemical reaction system. Study on autonomous systems.
    2. Four main parts focusing on:
        1. **Systems Biology** and **Systems ID**.
        2. **Systems Identification using PDf Shaping**: mathematical description of the proposed parameter estimation scheme.
            1. What assumptions are made.
            2. Mathematical formulation.
            3. PDF estimation method (KDE)
        3. **PDF shaping Using the Entropy**.
        4. **Examples**: based on a univariate description including cases where the noise component could have an arbitrary distribution.

Introduction: Systems Biology and Systems ID

1. Systems ID, an important aspect of Systems Biology:
    1. The signaling pathway is translated into a parametrized ODE model.
    2. Modeling procedure where the parameters are chosen so as to map the given data.
    3. Obtain insight into the cells internal characteristics.

The Cell in General



Example of an Autonomous System
Michaelis Manten Signaling pathway

$$E + S \underset{\theta 2}{\overset{\theta 1}{\rightleftarrows}} ES \xrightarrow{\theta 3} E + P$$

$$\frac{dE(t)}{dt} = -\theta_1 E(t)S(t) + \left(\theta_2 + \theta_3\right) ES(t)$$

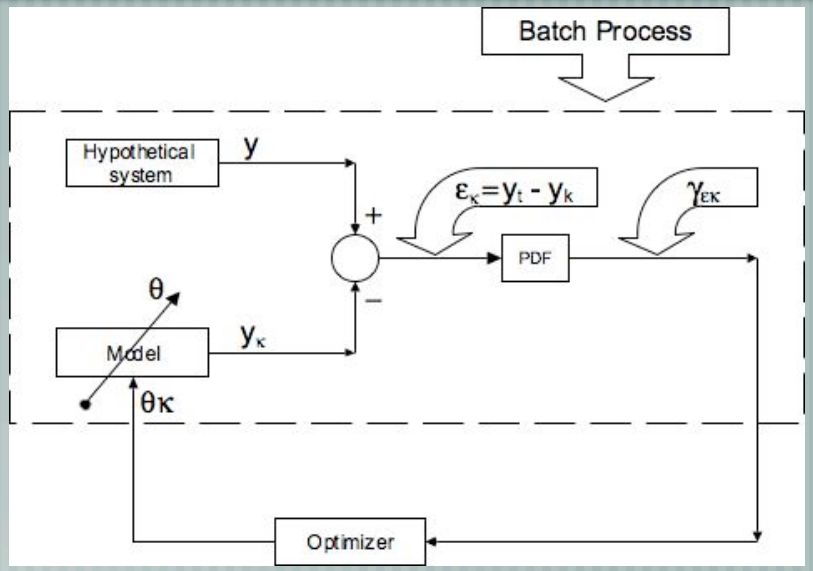
$$\frac{dS(t)}{dt} = -\theta_1 E(t)S(t) + \theta_2 ES(t)$$

$$\frac{dES(t)}{dt} = \theta_1 E(t)S(t) - \left(\theta_2 + \theta_3\right) ES(t)$$

$$\frac{dP(t)}{dt} = \theta_3 ES(t)$$

Introduction: Method Formulation

1. The schematic on the right depicts the basic configuration of the method.
2. The Entropy optimization variant is presented here.
3. Autonomous systems are considered.



$$J(\hat{\theta}) = argmin_{\hat{\theta}} \int_{\gamma_e \in [\alpha,\beta]} \cdots \int [\hat{\gamma}_e(\epsilon_1, \epsilon_2, \ldots, \epsilon_n, \hat{\theta}) log(\hat{\gamma}_e(\epsilon_1, \epsilon_2, \ldots, \epsilon_n, \hat{\theta}))$$

$\varepsilon_\varkappa = y - y_k \qquad t_k , k = 1,2,..$

Actual system

$$\dot{x} = f(x, \theta)$$

$$y = x + w$$

Model

$$\dot{\hat{x}} = f(x, \hat{\theta}_k)$$

$$y_k = \hat{x}$$

Kernel Density Estimation of the PDF

$$\hat{f}(\epsilon_k) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{\epsilon_k - \epsilon_k^{(i)}}{h}\right)$$

Using publicly available Toolbox

Why PDF Shaping?

1. Gene expression is *stochastic by nature* [2]:
    1. Consequence: gene regulatory networks as well as signal transduction networks follow a similar stochastic behavior.
    2. Moreover this stochasticity does not correspond to normally distributed events only [2], for example:
        1. **Protein randomization** due to mRNA fluctuations are of Poissonian nature [3].
        2. **Pareto-like distribution** in gene expression data sets examined in yeast, mouse and human cells [5].
        3. This will inevitably affect protein production as also studied in [4].
    3. Randomization of species concentration (since [<1000] molecules). Concentrations are now Random Variables.

Parameter Estimation Using PDF Shaping: Advantages

1. PDF shaping works directly with the joint PDF of the residual. Imposes less computational burden.
2. It does not require normally distributed noise component.
3. Direct access to the systems original parameter structure.
4. Kernel Density Estimation is used to approximate the residual PDF. Potentially express the stochastic properties of the residual as a superposition of Gaussian characteristics.

Model Justification

1. Define $P_n(t)$ as the probability of having n molecules at time t. We are interested in the joint event :

$$(\#S(t+\Delta t)=n, \#S(t)=m)$$

Expressing this in a probabilistic framework, addition law of probabilities:

$$P(\#S(t+\Delta t)=n)=\sum_m P(\#S(t+\Delta t)=n, \#S(t)=m)$$

and using the conditional law of probabilities:

$$P(\#S(t+\Delta t)=n, \#S(t)=m)=P(\#S(t)=m)\cdot P(\#S(t+\Delta t)=n|\#S(t)=m)$$

eventually leading to an expression of the probability $P_n(t)$ having m molecules at the preceding step.

$$P_n(t+\Delta t)=\sum_{m=1}^{\infty} p_{m,n}P_m(t)$$

and the corresponding expression for the ensemble average:

$$\langle S(t)\rangle=\sum_{n=1}^{\infty} n\cdot P_n(t)$$

Splitting the sum the expression becomes:

$$\langle S(t)\rangle=\sum_{n=1}^{m} n\cdot P_n(t)+\sum_{n=m+1}^{\infty} n\cdot P_n(t)$$

corresponding to the case where just m number of molecules are available:

$$S_m(t)=\langle S(t)\rangle - S_r(t)$$

2. Indicating that the assumed expression of the system:

$$\dot{x}=f(x,\theta,t)$$
$$y(t)=x(t)+w$$

where $y(t)$is the output concentration $S_m(t)$ $x(t)$ is the mean concentration $\langle S(t)\rangle$ and w the noise component assumed to be $S_r(t)$.

Modeling Assumptions

1. The system behaves as a “Well - Stirred” chemical reactor. Concentrations do not vary with space.
2. Available amount of molecules for each of the species is <1000 molecules. The Deterministic model cannot be used by itself.
3. The system is considered to be in a thermodynamic equilibrium. Resulting in a isochore and isothermal system in general.

Method Description: Assumptions

1. Bounded system in an interval $[\alpha,\beta]$.
2. The residual PDF γ_e is assumed to measurable and differentiable in all its arguments within $[\alpha,\beta]$.
3. The system is assumed to be identifiable in $[\alpha,\beta]$.
4. Formulation is based on Autonomous Systems.

Method Description: Formulation

1. The method employs Entropy minimization of the residual PDF.

$$J(\hat{\theta}) = argmin_{\hat{\theta}} \int_{\epsilon \in [\alpha,\beta]} \cdots \int [\hat{\gamma_e}(\epsilon_1, \epsilon_2, \ldots, \epsilon_n, \hat{\theta}) \log(\hat{\gamma_e}(\epsilon_1, \epsilon_2, \ldots, \epsilon_n, \hat{\theta})) d\epsilon_1 d\epsilon_2 \ldots d\epsilon_n$$

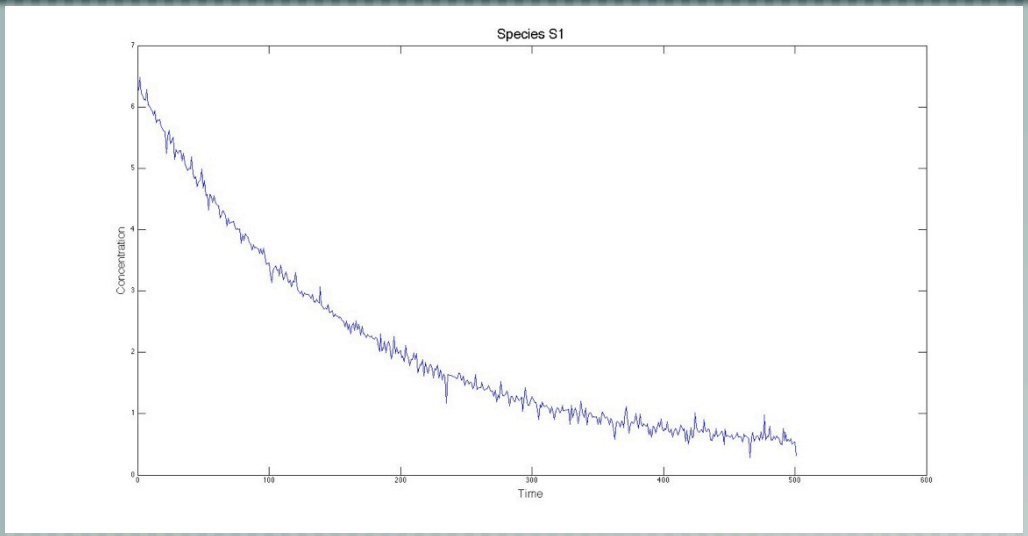
2. With the KDE estimation satisfying:

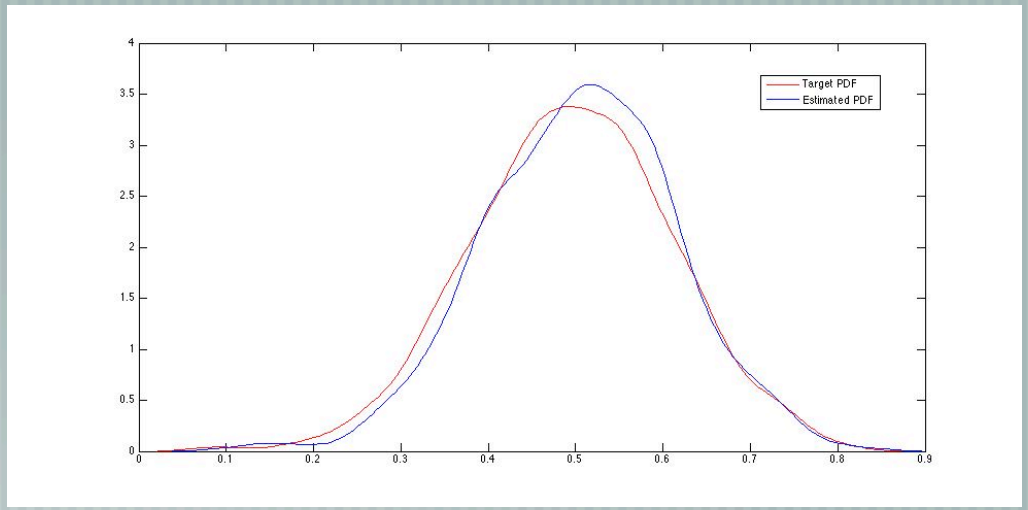
$$\int_{\gamma_e \in [\alpha,\beta]} \cdots \int \hat{\gamma_e}(\epsilon_1, \epsilon_2, \ldots, \epsilon_n, \hat{\theta}) = 1$$

Additional Notes

1. KDE was used for PDF estimation using a publicly available toolbox.
2. The examples include the implementation of the Entropy variant.
3. The univariate case was considered for better visualization of the process.

Examples: Simple Monomolecular Reaction Model

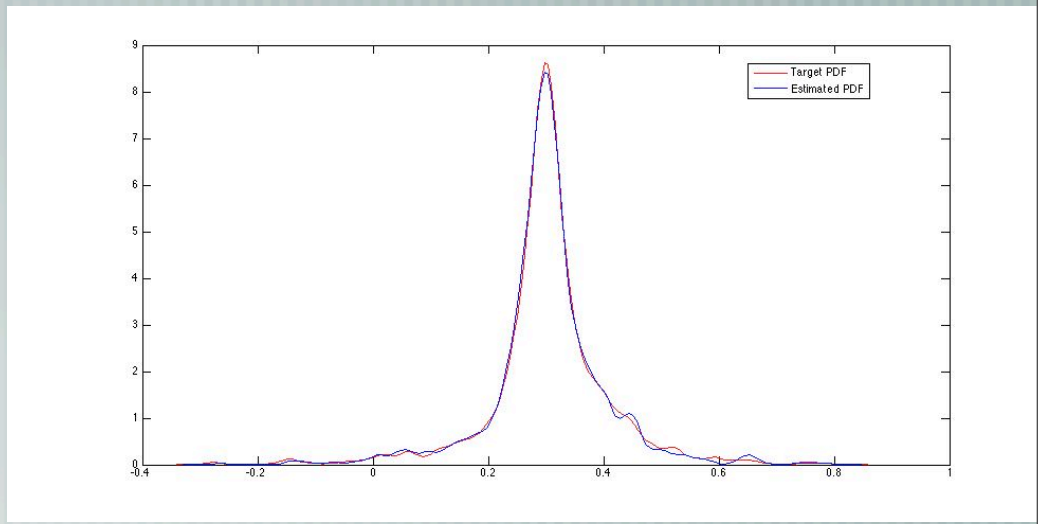




Simple Monomolecular Reaction

$$S_1 \xrightarrow{k} S_2 \quad \Longrightarrow \quad \frac{dS_1(t)}{dt} = -\theta_1 S_1(t)$$

1. The residual PDF is shaped so as to have minimum Entropy.
2. Two cases:
    1. Noise normally distributed.
    2. Noise having arbitrary distribution.



Conclusions

1. A *Stochastic Parameter Estimation framework* was presented where PDF Shaping [1] was used and the optimal parameter vector was estimated by *minimizing the entropy* of the residual vector.
2. The novelty of the method lies in the fact that the *noise* can now have an *arbitrary distribution*.
3. The method proves to be very robust even even when the *S/N ratio* is very low.

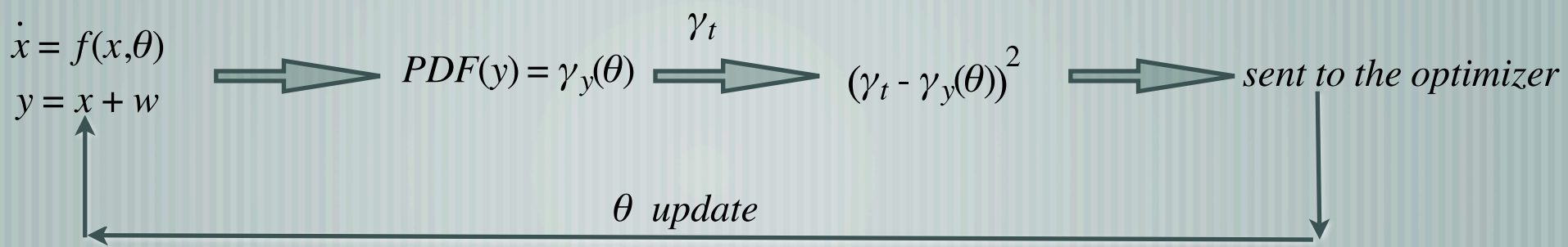
Future Research

1. Include a 1-norm complexity term on the optimization objective function:

    1. Introduce sparseness into the parameter space leading to a Sparse Stochastic Parameter Estimation framework where sparseness now has a probabilistic meaning.

$$J(\theta) = \text{argmin}_{\theta} \int_{\alpha}^{\beta} \ldots \int_{\alpha}^{\beta} [\gamma_e(\theta, y) - \gamma_t(y)]^2 dy + \lambda . \|\theta\|_1$$

2. Implement PDF shaping taking into on the states species concentration directly. The requirement is a target PDF and the biochemical pathway (Stochastic System) is now used.



References

1. Wang H. “Bounded Dynamic Stochastic Systems”, Modeling and Control, Springer Verlag,2000.
2. Thatai M,Van Oudenaaden A,”Intrinsic noise in gene regulatory networks”,Proceedings of the national academy of sciences of Unites States of America,vol 98 Jul 2001.
3. Salteli a, et al, “Sensitivity analysis in practice: a guide to assessing scientific models”, Wiley and sons, 2004.
4. McAdams A, Arkin A, “Its a noisy business: genetic regulation at the nanomolar scale”, Elsevier Science, vol 15, No 2, 1999.
5. Kusnetsov V A, Knott G A, Bonner R F, “General statistics of stochastic processes of gene expression inn Eukaryotic cells” Genetics society of America, vol 162, Jul 2002.
6. Elowitz M, Levine A, Siggia E, Swain P, “Stochastic gene expression in a single cell”, Science 297, 2002.
7. Brown M, Costen N P, “Exploratory Basis Pursuit Classification”, Pattern Recognition Letters, vol 26, No 12, 2005.
8. Papadopoulos G, Brown M, “Feature Sensitivity on Biochemical Signaling Pathways”, IEEE symposium on CIBCB, April 2007.