LECTURE #10

JOHN GOUTSIAS
WHITAKER BIOMEDICAL ENGINEERING INSTITUTE
THE JOHNS HOPKINS UNIVERSITY
BALTIMORE, MD 21218

STATISTICAL INFERENCE

Statistical Inference

- Statistical inference is the process of using data analysis techniques to estimate the values of unknown parameters associated with a Markovian reaction network using observations of the population process either at multiple time points (desirable but more difficult to obtain) or at steady-state (less desirable but easier to obtain).
- There are two fundamentally different types of parameters associated with a Markovian reaction network model:
  - the stoichiometric coefficients, which determine the structure of the network
  - the kinetic parameters, which determine the non-structural portion of the propensity functions.

https://en.wikipedia.org/wiki/Statistical_inference

Statistical Inference

- Some parameter values can be deduced experimentally or by using heuristic arguments.
- Most parameters however must be estimated from available data using statistical inference techniques.
- Since the predictive power of a given model fundamentally depends on the accuracy of its parameterization, inferring the unknown parameter values in a Markovian reaction network is a problem of key interest and practical importance.

Statistical Inference

- □ Although the problem of statistical inference has been extensively studied for reaction networks with deterministic dynamics, efficient inference of Markovian reaction networks is an open problem.
- □ Currently available methods have been primarily designed for biochemical reaction networks but can easily be adopted in other applications with little or no effort.
- □ These methods do not adequately address important issues, such as curse of dimensionality, thermodynamic consistency, and computational efficiency.

Statistical Inference

- It is quite common to assume known structural parameters and proceed with estimating the kinetic parameters using (most often noisy and sparse) measurements of system dynamics.
- This problem, known as model calibration, is much easier than the problem of estimating the structural parameters, which is often referred to as model selection.

Model Calibration

- The two most difficult issues associated with model calibration is the curse of dimensionality and the need for non-convex function optimization, which complicates numerical implementation.
- The curse of dimensionality refers to the problem of requiring an amount of data that dramatically increases with the dimensionality of the parameter space in order to ensure sufficiently acceptable parameter estimation.
- Consequently, the problem of finding good parameter values becomes difficult when the number of model parameters becomes large.
- This is further exacerbated by using non-convex optimization in order to find these values, which is a computationally difficult problem to solve in most cases of interest.

Model Calibration

- Therefore, development of statistical techniques for accurate and computationally efficient model calibration of Markovian reaction networks is an extremely challenging problem.
- Possible ways to attack this problem:
  - Effectively reduce the number of parameters that must be estimated by incorporating known constraints (such as thermodynamic constraints).
  - Identify a smaller set of "influential" parameters whose values must be estimated with sufficient precision.
- This is known as <u>dimensionality reduction</u> and must be combined with fast algorithms for solving the master equation, as well as with efficient optimization methods and appropriately designed experimental protocols for collecting data with high information content about the values of the unknown parameters.

Model Selection

- □ Model selection is a substantially more difficult problem.
- □ Solving this problem requires development of new statistical approaches for comparing between two competing network models (e.g., an originally proposed network and another network obtained by adding new reactions).
- □ This approach however requires that both models are calibrated before compared to each other, which adds to the difficulty of the problem.

Statistical Inference – SIR Model

- SIR model:
  - Susceptible (S) individuals.
  - Infected (I) individuals.
  - Resistant (R) individuals.

$$X_1 \leftrightarrow \mathrm{S}$$
$$X_2 \leftrightarrow \mathrm{I}$$
$$X_3 \leftrightarrow \mathrm{R}$$

- We have 3 species and 2 reactions:

$$\begin{aligned} X_1 + X_2 &\rightarrow 2X_2 \\ X_2 &\rightarrow X_3 \end{aligned}$$

$$\mathbf{S} = \begin{bmatrix} -1 & 0 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}$$

- Propensity functions:

$$\pi_1(x_1, x_2, x_3) = \kappa_1 x_1 x_2$$
$$\pi_2(x_1, x_2, x_3) = \kappa_2 x_2$$

$$\kappa_1 = 0.5$$
$$\kappa_2 = 0.1$$
$$X_1(0) = 99$$
$$X_2(0) = 1$$
$$X_3(0) = 0$$
$$T = 5 \text{ (max time)}$$

Statistical Inference – SIR Model

- Assume that we can observe the population process $\mathbf{x}(t)$ within a time interval $0 \le t \le T$.

- Since for the SIR model $\mathbf{S}^T\mathbf{S}$ is <u>invertible</u>, we can calculate the DA processes $z_1(t)$ and $z_2(t)$ from the population processes $x_1(t)$, $x_2(t)$, and $x_3(t)$, using

$$\begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} = \begin{bmatrix} -2/3 & 1/3 & 1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix} \begin{bmatrix} x_1(t) - 99 \\ x_2(t) - 1 \\ x_3(t) \end{bmatrix}$$

since

$$\mathbf{x}(t) = \mathbf{S}\mathbf{z}(t) + \mathbf{x}(0) \;\Leftrightarrow\; \mathbf{z}(t) = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T[\mathbf{x}(t) - \mathbf{x}(0)]$$

Statistical Inference – SIR Model

- □ Note that the DA processes can be uniquely determined from the random vector

$$\mathbf{V} = \begin{bmatrix} T_1 \\ M_1 \\ T_2 \\ M_2 \\ \vdots \\ T_K \\ M_K \end{bmatrix}$$

where $T_1, T_2, \ldots, T_K$ are the successive times of reaction firings and $M_1, M_2, \ldots, M_K$ are the reactions occurring at those times (all being random variables).

Statistical Inference – SIR Model

- Using this representation, we can compute estimates $\hat{\kappa}_1, \hat{\kappa}_2$ of κ_1, κ_2 by solving the following problem, known as maximum-likelihood (ML) estimation:

$$\{\hat{\kappa}_1, \hat{\kappa}_2\} = \arg\max_{\kappa_1, \kappa_2} \Pr[\mathbf{V} = \mathbf{v} \mid \kappa_1, \kappa_2]$$

where

$$\mathbf{V} = \begin{bmatrix} T_1 \\ M_1 \\ T_2 \\ M_2 \\ \vdots \\ T_K \\ M_K \end{bmatrix} \quad \text{and} \quad \mathbf{v} = \begin{bmatrix} t_1 \\ m_1 \\ t_2 \\ m_2 \\ \vdots \\ t_K \\ m_K \end{bmatrix}$$

https://en.wikipedia.org/wiki/Maximum_likelihood_estimation

Statistical Inference – SIR Model

- Using the results in Lecture #6, page 8, we can show that

$$
\begin{aligned}
&\Pr[\mathbf{V} = \mathbf{v} \mid \kappa_1, \kappa_2] \\
&\qquad = \prod_{k=1}^{K+1} \alpha_{m_k}(\mathbf{z}(t_{k-1})) \exp\left\{-(t_k - t_{k-1})[\alpha_1(\mathbf{z}(t_{k-1})) + \alpha_2(\mathbf{z}(t_{k-1}))]\right\} (t_k - t_{k-1})
\end{aligned}
$$

where $t_0 = 0$, $t_{K+1} = T$, and

$$
\begin{aligned}
\alpha_1(\mathbf{z}(t)) &= \kappa_1[99 - z_1(t)][z_1(t) - z_2(t) + 1] \\
\alpha_2(\mathbf{z}(t)) &= \kappa_2[z_1(t) - z_2(t) + 1]
\end{aligned}
$$

Statistical Inference – SIR Model

- This leads to the ML estimates $\hat{\kappa}_1$ and $\hat{\kappa}_2$ of the specific probability rate constants κ_1 and κ_2, given by:

$$\hat{\kappa}_1 = \frac{z_1(T)}{\sum_{k=1}^{K+1}(t_k - t_{k-1})[99 - z_1(t_{k-1})][z_1(t_{k-1}) - z_2(t_{k-1}) + 1]}$$

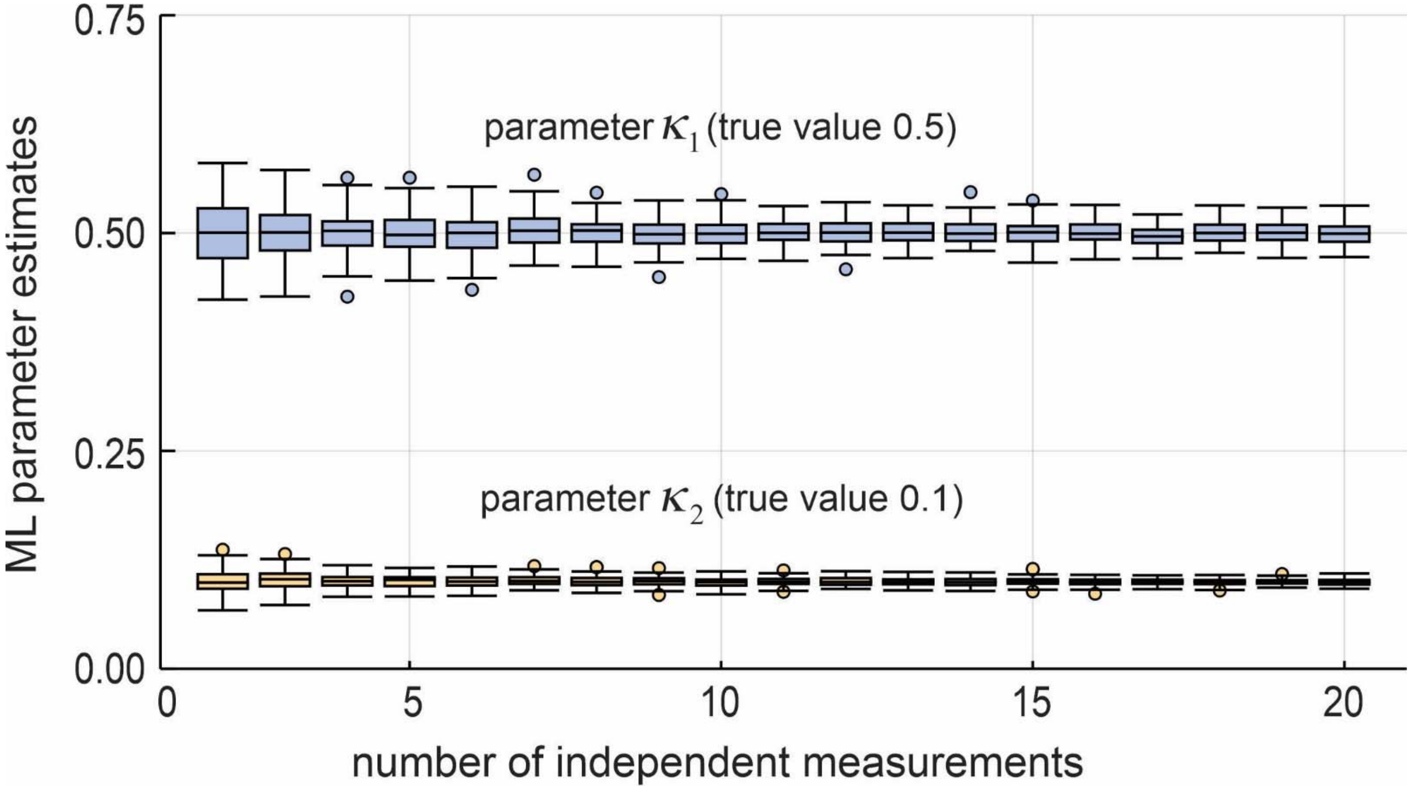
$$\hat{\kappa}_2 = \frac{z_2(T)}{\sum_{k=1}^{K+1}(t_k - t_{k-1})[z_1(t_{k-1}) - z_2(t_{k-1}) + 1]}$$

which shows that the ML estimate of κ_m is proportional to the total number $z_m(T)$ of occurrences of the m-th reaction within $[0,T]$.

- Similar formulas can be obtained when multiple independent measurements of the population processes are available.

Statistical Inference – SIR Model

- □ Box plots of ML estimated parameters (1000 times)



https://en.wikipedia.org/wiki/Box_plot

Remarks

- □ The previous example is not representative of typical statistical inference in Markovian reaction networks.
- □ The main reason is that <u>complete</u> observations of the population process are not available in most cases of interest.
- □ Moreover, the DA process cannot be computed from the population process, since the stoichiometry matrix is not invertible in general.
- □ These issues substantially complicate statistical inference, which must be addressed by developing more sophisticated approaches.