Bayesian Quantification of Uncertainty in Systems with Intrinsic Noise

KHACHIK SARGSYAN, COSMIN SAFTA, BERT DEBUSSCHERE, HABIB NAJM
Sandia National Laboratories, Livermore, CA

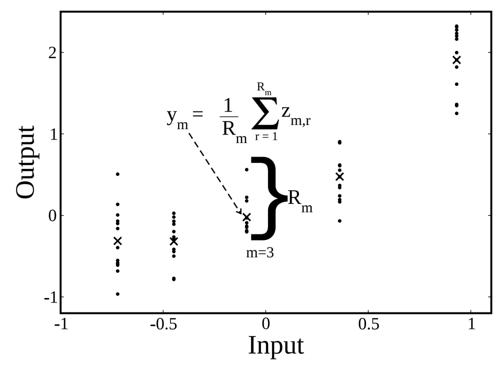
Stochasticity plays an important role in many phenomena

B. Subtilis endospore stain, by A. Schenkel, P. Justice and E. Suchman, Colorado State University.

- In stochastic reaction networks, intrinsic stochasticity is due to reactions between small number of molecules
- Applications
  - Gene regulatory networks, bioenergy and bioremediation
  - Interfacial reaction processes, fuel cells and batteries
  - Cellular signaling, immunology
- Uncertainty sources include intrinsic stochasticity, parametric uncertainty, sparsity of the available data, experimental noise.
- Questions that uncertainty quantification helps to answer
  - How predictive is the model?
  - If the model is good enough, what is the mismatch with experiments due to?
  - Does the system work in spite of the noise or because of it?

Problem formulation

- Stochastic model $Y(\boldsymbol{\lambda})$ with a d-dimensional input parameter vector $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)$
- Observable of interest $y = \mathbb{E}[Y]$
- *Training* runs at M input parameter values
- R_m replica runs for m-th input parameter
- A total of $N = \sum_{m=1}^{M} R_m$ model evaluations, $\{z_{m,r}\}$



Polynomial chaos spectral representation

To build a representation for input-output relationship, Polynomial Chaos (PC) spectral expansions are used; see Ghanem and Spanos, "Stochastic Finite Elements: A Spectral Approach",1991.

- Interprets input parameters as random variables
- Allows propagation of input parameter uncertainties to outputs of interest
- Serves as a computationally inexpensive surrogate for calibration or optimization

Input parameters are represented via their cumulative distribution function (CDF)

$$\eta_i = 2F_{\lambda_i}(\lambda_i) - 1, \qquad \text{for } i = 1, 2, \dots, d.$$

If input parameters are uniform $\lambda_i \sim \text{Uniform}[a_i, b_i]$, then

$$\eta_i = \frac{2}{b_i - a_i}\left(\lambda_i - \frac{a_i + b_i}{2}\right).$$

Output is represented with respect to Legendre polynomials

$$y(\boldsymbol{\eta}) \approx y_{\boldsymbol{c}}(\boldsymbol{\eta}) \equiv \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\eta}).$$

Sparse quadrature intergation fails with noisy data

Using orthogonality of the basis functions

$$\langle \Psi_i(\boldsymbol{\eta})\Psi_j(\boldsymbol{\eta})\rangle = \delta_{ij}\langle \Psi_i(\boldsymbol{\eta})^2\rangle,$$

one can compute *PC modes* via projection

$$c_k = \frac{\langle y\Psi_k(\boldsymbol{\eta})\rangle}{\langle \Psi_k^2(\boldsymbol{\eta})\rangle} = \frac{1}{2^d\langle \Psi_k^2(\boldsymbol{\eta})\rangle}\int_{[-1,1]^d} y(\boldsymbol{\eta})\Psi_k(\boldsymbol{\eta})d\boldsymbol{\eta}$$

Monte-Carlo estimation of the above integral converges slowly.

Quadrature approaches fail as well.

$$\sum_{q=1}^{Q} y_q \Psi_k(\boldsymbol{\eta}_q) w_q$$

- *Tensor product* quadrature suffers from the curse of dimensionality
- *Sparse grid* quadrature is infeasible for noisy systems due to negative weights. Even a very small error in function evaluation is amplified by a factor that increases with dimensionality!

Bayesian inference of PC modes

Bayesian framework allows quantifying different sources of uncertainties - parametric, intrinsic, or uncertainties associated with lack-of-sampling.

Bayes formula

$$p(\boldsymbol{c}|D) \propto L_{\mathcal{D}}(\boldsymbol{c})p(\boldsymbol{c})$$

relates prior distribution $p(\boldsymbol{c})$ of PC modes to the posterior $p(\boldsymbol{c}|\mathcal{D})$, where the data $\mathcal{D}$ is the set of all training runs $\{z_{m,r}\}$, $m = 1:M$, $r = 1:R_m$.

Estimates of the mean of the data $z_{m,r}$ and its variance at the m-th parameter location are, respectively,

$$y_m = \frac{1}{R_m}\sum_{r=1}^{R_m} z_{m,r},$$

$$s_m^2 = \frac{1}{R_m - 1}\sum_{r=1}^{R_m}(z_{m,r} - y_m)^2.$$

Prior distribution on $\boldsymbol{c}$ is uniform, $p(\boldsymbol{c}) = \text{const}$.

The likelihood accounts for the discrepancy between the averaged data and the model,

$$L_{\mathcal{D}}(\boldsymbol{c}) = L_{\mathcal{D}}(\boldsymbol{c}; \boldsymbol{s}^2) = \frac{1}{(2\pi)^{M/2}\prod_{m=1}^{M}(s_m/\sqrt{R_m})}\exp\left(-\sum_{m=1}^{M}\frac{(y_m - y_{\boldsymbol{c}}(\boldsymbol{\eta}_m))^2}{2s_m^2/R_m}\right)$$

The posterior is analytically tractable, it is a multivariate normal distribution,

$$\boldsymbol{c} \in \mathcal{MVN}(\underbrace{(\boldsymbol{\Psi}^T\boldsymbol{Q}^{-1}\boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}^T\boldsymbol{Q}^{-1}\boldsymbol{y}}_{\text{mean}}, \underbrace{(\boldsymbol{\Psi}^T\boldsymbol{Q}^{-1}\boldsymbol{\Psi})^{-1}}_{\text{covariance}}),$$

where $\boldsymbol{\Psi}$ is a $M \times (K+1)$ matrix with elements $\boldsymbol{\Psi}_{mk} = \Psi_k(\boldsymbol{\eta}_m)$ and $\boldsymbol{Q}$ is a diagonal weight matrix with entries $Q_{mm'} = \delta_{m,m'}R_m/(2s_m^2)$.

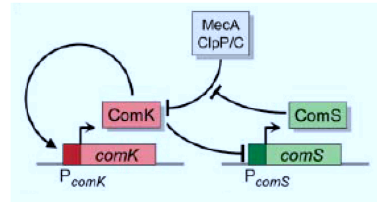
Mixture PC expansion based on nearest neighbor classification

- If data has quantitatively different behavior in different regions, global polynomial fit is inaccurate
- A mixture PC formulation is developed based on a nearest neighbor classification
  - The input set of points is clustered according to the corresponding output values
  - For each cluster, a separate PC expansion is obtained
  - The resulting expansion is a weighted sum of PC expansions for a certain number of nearest neighbors
- If the output values are bounded, a map to $(-\infty; +\infty)$ is utilized before PC representation to keep the approximation from exceeding physical bounds
  - For example, if $y \in [0, 1]$, the effective output is $\tilde{y} = \log\frac{y}{1-y}$

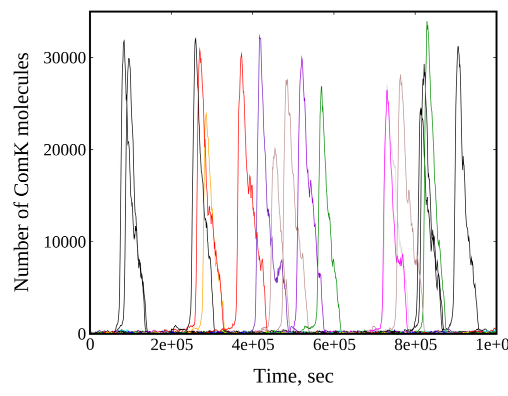
Sargsyan *et al.*, "Multiparameter spectral representation of noise-induced competence in Bacillus Subtilis", to be submitted to *Biophys J*, 2011.

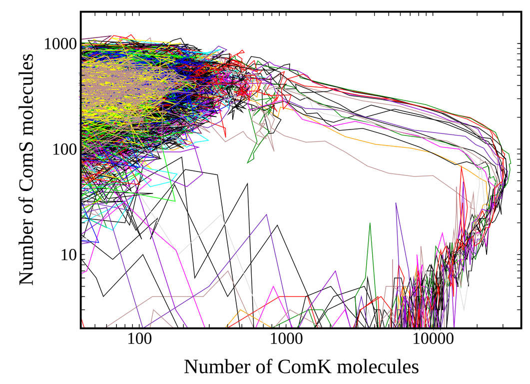
Bacillus Subtilis is a gram positive soil bacterium

- Competence in *B. Subtilis* is a state that allows uptake of external DNA
- It is characterized by a sporadic jump in the number of comK molecules
- Stochastic reaction network of competence dynamics consists of 11 species and 16 reactions, see Süel *et al.*, Science, 2007
- Input parameters are reaction rate parameters in logarithmic scale, $\boldsymbol{\eta} = \log\tilde{k} \pm \log f$, *i.e.* the range is $[\tilde{k}/f, \tilde{k}f]$ with a range factor $f > 1$ and a nominal parameter value $\tilde{k}$.



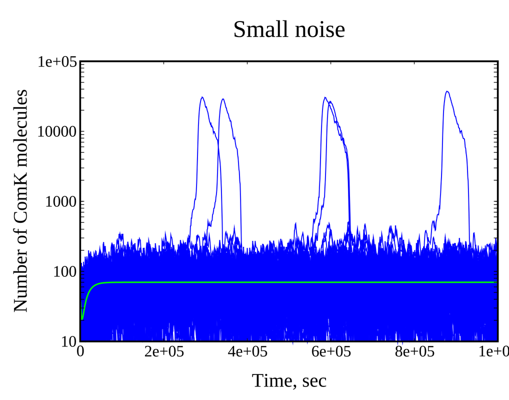
Süel et al., Science, 2007

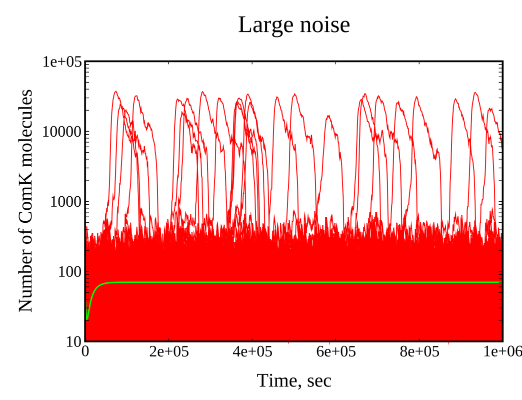




ODE limit and noise-induced transition to competence

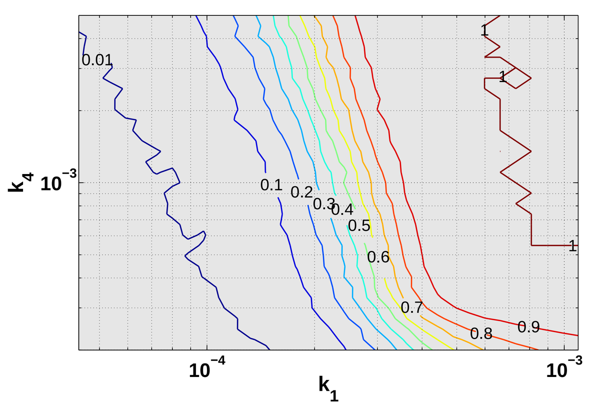
- Competence events, *i.e.* sporadic jumps in the number of comK molecules, are driven by noise
- In the limit of large volume, the system is described by a system of ODEs, called rate equations
- By tuning reaction network parameters in a special way, one can keep the corresponding ODE limit unchanged, focusing on pure noise dependence



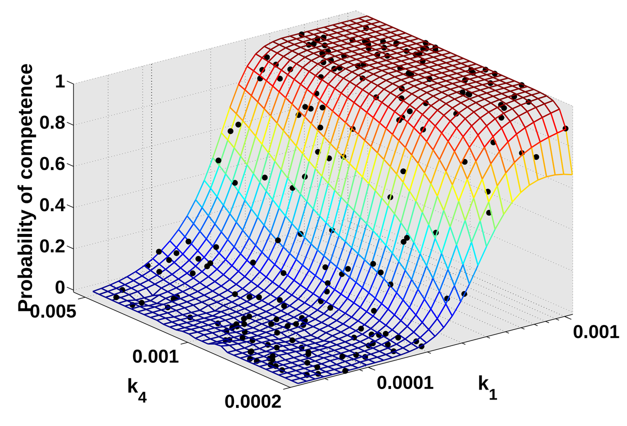


Two-dimensional study

- An output observable is the fraction of time, in steady state, the system spends in competence, i.e. $P_c = P(X_\infty > 5000)$. Note that $P_c \in [0, 1]$ by definition, and the map $\log\frac{P_c}{1-P_c}$ is employed
- Some regions in input space lead to a fully vegetative ($P_c = 0$) or a fully competent ($P_c = 1$) state
- Clustering approach fits a constant (0 or 1) in these trivial regions



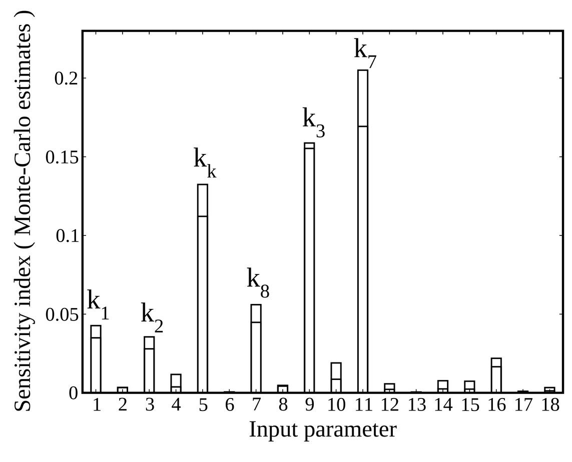
Contour plots of the probability of competence



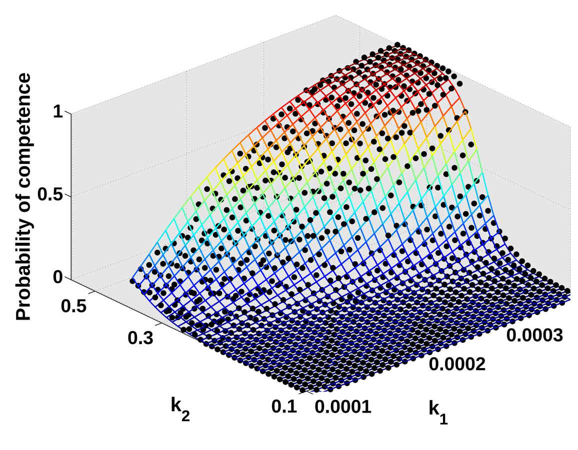
4-th order mean response surface

Dimensionality reduction using sensitivity indices

- All 18 reaction rate parameters are taken
- Due to sparsity of the data ($M = 1000$ points in 18-dimensional parameter space) the global PC expansion is more reliable than the clustering-based mixture PC
- Variance-based sensitivity indices $S_i = \frac{\text{Var}[E(y_{\boldsymbol{C}}(\boldsymbol{\eta})|\eta_i)]}{\text{Var}[y_{\boldsymbol{C}}(\boldsymbol{\eta})]}$ are computed from the global PC to down-select from 18 dimensions to 6 dimensions
- The comK-related reaction parameters have shown larger sensitivity indices
- For the resulting 6-dimensional problem, a mixture PC is constructed and shown to be more accurate
- For each of the $M = 1000$ input parameters, $R_m = 100$ replica simulations are taken
- The resulting *uncertain* response surface has a relative L_2 error of ~ 0.08



Sensitivity indices



4-th order mean response surface

This work was supported by the U.S. Department of Energy Office of Science through the Applied Mathematics program in the Office of Advanced Scientific Computing Research (ASCR) under contract 07-012783 with Sandia National Laboratories.

Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract No. DE-AC04-94AL85000.

Sandia National Laboratories