Uncertainty Propagation Using An Infinite Mixture of Gaussian Processes and Variational Bayesian Inference

Prof. Nicholas Zabaras
Warwick Centre for Predictive Modelling
The University of Warwick
nzabaras@gmail.com
http://www2.warwick.ac.uk/wcpm/
http://www.zabaras.com

Joint Work with: Dr. Peng Chen (CGG) and Prof. Ilias Bilionis (Purdue)
Introduction

Reservoir simulation (secondary production by engineered drive)

Limitations of current UQ studies:

- Limited measurements of the reservoir properties
- Expensive simulation tools
- High dimensional stochastic space
- Multimodal behavior and discontinuities in stochastic space
Review of Current Approaches

Spectral finite element method:
- Generalized polynomial chaos [Xiu & Karniadakis, 2002]
- Multi-element generalized polynomial chaos [Wan & Karniadakis, 2005]

Stochastic Collocation:
- Probabilistic Collocation [Li & Zhang, 2010]
- Adaptive Sparse Grid Collocation [Ma & Zabaras, 2009]

Markov Chain Monte Carlo:
- Two-stage MCMC [Ginting et al., 2013]
- Continuous reservoir simulation via MCMC [Liu, 2009]

Probabilistic graphical models:
- Nonparametric probabilistic graphical model [Chen & Zabaras, 2013]
- Gaussian probabilistic graphical model [Wan & Zabaras, 2013]

Gaussian Process:
- Multi-output Gaussian Process [Bilionis & Zabaras, 2012]
**Difficulties**

**Challenges of UQ problem:**

- Expensive deterministic solver
- High dimensional Data
- Multi-modal behavior and discontinuities
- Complex target posterior distributions

**Solutions:**

- Surrogate Model
- Decomposition of space
- Mixture Model
- Variational Inference
Problem Definition

Consider a general physical problem

Input: \( \mathbf{x} = (\xi, s, t) \in \mathbb{R}^d \) \( \xrightarrow{f : X \rightarrow \mathbb{R}^q} \) Output: \( \mathbf{y} \in \mathbb{R}^q \)

\( \xi \in X_{\xi} \subset \mathbb{R}^{d_{\xi}} \) \( \mathbf{x}_s \subset \mathbb{R}^{d_s} \) \( \mathbb{X}_t = [0, T] \)

stochastic input \hspace{1cm} \text{spatial points} \hspace{1cm} \text{time steps}

Denote the training data \( \mathcal{D} = \{\mathbf{X}, \mathbf{Y}\} \)

\[ \mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times d} \quad \mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^T \in \mathbb{R}^{N \times q} \]

latent indicator variable \( \mathbf{z} = [z_1, \ldots, z_N]^T \in \mathbb{R}^N \)

\( \mathbf{X} \xrightarrow{\mathbf{f}^{(k)}(\mathbf{x})} \mathbf{y} \)

\( K \) unknown

Multi-output Gaussian Process [Conti & O'Hagan, 2010]

Dirichlet Process [Ferguson, 1973]

\( N = n_{\xi} n_{s} n_{t} \)
Objectives of the UQ Problem

Computing the statistics of the system response

Mean:

\[ m_f(s, t) = \int f(\xi, s, t) p(\xi) \, d\xi, \]

Covariance:

\[ C_f(s, t, s', t') = \int (f(\xi, s, t) - m_f(s, t)) (f(\xi, s', t') - m_f(s', t'))^T p(\xi) \, d\xi, \]

PDF of the output at a particular spatial location \( s_0 \) and time instant \( t_0 \) :

\[ p(y|s = s_0, t = t_0) = \int \delta(f(\xi, s_0, t_0) - y) \, p(\xi) \, d\xi, \]
Multi-output Gaussian Process Regression

Model the $m$th latent function\(^1\) as a $q$-dim GP (before seeing any data)

$$f(\cdot) \mid B, \Sigma, \theta \sim \mathcal{N}_q(f(\cdot) \mid \mu(\cdot; B), c(\cdot, \cdot; \theta) \Sigma)$$

conditional on hyper-parameters $\phi = (B, \Sigma, \theta)$

$$\mu(x; B) = B^T h(x)$$

$h(\cdot) = (1, x^T)$

The covariance function is taken to be

$$c(x, x'; \theta) = \exp \left\{ -\frac{1}{2} \sum_{l=1}^{d} \left( \frac{(x_l - x'_l)^2}{(r_l)^2} \right) \right\} + \epsilon \delta(x - x')$$

Nugget: Variance of the noise

$$\theta = (r_1, \ldots, r_d, \epsilon) \in \mathbb{R}^{d+1}_{+}$$

Given $D_m = \{X_m, Y_m\}$, the likelihood follows a matrix-normal distribution

$$Y \mid X, \phi \sim \mathcal{N}_{p \times q}(Y \mid HB, A, \Sigma)$$

\(^1\)To keep the notation simple, we ignore the label $m$

[Conti & O’Hagan, 2010]
Multi-output Gaussian Process Regression

- Prior selection:
  \[ p(B, \Sigma) \propto |\Sigma|^{-\frac{q+1}{2}} \quad , \quad p(\theta | \gamma) = \left[ \prod_{i=1}^{d} \mathcal{E}(r_i | \gamma_r) \right] \mathcal{E}(\varepsilon | \gamma_\varepsilon) \]

- The posterior distribution:
  \[ f(\cdot) | \phi \sim \mathcal{N}_q (f(\cdot) | \mu^* (\cdot; B), c^* (\cdot, \cdot; \theta) \Sigma) \]

where
\[ \mu^* (x; B) = B^T h(x) + (Y - HB)^T A^{-1} a(x) \]
\[ c^* (x, x'; \theta) = c(x, x'; \theta) - a(x)^T A^{-1} a(x') \]

\( B^{(m)} \) and \( \Sigma^{(m)} \) can be integrated out as
\[ f(\cdot) | \theta, \mathcal{D} \sim \mathcal{T}_q (f(\cdot) | \mu^{**} (\cdot; \widehat{B}), c^{**} (\cdot, \cdot; \theta) \widehat{\Sigma}; n-p) \]
\[ \mu^{**} (x) = \mu^* (x; \widehat{B}) \]
\[ c^{**} (x, x'; \theta) = c^*(x, x'; \theta) + \left( h(x) - H^T A^{-1} a(x) \right)^T (H^T A^{-1} H)^{-1} \left( h(x') - H^T A^{-1} a(x') \right) \]

Mean of the posterior

\[ a(\cdot) = (c(\cdot, x_1; \theta), \ldots, c(\cdot, x_n; \theta)) \in \mathbb{R}^n \]
A DP defines a probability measure on a space of probability measures.

\[ \mathcal{P}(\Omega, \mathcal{F}) = \{G : \mathcal{F} \to \mathbb{R}_+ \text{ is a probability measure}\} \]

A sample \( G(\cdot) \) from a DP is in \( \mathcal{P}(\Omega, \mathcal{F}) \).

Let \( G_0(\cdot) \in \mathcal{P}(\Omega, \mathcal{F}) \) and \( \alpha_0 > 0 \) some constant base distribution.

\[ G(\cdot) \sim \text{DP}(G_0(\cdot), \alpha_0) \]

scaling factor

if and only if for any partition \( \{A_1, \ldots, A_k\} \subset \mathcal{F} \) of \( \Omega \), we have:

\[ (G(A_1), \ldots, G(A_k)) \sim \text{Dir}_k(G(A_1), \ldots, G(A_k) | \alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_k)) \]
The “Stick-Breaking” DP Construction

- The “stick-breaking” construction allows us to define the DP in a generative way by the use of some intermediate random variables.

In particular, define the random variables \( v = (v_1, v_2, \ldots) \in [0, 1]^\infty \) by

\[
v|\alpha_0 \sim \prod_{m=1}^{\infty} \text{Beta}(v_m|1, \alpha_0)
\]

the sequence of numbers:

\[
\pi_m(v) = v_m \prod_{i=1}^{m-1} (1 - v_i),
\]

and the random variables \( \omega = (\omega_1, \omega_2, \ldots) \)

\[
\omega|G_0(\cdot) \sim \prod_{m=1}^{\infty} G_0(\omega_m).
\]

Then,

\[
G(\cdot; v, \omega) = \sum_{m=1}^{\infty} \pi_m(v) \delta_{\omega_m}(\cdot).
\]

[\text{Ferguson, 1973}]
Each MGP model characterized by $\phi \in \Omega_\phi$

The role of this DP is to generate the components of the MGP mixture prior to observing any data.

Let $\mathcal{F}_\phi$ be a $\sigma$-algebra on $\Omega_\phi$, and $P_{\phi,0}(\cdot | \gamma)$ be the probability measure on $(\Omega_\phi, \mathcal{F}_\phi)$ induced by the prior $p(\phi | \gamma)$.

$$P_{\phi,0}(A | \gamma) = \int_A p(\phi | \gamma) d\phi.$$ 

$\mathcal{D}P(\cdot | P_{\phi,0}(\cdot ; \gamma), \alpha_0)$ prior on $\mathcal{P}(\Omega_\phi, \mathcal{F}_\phi)$.

$P_{\phi}(\cdot) \sim \mathcal{D}P(\cdot | P_{\phi,0}(\cdot ; \gamma), \alpha_0)$ generate the mixture components.

$$\Phi | \gamma \sim \prod_{m=1}^{\infty} P_{\phi,0}(\phi_m | \gamma) = \prod_{m=1}^{\infty} p(\phi_m | \gamma).$$

$$P_{\phi}(\cdot ; \nu, \Phi) = \sum_{m=1}^{\infty} \pi_m(\nu) \delta_{\phi_m}(\cdot).$$
For N observations, we thus introduce indicator variables $z_i \in \mathbb{N}$,

$$z = (z_1, \ldots, z_N).$$

The prior assigned to $z$

$$p(z|\nu) = \prod_{n=1}^{N} \sum_{m=1}^{\infty} \pi_m(\nu)^{1[z_n=m]}$$

$$= \prod_{i=1}^{N} Multi(z_i|\pi_1(\nu),\pi_2(\nu),\ldots) = \prod_{n=1}^{N} \pi_{z_n}(\nu).$$

We consider the indicator vector to be dependent on the input:

$$p(z|X, m, R, \nu) = \prod_{n=1}^{N} \prod_{m=1}^{\infty} \left( \frac{p(x_n|m_m, R_m)\pi_m(\nu)}{\sum_{j=1}^{\infty} p(x_n|m_j, R_j)\pi_j(\nu)} \right)^{1[z_n=m]}$$

Conjugate Priors

and

$$p(x_n|z_n = m, m_m, R_m) = \mathcal{N}_d(x_n|m_m, R_m^{-1}).$$

$m_m = \mathcal{N}_d(u_0, R_0^{-1})$

$R_m = W_d(W_0, \nu_0)$
The likelihood of the full model can be described as:

\[ p(Y, z|X, \Phi) = p(Y|X, z, \Phi) p(z|X, m, R, \nu) \]

where \( p(Y|X, z, \Phi) = \prod_{m: |z_n = m| > 0} p(S_m(Y; z)|S_m(X; z), \phi_m) \)

The prior of the full model is:

\[ p(\Phi, \nu, m, R|\mathcal{I}) = p(\Phi|\gamma) p(\nu|\alpha_0) p(m|u_0, R_0) p(R|W_0, \nu_0) \]

\[ \mathcal{I} = (\gamma, \alpha_0, u_0, R_0, W_0, \nu_0). \]

The posterior of the full model is:

\[ p(\Phi, z, \nu, m, R|\mathcal{D}, \mathcal{I}) = \frac{p(Y|X, z, \Phi) p(z|X, m, R, \nu) p(\Phi, \nu, m, R|\mathcal{I})}{p(\mathcal{D}|\mathcal{I})} \]

where the evidence is:

\[ p(\mathcal{D}|\mathcal{I}) = \sum_z \int p(Y|X, z, \Phi) p(z|X, m, R, \nu) p(\Phi, \nu, m, R|\mathcal{I}) d\Phi d\nu dm dR. \]
The joint distribution is represented as

\[
p(D, \Psi | I) = p(Y | X, z, \Phi) p(z | X, m, R, v) p(\Phi, v, m, R | I)
\]

\[
= \prod_{n=1}^{\infty} \left\{ p(S_m(Y; z) | S_m(X; z), \phi_m) \prod_{n=1}^{N} \left( \frac{p(x_n | m_m, R_m) \pi_m(v)}{\sum_{j=1}^{\infty} p(x_n | m_j, R_j) \pi_j(v)} \right) 1[z_n=m] \right\} p(\Phi, v, m, R | I)
\]

\[
= \prod_{n=1}^{\infty} \left\{ p(S_m(Y; z) | S_m(X; z), \phi_m) \prod_{n=1}^{N} \left( \frac{p(x_n | m_m, R_m) \pi_m(v)}{C_n(m, R, v)} \right) 1[z_n=m] \right\} p(\Phi, v, m, R | I),
\]

where we define

\[C_n(m, R, v) = \sum_{j=1}^{\infty} p(x_n | m_j, R_j) \pi_j(v).\]

However, \(C_n(m, R, v)\) is hard to explicitly calculate.

Thus, we follow the spirit of EM algorithm by using \(\hat{C}_n(\hat{m}, \hat{R}, \hat{v})\) to approximate \(C_n(m, R, v)\), where \(\hat{m}, \hat{R},\) and \(\hat{v}\) are the means of the corresponding posterior distributions.
Probabilistic Graphical Model Representation
Variational Inference

Denote $\Psi = \{z, \nu, m, R, \Phi\}$

Variational inference makes a proposal of the distribution of interest by factorization and finds the optimal proposal by maximizing the lower bound of model evidence $p(D|I)$

$$\ln p(D|I) = KL[q \parallel p] + \mathcal{L}[q]$$

Where the lower bound is

$$\mathcal{L}[q] = \int q(\Psi) \ln \frac{p(D, \Psi|I)}{q(\Psi)} d\Psi$$

Proposed distributions:

$$q(\Psi) = \prod_{n=1}^{N} q(z_n) \prod_{m=1}^{M} q(\nu_m) \prod_{n=1}^{M} q(m_m)q(R_m) \prod_{m=1}^{M} q(B_m, \Sigma_m) q(\theta_m)$$

[Blei & Jordan, 2006]
For a general case, assume $q(\Psi) = \prod_k q_k(\omega_k)$

Denote $q_k(\omega_k) = q_k$, the lower bound can be written as

$$
\mathcal{L}(q, \mathcal{D}) = \int \prod_k q_k \left[ \ln p(\Psi, \mathcal{D}) - \sum_k \ln q_k \right] d\Psi
$$

$$
= \int \prod_k q_k \ln p(\Psi, \mathcal{D}) \prod_k d\omega_k - \sum_k \int \prod_j q_j \ln q_k d\omega_j
$$

$$
= \int q_j \left[ \ln p(\Psi, \mathcal{D}) \prod (q_k d\omega_k) \right] d\omega_j
$$

$\qquad q(\Psi) = \prod_k q_k(\omega_k)$

$$
- \int q_j \ln q_j d\omega_j - \sum_{k \neq j} \int q_k \ln q_k d\omega_k
$$

$$
= \int q_j \ln \frac{\exp \mathbb{E}_{F_\psi \omega_j} [\ln p(\Psi, \mathcal{D})]}{q_j} d\omega_j - \sum_{k \neq j} \int q_k \ln q_k d\omega_k
$$

$$
= -\text{KL}(q_j \parallel \exp \mathbb{E}_{F_\psi \omega_j} [\ln p(\Psi, \mathcal{D})]) - \sum_{k \neq j} \int q_k \ln q_k d\omega_k.
$$

Lower bound is maximized when KL distance goes to zero

Update rule

$$
\ln q(\omega) = \mathbb{E}_{F_\psi \omega} [\ln p(\Psi, \mathcal{D} | \mathcal{I})] + \text{const}
$$

[Bishop, 2006]
Update of $q(\nu_m)$

To update $\nu_m$, we throw any terms that are independent of $\nu_m$ in the joint distribution. We have

$$\ln q(\nu_m) = \ln p(\nu_m|\alpha_0) + \sum_{n=1}^{N} \mathbb{E}_{F_\psi \setminus \nu_m}[\ln p(z_n|X, m, R, \nu)] + \text{const}$$

where

$$\mathbb{E}_{F_\psi \setminus \nu_m}[\ln p(z_n|X, m, R, \nu)] = \mathbb{E}_{F_\psi \setminus \nu_m} \left[ \ln \left( \prod_{l=1}^{M} \left( \frac{p(x_l|m_m, R_m)}{C_n(m, R, \nu)} \right)^{v_l} \prod_{j=1}^{l-1} \left( 1 - v_j \right)^{1[z_n = j]} \right) \right]$$

$$= \mathbb{E}_{F_\psi \setminus \nu_m} \left[ q(z_n = m) \ln \nu_m + \sum_{i=m+1}^{M} q(z_n = i) \ln (1 - \nu_m) \right] + \text{const}$$

$$= \mathbb{E}_{z_n}[q(z_n = m)] \ln \nu_m + \mathbb{E}_{z_n}[q(z_n > m)] \ln (1 - \nu_m) + \text{const}.$$ 

$$\ln q(\nu_m) = (\alpha_0 - 1) \ln (1 - \nu_m) + \sum_{n=1}^{N} \left[ \mathbb{E}_{z_n}[q(z_n = m)] \ln \nu_m + \mathbb{E}_{z_n}[q(z_n > m)] \ln (1 - \nu_m) \right] + \text{const}$$

$$= \left[ \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n = m)] \right] \ln \nu_m + \left[ \alpha_0 - 1 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n > m)] \right] \ln (1 - \nu_m) + \text{const}.$$

$$\nu_m \sim \text{Beta} \left( 1 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n = m)], \alpha_0 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n > m)] \right)$$
Update of $q(m_m)$

To update $m_m$, we throw any terms that are independent of $m_m$ in the joint distribution, we have

$$\ln q(m_m) = \ln p(m_m | u_0, R_0) + \sum_{n=1}^{N} \mathbb{E}_{F \setminus m_m} [\ln p(z_n | X, m, R, \nu)] + \text{const}$$

The second term can be calculated as:

$$\mathbb{E}_{F \setminus m_m} [\ln p(z_n | X, m, R, \nu)] = \mathbb{E}_{F \setminus m_m} \left[ \ln \left( \prod_{i=1}^{M} \left( \frac{p(x_n_i | m_i, R_i) \pi_i(\nu)}{C_n(\bar{m}, \bar{R}, \bar{\nu})} \right) 1[z_n = i] \right) \right]$$

$$= \mathbb{E}_{F \setminus m_m} \left[ \ln \left( \prod_{i=1}^{M} p(x_n_i | m_i, R_i) 1[z_n = i] \right) \right] + \text{const}$$

$$= \sum_{j=1}^{M} \mathbb{E}_{z_n}[q(z_n = j)] \mathbb{E}_{F \setminus m_m} [\ln p(x_n_i | m_i, R_i)]$$

$$= \mathbb{E}_{z_n}[q(z_n = m)] \mathbb{E}_{F \setminus m_m} [\ln p(x_n | m_m, R_m)]$$

$$= \mathbb{E}_{z_n}[q(z_n = m)] \left( -\frac{1}{2} (x_n - m_m)^T \mathbb{E}_{R_m}[R_m] (x_n - m_m) \right) + \text{const.}$$

denote $R_{m1} = \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n = m)] \mathbb{E}_{R_m}[R_m]$ and $R_{m2} = \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n = m)] \mathbb{E}_{R_m}[R_m] x_n$.

$m_m \sim \mathcal{N}_d \left( u_m, (R_0 + R_{m1})^{-1} \right)$
Update of $q(R_m)$

- To update $R_m$, we throw any terms that are independent of $R_m$ in the joint distribution. We have

$$
\ln q(R_m) = \ln p(R_m | W_0, v_0) + \sum_{n=1}^{N} \mathbb{E}_{F_\psi \setminus R_m} \left[ \ln p(z_n | X, m, R, \nu) \right] + \text{const}
$$

where

$$
\mathbb{E}_{F_\psi \setminus R_m} \left[ \ln p(z_n | X, m, R, \nu) \right] \\
= \mathbb{E}_{F_\psi \setminus R_m} \left[ \ln \left( \prod_{i=1}^{M} \left( \frac{p(x_n | m_i, R_i) \pi_i(\nu)}{C_n(m, R, \nu)} \right)^{1[z_n=i]} \right) \right] \\
= \mathbb{E}_{F_\psi \setminus R_m} \left[ \ln \left( \prod_{i=1}^{M} p(x_n | m_i, R_i)^{1[z_n=i]} \right) \right] + \text{const} \\
= \mathbb{E}_{z_n} [q(z_n = m)] \mathbb{E}_{m_m} \left[ \ln p(x_n | m_m, R_m) \right] + \text{const} \\
= \frac{1}{2} \mathbb{E}_{z_n} [q(z_n = m)] \left( \ln |R_m| - \mathbb{E}_{m_m} \left[ (x_n - m_m)^T R_m (x_n - m_m) \right] \right) + \text{const.}
$$

$R_m \sim W_d(W_m, u_m)$

$$
u_m = \sum_{n=1}^{N} \mathbb{E}_{z_n} [q(z_n = m)] + v_0,$$

$$W_m^{-1} = W_0^{-1} + \sum_{n=1}^{N} \mathbb{E}_{z_n} [q(z_n = m)] \mathbb{E}_{m_m} \left[ (x_n - m_m)(x_n - m_m)^T \right]$$
Similarly, to update $q(z_n)$, we throw any terms that are independent of $z_n$

\[
\ln q(z_n) = \mathbb{E}_{F_{\theta} \mid z_n} \left[ \ln p(z_n \mid X, m, R, \nu) + \ln p(y_n \mid x_n, z_n, \Phi, D) \right] + \text{const}
\]

where

\[
\mathbb{E}_{F_{\theta} \mid z_n} \left[ \ln \left( \prod_{m=1}^{M} \left( \frac{p(x_n \mid m_m, R_m) \pi_m(\nu)}{C_n(m, R, \nu)} \right)^{1[z_n=m]} \right) \right] = \sum_{m=1}^{M} \left\{ 1[z_n = m] \left( \mathbb{E}_{\nu_m} [\ln \nu_m] + \sum_{j=1}^{m-1} \mathbb{E}_{\nu_j} [\ln (1 - \nu_j)] + \frac{1}{2} \mathbb{E}_{R_m} [R_m] \right. \right.
\]

\[
- \frac{d}{2} \ln(2\pi) - \frac{1}{2} \mathbb{E}_{m_n, R_n} \left[ (x_n - m_m)^T R_m (x_n - m_m) \right] - \ln \left( \mathbb{E}_{n} \right)
\]

and

\[
\mathbb{E}_{F_{\theta} \mid z_n} \left[ \ln p(y_n \mid x_n, z_n, \Theta, D) \right]
\]

\[
= \sum_{m=1}^{M} \left\{ 1[z_n = m] \left( \ln c_m - \frac{1}{2} \mathbb{E}_{\theta_m} [\ln |\Lambda_m|] \right. \right.
\]

\[
- \frac{n_m - p + q}{2} \mathbb{E}_{\theta_m} \left[ \ln \|q + \frac{1}{n_m - p} (y_n - \mu^{**}(x_n))^T \Lambda_m^{-1} (y_n - \mu^{**}(x_n)) \| \right] \right\}
\]

\[
q(z_n) = \prod_{m=1}^{M} \left( \frac{\rho_{n,m}}{\sum_{m=1}^{M} \rho_{n,m}} \right)^{1[z_n=m]}
\]
Upon convergence of the VI algorithm and computation of the responsibilities \( \widehat{\rho}_{n,m} \), the classification of the observation data \( D \) is based on the maximum responsibilities.

\[
\text{assign } (x_n, y_n) \text{ to the } i-th \text{ model if } \arg \max_m \widehat{\rho}_{n,m} = i,
\]

each mixture component is taken to completely model a subset of the observation data.

**Note:** The number of data points assigned to each model has to be larger than \( p + q \) in order to ensure a proper posterior distribution for the covariance matrix \( \Sigma_m \).

Therefore, any model that is assigned less than \( p + q \) points is removed and the corresponding data are re-assigned to the models with the next greatest responsibility in explaining them.
The update of \( q(\theta_m) \) is slightly more complex.

**Issue:** The posterior of \( \theta_m \) does not belong to the variational family, i.e., \( q(\theta_m) \) cannot be constrained to a family of simple distributions.

**Solution:** Nonparametric variational inference [Gershman et al., 2012]

The lower bound of the local evidence \( \ln p(D_m) \) can be re-expressed as

\[
\mathcal{L}[q] = \mathbb{E}_q \left[ \ln \frac{p(\theta_m, D_m)}{q(\theta_m)} \right] = \mathcal{H}[q] + \mathbb{E}_q [g(\theta_m)]
\]

where

\[
g(\theta_m) = \mathbb{E}_{\psi \setminus \theta_m} \left[ \ln \pi(\theta_m|\gamma) + \ln p(S_m(Y; z)|S_m(X; z), \phi_m) \right] + \text{const}
\]

\[
= \ln \pi(\theta_m|\gamma) + \mathbb{E}_{B_m, \Sigma_m} \left[ \ln p(S_m(Y; z)|S_m(X; z), \theta_m, B_m, \Sigma_m) \right] + \text{const},
\]
Nonparametric Variational Inference

The lower bound of the local evidence \( \ln p(D_m) \) can be re-expressed as

\[
\mathcal{L}[q] = \mathbb{E}_q \left[ \ln \frac{p(\theta_m, D_m)}{q(\theta_m)} \right] = \mathcal{H}[q] + \mathbb{E}_q [g(\theta_m)]
\]

The proposal distribution \( q(\theta^{(m)}) \) is chosen to be,

\[
q(\theta_m) = \frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \omega_l^2 \mathcal{N}(\theta_m; \mathbf{m}_l, \sigma_l^2 \mathbb{I}_{d+1}).
\]

\[
\mathcal{H}[q] = -\int_{\theta_m} q(\theta_m) \ln q(\theta_m) d\theta_m
\]

\[
= -\int_{\theta_m} \frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \omega_l^2 \mathcal{N}(\theta_m; \mathbf{m}_l, \sigma_l^2 \mathbb{I}_{d+1}) \ln q(\theta_m) d\theta_m
\]

\[
= -\frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \omega_l^2 \int_{\theta_m} \mathcal{N}(\theta_m; \mathbf{m}_l, \sigma_l^2 \mathbb{I}_{d+1}) \ln q(\theta_m) d\theta_m
\]

\[
\geq -\frac{1}{(\sum_{l=1}^{L} \omega_l^2)^2} \sum_{l=1}^{L} \omega_l^2 \ln q_l, \quad q_l = \sum_{j=1}^{L} \omega_j^2 q_{lj}, \quad q_{lj} = \mathcal{N}(\mathbf{m}_l; \mathbf{m}_j, (\sigma_l^2 + \sigma_j^2) \mathbb{I}).
\]

and \( \mathbb{E}_q [g(\theta_m)] = \frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \omega_l^2 \int_{\theta_m} \mathcal{N}(\theta_m; \mathbf{m}_l, \sigma_l^2 \mathbb{I}_{d+1}) g(\theta_m) d\theta_m \) 

\[\text{[Gershman et al., 2012]}\]
Nonparametric Variational Inference

$g(\theta_m)$ around $m_l$ can be approximated by a second-order Taylor expansion

$$g(\theta_m) \approx \hat{g}_l(\theta_m) = g(m_l) + \nabla g(m_l)(\theta_m - m_l) + \frac{1}{2}(\theta_m - m_l)^T \mathcal{H}_l(\theta_m - m_l)$$

then

$$\mathbb{E}_q[g(\theta_m)] \approx \frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \int_{\theta_m} \omega_l^2 N(\theta_m; m_l, \sigma_l^2 \mathbb{I}_{d+1}) \hat{g}_l(\theta_m) d\theta_m.$$

$$= \frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \omega_l^2 \left\{ \frac{g(m_l) + \sigma_l^2}{2} tr(\mathcal{H}_l) \right\}.$$

The variational parameters $\omega_l, m_l$ and $\sigma_l$ are going to be learnt by a gradient ascent method.

- update $m_l$, use

$$\mathcal{L}_1[q(\theta_m)] = \frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \omega_l^2 g(m_l) - \frac{1}{(\sum_{l=1}^{L} \omega_l^2)^2} \sum_{l=1}^{L} \omega_l^2 \ln q_l.$$

- update $\omega_l$ and $\sigma_l$, use

$$\mathcal{L}_2[q(\theta)] = \frac{1}{\sum_{l=1}^{L} \omega_l^2} \sum_{l=1}^{L} \omega_l^2 \left\{ g(m_l) + \frac{\sigma_l^2}{2} tr(\mathcal{H}_l) \right\} - \frac{1}{(\sum_{l=1}^{L} \omega_l^2)^2} \sum_{l=1}^{L} \omega_l^2 \ln q_l.$$
Summary: Variational Inference Algorithm

- **Update $\nu_m$**
  
  $$
  \nu_m \sim \text{Beta} \left( 1 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n = m)], \alpha_0 + \sum_{n=1}^{N} \mathbb{E}_{z_n}[q(z_n > m)] \right)
  $$

- **Update $m_m$**
  
  $$
  m_m \sim \mathcal{N}_d(u_m, (R_0 + R_{m1})^{-1})
  $$

- **Update $R_m$**
  
  $$
  R_m \sim \mathcal{W}_d(W_m, \nu_m)
  $$

- **Update $q(z_n)$**
  
  $$
  q(z_n) = \prod_{m=1}^{M} \left( \frac{\rho_{n,m}}{\sum_{m=1}^{M} \rho_{n,m}} \right)^{[z_n = m]}
  $$

- **Cluster the data**

- **Update $q(B_m, \Sigma_m)$**
  
  $$
  q(B_m|\Sigma_m) = \mathcal{N}_{p \times q} \left( B_m | \widehat{B}_m, \mathbb{E}_{\theta_m} \left[ \left( H_m^T A_m^{-1} H_m \right)^{-1} \right], \Sigma_m \right)
  $$

- **Update $q(\theta_m)$**
  
  $$
  q(\Sigma_m) = \mathcal{W}^{-1}_{q} \left( \Sigma_m | (n_m - p) \widehat{\Sigma}_m, n_m - p \right)
  $$

- **Nonparametric variational inference**
Application to Uncertainty Quantification

- From the previous steps, we obtained the approximated posterior \( q(\Theta) \). Given the input distribution \( p(\xi) \), the objective is to find the statistics of responses.

- The training data is observed on \( S \in \mathbb{R}^{n_s \times d_s} \) and \( t \in \mathbb{R}^{n_t} \).

- The prediction is made on a denser spatial and temporal design, \( S^* \in \mathbb{R}^{n_s' \times d_s} \) and \( t^* \in \mathbb{R}^{n_t'} \).

Let \( \xi^* \sim p(\xi) \), \( s^* \in S^* \), and \( t^* \in t^* \) denote \( y^* \) be the output at \( x^* = (\xi^*, s^*, t^*) \).

The predictive response surface at \( \Theta^* \) is written as:

\[
p(y^*, z^* | x^*, \Psi, \mathcal{D}) = p(y^* | x^*, z^*, \Phi, \mathcal{D}) p(z^* | x^*, m, R, v)
\]

\[
= p(y^* | x^*, S_{z^*}(X; z), S_{z^*}(Y; z), \phi_{z^*}) p(z^* | x^*, m, R, v).
\]

where

\[
p(y^* | x^*, S_{z^*}(X; z), S_{z^*}(Y; z), \theta_{z^*}) = p(y^* | x^*, S_{z^*}(X; z), S_{z^*}(Y; z), \theta_{z^*}, \tilde{B}_{z^*}, \Sigma_{z^*}).
\]

\[
p(z^* | x^*, m, R, v) = \frac{p(x^* | z^*, m_{z^*}, R_{z^*}) p(z^* | v)}{\sum_{z^*} p(x^* | z^*, m_{z^*}, R_{z^*}) p(z^* | v)},
\]

[ Bilinois & Zabaras, 2013 ]
The 1\textsuperscript{st} and 2\textsuperscript{nd} order statistics of interest can then be evaluated analytically

- first-order statistics

\[
M^*(s^*, t^*) := \sum_{m=1}^{M} \int M^{(m)}(x^*) p(z^* = m|\mathbf{x}^*) p(\xi^*) d\xi^*
\]

\[
M^*(s^*, t^*) = \sum_{m=1}^{M} \left[ \zeta^{(m)T}_h \mathbf{B}_m + \zeta^{(m)T}_a A^{-1}_m (Y_m - \mathbf{H}_m \mathbf{B}_m) \right]
\]

where

\[
\zeta^{(m)}_h = \int \mathbf{h}(\mathbf{x}^*) p(z^* = m|\mathbf{x}^*) p(\xi^*) d\xi^*,
\]

\[
\zeta^{(m)}_a = \int \mathbf{a}_m(\mathbf{x}^*) p(z^* = m|\mathbf{x}^*) p(\xi^*) d\xi^*
\]
Application to uncertainty quantification

- Second-order statistics

\[ C_{ij}(s^*, t^*) : = \int (Y_i^* - M_i^*)(Y_j^* - M_j^*)^T p(\xi^*) d\xi^*, \]

\[ = \int \left( \sum_{m=1}^{M} M_i^{(m)}(x^*) p(z^* = m|x^*) - M_i^* \right) \left( \sum_{r=1}^{M} M_j^{(r)}(x^*) p(z^* = r|x^*) - M_j^* \right)^T p(\xi^*) d\xi^*. \]

\[ = \int \left( \sum_{m=1}^{M} \sum_{r=1}^{M} \beta_m \beta_r M_i^{(m)}(x^*) M_j^{(m)}(x^*)^T p(\xi^*) d\xi^* - M_i^*(M_j^*)^T, \right. \]

\[ C_{ij}^*(s^*, t^*) = \sum_{m=1}^{M} \sum_{r=1}^{M} \left[ \kappa_{hh}^{mr} + \kappa_{ha}^{mr} + \kappa_{ah}^{mr} + \kappa_{aa}^{mr} \right] - M_i^*(M_j^*)^T, \]

where

\[ \kappa_{hh}^{mr} : = \int \beta_m \beta_r \left( h(x^*) \tilde{B}_m \right)_i \left( \tilde{B}_r^T h(x^*)^T \right)_j p(\xi^*) d\xi^*, \]

\[ \kappa_{ha}^{mr} : = \int \beta_m \beta_r \left( h(x^*) \tilde{B}_m \right)_i \left( \tilde{Y}_r A_r^{-1} a_r(x^*) \right)_j p(\xi^*) d\xi^*, \]

\[ \kappa_{aa}^{mr} : = \int \beta_m \beta_r \left( a_m(x^*)^T A_m^{-1} \tilde{Y}_m \right)_i \left( \tilde{Y}_r A_r^{-1} a_r(x^*) \right)_j p(\xi^*) d\xi^*. \]
**Numerical Examples: KO-1d Problem**

- Kraichnan-Orszag (KO) problem

**Governing equation:**
\[
\begin{align*}
\frac{dy_1}{dt} &= y_1 y_3, \\
\frac{dy_2}{dt} &= -y_2 y_3, \\
\frac{dy_3}{dt} &= -y_1^2 + y_2^2,
\end{align*}
\]

**1D initial condition:**
\[
\begin{align*}
y_1(0) &= 1, \\
y_2(0) &= 0.1 \xi, \\
y_3(0) &= 0, \\
\xi &\sim U([-1, 1]).
\end{align*}
\]

Predicted variance \((n_\xi = 91)\) v.s MC variance \((n=1e6)\)

- **Data-selection process**

![Graphs showing variance and clusters](image)

- **Truncation level**
  \[M = 200\]
**Numerical Examples: KO-2d problem**

- **KO-2d problem**

  Initial conditions \( y_1(0) = 0, \ y_2(0) = 0.1\xi_1, \ y_3(0) = \xi_2, \ \xi_i \sim \mathcal{U}([-1, 1]), \ i = 1, 2. \)

  Predicted variance (\( n_\xi = 200, 400 \)) v.s MC variance (\( n=1e6 \))

  **Truncation level** \( M = 800 \)

  - \( n_\xi = 200 \)
  - \( n_\xi = 400 \)
KO-2d pdf Prediction

\[ y_2(t) \]

- \( n_\xi = 100 \)
- \( n_\xi = 200 \)
- \( n_\xi = 400 \)
KO-3d Variance Prediction

- **KO-3d problem**

Initial conditions \( y_1(0) = \xi_1, \ y_2(0) = \xi_2, \ y_3(0) = \xi_3, \ \xi_i \sim \mathcal{U}([-1, 1]), \ i = 1, 2, 3. \)

Predicted variance (\( n_\xi = 1000, 4000 \)) v.s MC variance (\( n=1e6 \))

**Truncation level** \( M = 2000 \)

- \( n_\xi = 1000 \)

- \( n_\xi = 4000 \)
KO-3d pdf Prediction

Predicted pdf ($n_\xi = 4000$) v.s MC estimates ($n=1e5$)
The porosity is correlated with the permeability

In this work, we consider immiscible, incompressible two-phase (water-oil) flow problem

Quarter-five spot

no capillary pressure and no gravity

[SPE-10 layer 1] 60x220

Log permeability

corner prod. well

injection well

[V. Ginting et al., 2013]
The governing equations are given as

**Pressure equation:**

\[ \nabla \cdot \mathbf{v} = q \]

\[ \mathbf{v} = - (\lambda_o + \lambda_w) K \nabla p \]

**Saturation equation:**

\[ \phi \frac{\partial s_\alpha}{\partial t} + \nabla \cdot (f_\alpha \mathbf{v}) = q_\alpha \]

where

- \( k_{rw} = (s')^2 \)
- \( k_{ro} = (1 - s')^2 \)
- \( s' = \frac{s - s_{wc}}{1 - s_{wc} - s_{or}} \)

The fractional-flow function (water-cut curve) at time \( t \) is defined as

\[ F(t) = \frac{\int_{\partial \Omega_{out}} f_w (\mathbf{v} \cdot \mathbf{n}) \, dx}{\int_{\partial \Omega_{out}} (\mathbf{v} \cdot \mathbf{n}) \, dx} \]

In this work, we take

- \( s_{wc} = s_{or} = 0.2 \)
- \( s_0 = s_{wc} \)
Discretize the Governing Equation

Pressure equation:

The mixed-finite element discretization

\[ \int_{\Lambda} \mathbf{v} \cdot (\lambda \mathbf{K})^{-1} \mathbf{u} d\mathbf{x} - \int_{\Lambda} P \nabla \cdot \mathbf{u} d\mathbf{x} = 0, \quad \text{for all } \mathbf{u} \in U, \]

\[ \int_{\Lambda} \nabla \cdot \mathbf{v} d\mathbf{x} = \int_{\Lambda} l q d\mathbf{x}, \quad \text{for all } l \in V. \]

Basis function

\[ p = \sum p_m \chi_m \text{ and } \mathbf{v} = \sum \mathbf{v}_{ij} \psi_{ij}. \]

Saturation equation:

finite-volume discretization

\[ s_{i}^{n+1} = s_{i}^{n} + \frac{\Delta t}{\phi_i \Lambda_i} \left( Q_i(s_{i}^{n+1}) - \sum_{j \neq i} F_{ij}(s_{j}^{n+1}) v_{ij} \right) \]

Newton-Raphson iterative method is employed to solve the implicit system

\[ Q_i(s_{i}^{n+1}) = \int_{\Lambda_i} q_w(s_{w}^{n+1}) d\mathbf{x} \]

\[ F_{ij}(s_{j}^{n+1}) = \max \left\{ \text{sign}(v_{ij}) f_w(s_{i}^{n+1}), -\text{sign}(v_{ij}) f_w(s_{j}^{n+1}) \right\} \]
Stochastic Input Model

From SPE-10 data set $1200 \times 2200 \times 170$ (ft$^3$) measurements in 60x220x85 grid, decompose into 340 2D reservoir domain in 60x54 grid, and assume they follow a second-order random field $G(x, \omega)$.

- Karhunen-Loève Expansion (KLE) [Loève, 1977]

$$G(x, \omega) = \mathbb{E}[G(x)] + \sum_{i=1}^{N} \sqrt{\lambda_i} \phi_i(x) \xi_i$$

where

- $\{\xi_i\}_{i=1}^{N}$ uncorrelated random variables
- $\phi_i(x)$ Eigen-functions
- $\lambda_i$ Eigen-values

The input stochastic dimension is set as 100

81% energy preserved at 50
91% energy preserved at 100

Normalized energy plot

81% energy preserved at 50
91% energy preserved at 100

Normalized energy plot
Choice of Simulation Parameters

The response is observed on 30 × 27 square spatial grid, and recorded every 30 days.

The predictions are made on a 60 × 54 grid and every 20 days.

The model is trained with $n_\xi = 40, 80$ and 160 observations.

The truncation level $M$ is set to be 20,000.

Remark: In the SPE10 projection, the water injection rate was set as 5000 fluid barrel per day and the reservoir domain was fully saturated around 2000 days.

\[ \begin{align*}
\mu_w &= 3.0 \text{ cP} \\
\mu_o &= 3.0 \text{ cP} \\
s_{wc} &= s_{or} = 0.2 \\
s_0 &= s_{wc} \\
T &= [0, 2100] \text{ days} \\
\xi &\sim N(0, 1)^d_\xi \\
d_\xi &= 100
\end{align*} \]
Predictions of the Mean of Saturation

Mean of Saturation at $T = 1000$ days

(a) $N_\xi = 40$

(b) $N_\xi = 80$

(c) $N_\xi = 160$

(d) Error bars for $N_\xi = 160$

(e) MC estimates with $N = 1e5$
Predictions of the std of Saturation

std of Saturation at T = 1000 days

(a) $N_{\xi} = 40$

(b) $N_{\xi} = 80$

(c) $N_{\xi} = 160$

(d) Error bars for $N_{\xi} = 160$

(e) MC estimates with $N = 1e5$
Predictions of the Mean of Saturation

Mean of Saturation at $T = 2000$ days

(a) $N_\xi = 40$

(b) $N_\xi = 80$

(c) $N_\xi = 160$

(d) Error bars for $N_\xi = 160$

(e) MC estimates with $N = 1e5$
Predictions of the std of Saturation

std of Saturation at $T = 2000$ days

(a) $N_\xi = 40$

(b) $N_\xi = 80$

(c) $N_\xi = 160$

(d) Error bars for $N_\xi = 160$

(e) MC estimates with $N = 1e5$
Predictions of the Mean of log x-velocity

Mean of log x-velocity at T = 1000 days

(a) \(N_\xi = 40\)

(b) \(N_\xi = 80\)

(c) \(N_\xi = 160\)

(d) Error bars for \(N_\xi = 160\)

(e) MC estimates with \(N = 1e5\)
Predictions of the std of log x-velocity

std of log x-velocity at $T = 1000$ days

(a) $N_\xi = 40$

(b) $N_\xi = 80$

(c) $N_\xi = 160$

(d) Error bars for $N_\xi = 160$

(e) MC estimates with $N = 1e5$
Predictions of the Mean of log y-velocity

Mean of log y-velocity at T = 2000 days

(a) $N_{\xi} = 40$

(b) $N_{\xi} = 80$

(c) $N_{\xi} = 160$

(d) Error bars for $N_{\xi} = 160$

(e) MC estimates with $N = 1e5$
Predictions of the std of log y-velocity

std of log y-velocity at $T = 2000$ days

(a) $N_\xi = 40$
(b) $N_\xi = 80$
(c) $N_\xi = 160$
(d) Error bars for $N_\xi = 160$
(e) MC estimates with $N = 1e5$
The PDFs of the Saturation at $T=1000\text{days}$

The pdfs of saturation at various spatial locations at $T = 1000$ days

(a) At location $(10, 10)$

(b) At location $(30, 22)$
The PDFs of the Saturation at $T=2000$ days

The pdfs of saturation at various spatial locations at $T = 2000$ days

(a) At location (10, 10)

(b) At location (30, 22)

(c) At location (5, 50)
The Water-Cut Curve

The water-cut curve from day 1 to day 2000

Mean of the water-cut curve

std of the water-cut curve
Closing Remarks

- A new outlook to high dimensional UQ was presented – rather than partitioning the stochastic space, we clustered the data. “Letting the data speak for themselves” is the key ingredient of the infinite GP mixture.

- Coupling an infinite mixture of GPs with variation inference was shown to be effective in high dimensional UQ tasks.

- Many of the difficulties associated with the use of a single GP are addressed.

- The input-dependent selection of clusters remains an open problem.

- Exploring (learning) the correlated nature of output responses is key to reduce the size of needed data.

- Using separable covariance function will further simplify the calculations and efficiency of the algorithm at the cost of modelling only linear correlations.