Bayesian Inference with Optimal Maps

Tarek El Moselhy and Youssef Marzouk, Massachusetts Institute of Technology

Motivation

- Conditioning models on data, via statistical inference, is central to many engineering and science applications
- **Bayesian approach:** foundation for inverse problems, data assimilation, assessing uncertainty in computational predictions, decision-making under uncertainty
- **Goal:** develop computationally efficient approach to Bayesian inference in large-scale complex systems
  - Would like algorithms that are theoretically sound, computationally efficient, and endowed with clear measures of accuracy and convergence
  - Overcome bottlenecks associated with traditional Bayesian computation (e.g., MCMC)

Bayesian framework

- Model parameters represented by random variable *x*; data *d*
- Bayes' rule

likelihood function

prior density

$$\pi(x \mid d) = \frac{L(d \mid x)\, p(x)}{\beta}$$

posterior density

evidence or marginal likelihood

Challenges

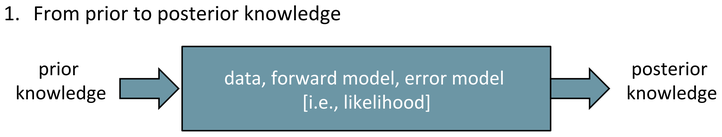
- Extracting information from the posterior: moments, marginals, realizations, set probabilities
- Evidence/normalization is difficult to compute
- Posterior evaluations may be expensive: forward model (PDE) appears inside the likelihood
- Parameter *x* may be high-dimensional

State of the art

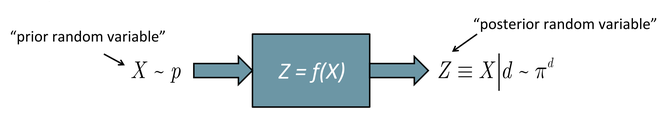
- Markov chain Monte Carlo (MCMC) simulation of the posterior: flexible and broadly applicable, but several drawbacks
- Variational Bayes: suitable for a special set of inference problems

Main idea: inference with maps

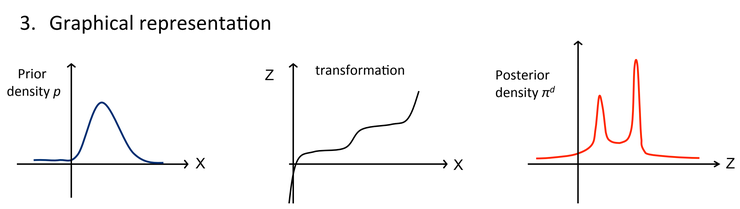
1. From prior to posterior knowledge



2. As a map or random variable transformation



3. Graphical representation



Map *pushes forward* the prior measure to the posterior measure

Formulation

$$\pi(z) = \frac{L(z;d)p(z)}{\beta} \quad \xrightarrow{Z = f(X)} \quad q(x;f) = \frac{L(f(x);d)\,p(f(x))}{\beta}\left|\det\frac{\partial f}{\partial x}\right|$$

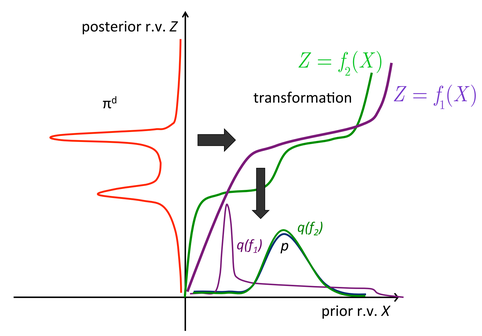
a probability density for the prior random variable *X*, parameterized by *f*

Find a monotone transformation $f(x)$ such that $q(x;f) = p(x)$

- Putting $q \to p$ is equivalent to requiring $T \to \text{constant}$

$$T(x;f) \equiv \log L(f(x);d) + \log p(f(x)) + \log\left|\det\frac{\partial f}{\partial x}\right| - \log p(x) = \log\beta$$

- As a byproduct of inference, we will **calculate the evidence β**



Optimization problems

- Map exists (under weak conditions) but is not unique
  - Example: linear Gaussian case, identity prior covariance, posterior $Z \sim N(\mu, \Sigma)$
  - Any map of the form $z = f(x) = \mu + Lx$, such that $LL^T = \Sigma$, is valid
- Two optimization formulations guarantee existence and uniqueness of a monotone map
  - Penalty from optimal transport theory [Caffarelli, McCann]:

$$\min_f \mathrm{Var}\left[T(X;f)\right] + \lambda\mathbb{E}\left[\left(f(X)-X\right)^T\left(f(X)-X\right)\right]$$

  - Knothe-Rosenblatt transport, i.e., "triangular" construction:

$$\min_f \mathrm{Var}\left[T(X;f)\right] \text{ such that } z_i = f_i(x_1, x_2, \ldots, x_i)$$

- These are stochastic optimization problems; use sample-average approximation with prior samples of *X*
- Represent *f* using an orthogonal polynomial expansion (e.g., Hermite chaos)

$$f(x) = G^T \Psi(x)$$

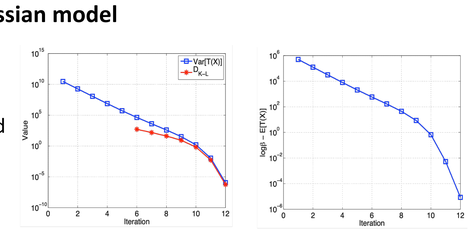
Vector of orthogonal polynomials

Matrix of unknown coefficients

Numerical demonstrations

Example 1: linear-Gaussian model

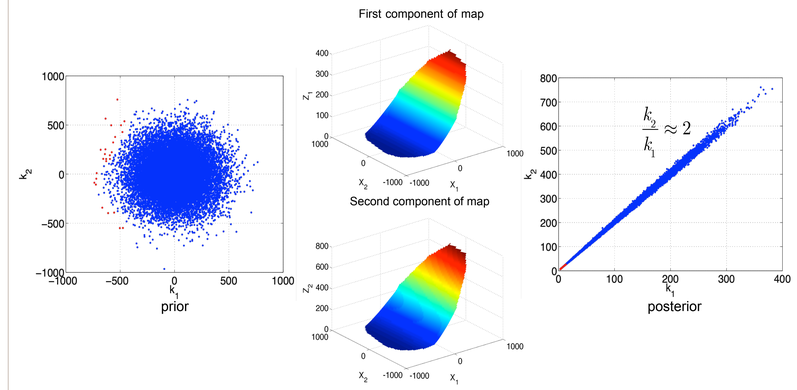
- $d = Ax + \epsilon$
  $A \in \mathbb{R}^{8\times100},\ d \in \mathbb{R}^8$
- *A* is randomly generated
- Convergence in < 12 iterations



Numerical demonstrations (cont.)

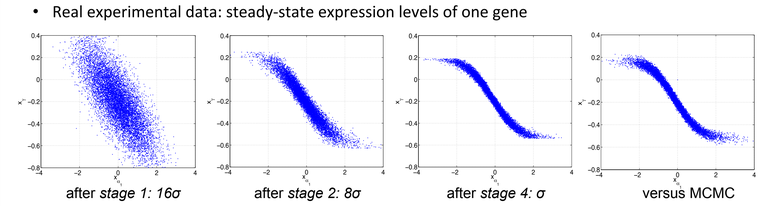
Example 2: Nonlinear reaction kinetics

- Five late-time observations of species A
- Infer parameters k_1 and k_2, truth is $k_1 = 1$, $k_2 = 2$ $\quad \frac{dA}{dt} = -k_1 A + k_2 B \quad \frac{dB}{dt} = k_1 A - k_2 B$
- 5th order polynomial map



Example 3: Cascaded maps

- DAE model of a genetic toggle switch from [Gardner 2000]; infer six parameters
- Real experimental data: steady-state expression levels of one gene

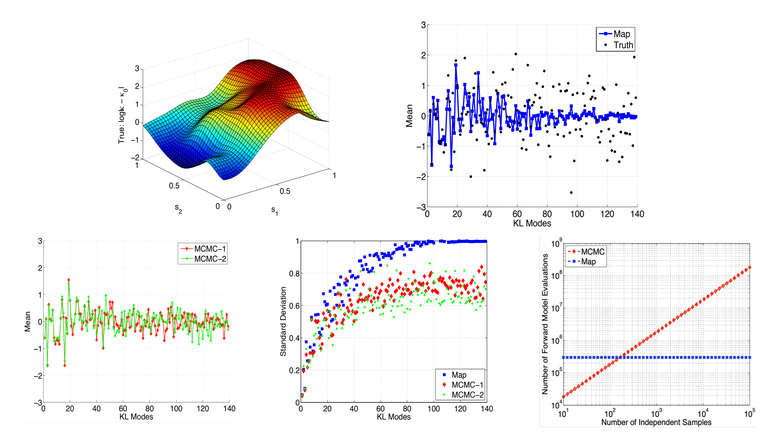


Example 4: PDE-constrained inverse problem (139 dimensions)

- Elliptic PDE: infer heterogeneous diffusivity *κ* from limited/noisy observations of pressure *p*
- 2-D spatial domain; *κ(r)* parameterized by Karhunen-Loève expansion with 139 independent modes
- Optimization algorithm identifies polynomial map up to 5th order
- Substantial gains in both *accuracy* and *speed* over MCMC

$$\nabla\cdot\left(\kappa(r,x)\nabla p(r)\right) = -g(r)$$

$$\kappa = \kappa_0 + \exp\left(\sum_{i=1}^{n}\phi_i(r)\sqrt{\lambda_i}\,x_i\right)$$



Conclusions

- An optimal transport interpretation of Bayesian statistical inference
- Posterior expressed as pushforward measure of prior
- Map computed by solving an optimization problem
- **Advantages:** clear convergence criterion; arbitrary numbers of *independent* posterior samples; embarrassingly parallel; exploit gradient/Hessian information; marginal likelihood computed for free (model selection); posterior moments computed analytically; enables sequential uncertainty propagation and inference

Support from DOE Office of Science/ASCR Early Career Research Program

ymarz@mit.edu