Variational Bayesian formulations with sparsity-enforcing priors for model calibration

I. Franck, P.S. Koutsourelakis
Continuum Mechanics Group
Technical University of Munich
p.s.koutsourelakis@tum.de

WCCM XI
Barcelona, July 23 2014
Can we use (continuum) models from solid mechanics to make/assist medical diagnosis?

**Model $\mathcal{M}$**

- Governing equation: $\nabla \cdot (FS) = 0$, $B$
- Boundary conditions: $u = u_0$, $\partial B$
- Constitutive law: $S = S(C; \Psi)$
  (In-compressibility: $J = 1$)

$\rightarrow$ noisy displacements (velocities etc) $\hat{u}$

$\Psi = ?$
Probabilistic approach

Bayes’ rule:

\[ p(\Psi | \hat{u}, M) = \frac{p(\hat{u} | \Psi, M) p(\Psi | M)}{p(\hat{u} | M)} \]

Goal: Find posterior density \( p(\Psi | \hat{u}, M) \)

- The posterior quantifies how likely a \( \Psi \) is to be the solution
- Provides a generalization over deterministic optimization strategies
- Evidence \( p(\hat{u} | M) \) quantifies how likely is for the data to have arisen from our model \( M \)
Probabilistic approach

Bayes’ rule:

\[ p(\Psi \mid \hat{u}, M) = \frac{p(\hat{u} \mid \Psi, M) p(\Psi \mid M)}{p(\hat{u} \mid M)} \]

Challenges:
- computational efficiency
- regularization (i.e. prior specification)
- dimensionality reduction
Variational inference attempts to *approximate* the posterior \( p(\Psi|\hat{u}, M) \) with a density \( q^*(\Psi) \) (belonging to an appropriate family of distributions \( Q \)) such that (Bishop 2006):

\[
q^*(\Psi) = \arg\min_{q \in Q} KL(q(\Psi)||p(\Psi|\hat{u}, M)) = -\int q(\Psi) \log \frac{p(\Psi|\hat{u}, M)}{q(\Psi)} \, d\Psi
\]
Minimizing the Kullback-Leibler divergence is equivalent to maximizing $\mathcal{F}(q, M)$:

$$
\log p(\hat{\mathbf{u}} | \mathcal{M}) = \log \int p(\hat{\mathbf{u}} | \Psi, \mathcal{M}) \, p(\Psi | \mathcal{M}) \, d\Psi \\
\geq \int q(\Psi) \frac{p(\hat{\mathbf{u}} | \Psi, \mathcal{M}) \, p(\Psi | \mathcal{M})}{q(\Psi)} \, d\Psi \quad (\text{Jensen's inequality}) \\
= \mathcal{F}(q, \mathcal{M})
$$

where:

$$
\mathcal{F}(q, \mathcal{M}) = \log p(\hat{\mathbf{u}} | \mathcal{M}) + KL(q(\Psi) || p(\Psi | \hat{\mathbf{u}}, \mathcal{M}))
$$
Variational Bayes

- If $\mathbb{E}$ implies expectation with $q(\Psi)$:

$$
\mathcal{F}(q, \mathcal{M}) = \int q(\Psi) \log \frac{p(\hat{u}|\Psi, \mathcal{M}) p(\Psi|\mathcal{M})}{q(\Psi)} \, d\Psi \\
= \left< \log p(\hat{u}|\Psi, \mathcal{M}) \right> + \left< \log p(\Psi|\mathcal{M}) \right> - \left< \log q \right>
$$

- Likelihood for data $\hat{u} \in \mathbb{R}^n$:

$$
\hat{u} = u(\Psi) + Z \Rightarrow p(\hat{u}|\Psi, \mathcal{M}) \propto \tau^{n/2} \exp\left\{ -\frac{\tau}{2} |\hat{u} - u(\Psi)|^2 \right\}
$$

where:

- $u(\Psi)$: model $\mathcal{M}$-predicted displacements for given material properties $\Psi$
- $Z$: observation noise, e.g. $Z \sim \mathcal{N}(0, \tau^{-1} I)$
If \( < . > \) implies expectation with \( q(\Psi) \):

\[
\mathcal{F}(q, \mathcal{M}) = \int q(\Psi) \log \frac{p(\hat{u}|\Psi, \mathcal{M}) p(\Psi|\mathcal{M})}{q(\Psi)} d\Psi \\
= < \log p(\hat{u}|\Psi, \mathcal{M}) > + < \log p(\Psi|\mathcal{M}) > - < \log q >
\]

Likelihood for data \( \hat{u} \in \mathbb{R}^n \):

\[
\hat{u} = u(\Psi) + Z \rightarrow p(\hat{u}|\Psi, \mathcal{M}) \propto \tau^{n/2} \exp\{ -\frac{\tau}{2} |\hat{u} - u(\Psi)|^2 \}
\]

where:
- \( u(\Psi) \): model \( \mathcal{M} \)-predicted displacements for given material properties \( \Psi \)
- \( Z \): observation noise, e.g. \( Z \sim \mathcal{N}(0, \tau^{-1} I) \)
Variational Bayes

- **Assumption 1:** One possible solution is to linearize $u(\Psi)$ using $G = \frac{\partial u}{\partial \Psi}$ using adjoint PDE (Chappelle et al 2009):

  $$u(\Psi) \approx u(\Psi_0) + G(\Psi - \Psi_0)$$

- As a result:

  $$\log p(\hat{u} | \Psi, M) = -\frac{T}{2}|\hat{u} - u(\Psi)|^2$$

  $$= -\frac{T}{2}(|u(\Psi) - u(\Psi_0)|^2 - 2(u(\Psi) - u(\Psi_0))^T G(\Psi - \Psi_0) + (\Psi - \Psi_0)^T G^T G(\Psi - \Psi_0))$$

- **Assumption 2:** Family of approximating distributions $q \in \mathcal{Q}$ are *multivariate Gaussians* $\mathcal{N}(\mu, S)$. 
Variational Bayes

Algorithm

\[
\max_{\mu, S} F(q, \mathcal{M}) = \langle \log p(\hat{u} | \psi, \mathcal{M}) \rangle + \langle \log p(\psi | \mathcal{M}) \rangle - \langle \log q \rangle
\]

0. Suppose a prior \( p(\psi | \mathcal{M}) \equiv \mathcal{N}(\mu_0, S_0) \). Initialize \( q(\psi) \equiv \mathcal{N}(\mu, S) \).

1. Set \( \psi_0 = \mu \) and linearize \( u(\psi) \approx u(\psi_0) + G(\psi - \psi_0) \).

2. Update for \( q(\psi) \):

\[
S^{-1} = \tau G^T G + S^{-1} \quad \text{and} \quad S^{-1} \mu = \tau G^T (\hat{u} - u(\psi_0)) + S_0^{-1} \mu_0
\]

3. Goto 1. until convergence
Figure: MCMC: 20,000 forward runs vs Variational Bayes: 50 forward runs
What should the prior be for an **undetermined problem** i.e. when data \( \hat{u} \in \mathbb{R}^n \) and unknowns \( \Psi \in \mathbb{R}^N, ~ N >> n \):

1) Smoothness-enforcing prior:

\[
p(\Psi|\mathcal{M}) \equiv \mathcal{N}(\mu_0, S_0)
\]

where the covariance \( S_0 \) enforces some smoothness/correlation.

- How big/small should that correlation be?
- Should I be using a different norm?

2) Introduce hyper-parameter(s) that penalize the jumps between neighboring \( \Psi \), which leads to (Bardsley 2013):

\[
p(\Psi|\mathcal{M}) \propto \exp\{-\frac{\delta}{2} \Psi^T L \Psi\}, \quad L : \text{Laplacian of graph}
\]

- How big/small should the neighborhoods be?
- Must also infer the hyperparameters (same or different for each jump).
What should the prior be for an undetermined problem i.e. when data \( \hat{u} \in \mathbb{R}^n \) and unknowns \( \Psi \in \mathbb{R}^N, N >> n \):

1) Smoothness-enforcing prior:

\[
p(\Psi|M) \equiv \mathcal{N}(\mu_0, S_0)
\]

where the covariance \( S_0 \) enforces some smoothness/correlation.

- How big/small should that correlation be?
- Should I be using a different norm?

2) Introduce hyper-parameter(s) that penalize the jumps between neighboring \( \Psi_i \), which leads to (Bardsley 2013):

\[
p(\Psi|M) \propto \exp\left\{ -\frac{\delta}{2} \Psi^T L \Psi \right\}, \quad L : \text{Laplacian of graph}
\]

- How optimal should the neighborhoods be?
- Must also infer the hyperparameters (same or different for each jump).
What should the prior be for an **undetermined problem** i.e. when data \( \hat{u} \in \mathbb{R}^n \) and unknowns \( \Psi \in \mathbb{R}^N \), \( N >> n \):

1) Smoothness-enforcing prior:

\[
p(\Psi|\mathcal{M}) \equiv \mathcal{N}(\mu_0, S_0)
\]

where the covariance \( S_0 \) enforces some smoothness/correlation.

- How big/small should that correlation be?
- Should I be using a different norm?

2) Introduce hyper-parameter(s) that penalize the jumps between neighboring \( \Psi_i \) which leads to (Bardsley 2013):

\[
p(\Psi|\mathcal{M}) \propto \exp\{-\frac{\delta}{2} \Psi^T L \Psi\}, \quad L : \text{Laplacian of graph}
\]

- How big/small should the neighborhoods be?
- Must also infer the hyperparameters (same or different for each jump).
What should the prior be for an undetermined problem i.e. when data \( \hat{u} \in \mathbb{R}^n \) and unknowns \( \Psi \in \mathbb{R}^N \), \( N \gg n \):

1) Smoothness-enforcing prior:

\[
p(\Psi|M) \equiv \mathcal{N}(\mu_0, S_0)
\]

where the covariance \( S_0 \) enforces some smoothness/correlation.

- How big/small should that correlation be?
- Should I be using a different norm?

2) Introduce hyper-parameter(s) that penalize the jumps between neighboring \( \Psi_i \) which leads to (Bardsley 2013):

\[
p(\Psi|M) \propto \exp\{-\frac{\delta}{2} \Psi^T L \Psi\}, \quad L : \text{Laplacian of graph}
\]

- How big/small should the neighborhoods be?
- Must also infer the hyperparameters (same or different for each jump).
Can one infer \( \Psi \in \mathbb{R}^N \) on a (much lower) dimensional subspace?

\[
\Psi_{N \times 1} = \mu_{N \times 1} + W_{N \times k} \theta_{k \times 1}, \quad k << N
\]

The basis vectors \( W = [w_1, w_2, \ldots, w_k] \) should depend on the data and the model \( \mathcal{M} \).

Given data \( \hat{u} \) and a forward model \( \mathcal{M} \), the best \( (\mu, W) \) should maximize the evidence:

\[
p(\hat{u}|\mathcal{M}) = p(\hat{u}|\mu, W)
\]

The advantage of the Variational Bayesian formulation adopted is that we also obtain an estimate (lower bound) on the evidence:

\[
p(\hat{u}|\mathcal{M}) \approx \mathcal{F}(q(\theta), \mu, W)
\]
\[
= - \langle \log p(\hat{u}|\theta, \mu, W) \rangle + \langle \log p(\theta|\mathcal{M}) \rangle - \langle \log q(\theta) \rangle
\]
\[
= - \langle \frac{1}{2} |\hat{u} - u(\mu + W\theta)|^2 \rangle + \ldots...
\]

where the expectation \( \langle . \rangle \) is with respect to the approximate posterior \( q(\theta) \) of the reduced coordinates \( \theta \).
Can one infer $\Psi \in \mathbb{R}^N$ on a (much lower) dimensional subspace?

$$\Psi_{N \times 1} = \mu_{N \times 1} + W_{N \times k} \theta_{k \times 1}, \quad k << N$$

The basis vectors $W = [w_1, w_2, \ldots, w_k]$ should depend on the data and the model $\mathcal{M}$.

Given data $\hat{u}$ and a forward model $\mathcal{M}$, the best $(\mu, W)$ should maximize the evidence:

$$p(\hat{u}|\mathcal{M}) = p(\hat{u}|\mu, W)$$

The advantage of the Variational Bayesian formulation adopted is that we also obtain an estimate (lower bound) on the evidence:

$$p(\hat{u}|\mathcal{M}) \approx \mathcal{F}(q(\theta), \mu, W)$$

$$= <\log p(\hat{u}|\theta, \mu, W)> + <\log p(\theta|\mathcal{M})> - <\log q(\theta)>$$

$$= - <\frac{\tau}{2}|\hat{u} - \mu + W\theta|^2> + \ldots$$

where the expectation $<.>$ is with respect to the approximate posterior $q(\theta)$ of the reduced coordinates $\theta$. 
Regularization & Dimensionality reduction

\[ \mathbf{\Psi} = \mathbf{\mu} + \mathbf{W} \mathbf{\theta}, \quad k \ll N \]

How can one infer the effective dimensionality \( k \)?

- Hierarchical heavy-tailed prior:
  \[
  p(\mathbf{w}_j | a_j) \equiv \mathcal{N}(0, a_j^{-1} \mathbf{I}_{N \times N})
  
  p(a_j) \equiv \text{Gamma}(\alpha, \beta), \quad j = 1, \ldots, k
  
  \text{Automatic Relevance Determination priors (ARD, MacKay 1994))}: \quad a_j \to \infty \text{ then } \mathbf{w}_j \to \mathbf{0} \text{ (i.e. basis vector } j \text{ is inactive)}
  
  \text{Closely related to LASSO (Tibshirani 1996), Compressive Sensing (Candés et al 2006, Donoho et al 2006)}
Variational Expectation-Maximization

\[
\max \mathcal{F}(q(\theta, a, \tau), \mu, W) = < \frac{n}{2} \log \tau >_{q(\tau)} - < \frac{\tau}{2} | \hat{u} - u(\mu + W\theta) |^2 >_{q(\theta, \tau)} \text{ (likelihood)} \\
+ < \log p(\theta) >_{q(\theta)} + < \log p(W|a)p(a) >_{q(a)} \text{ (priors)} \\
- < \log q(\theta, a, \tau) >_{q(\theta, a, \tau)}
\]

- **Assumption 1**: Mean-field approximation \( q(\theta, a, \tau) \approx q(\theta) q(\tau) q(a) \) (Wainwright 2008)

- **Assumption 2**: Linearize \( u(\mu + W\theta) \approx u(\mu) + GW\theta \)

**Algorithm \( O(N) \):**

0. Initialize \( \mu, W \)

1. Repeat until convergence:
   - Fix \( \mu, W \) and update \( q(\theta), q(\tau), q(a) \)
   - Fix \( W, q(\theta), q(\tau), q(a) \) and update \( \mu \)
   - Fix \( \mu, q(\theta), q(\tau), q(a) \) and update \( W \)
Example:

- large deformation, incompressible non-linear elasticity
- Mooney-Rivlin constitutive law: \( \Phi = c_1 (I_1 - 3) + c_2^0 (I_2 - 3) + \frac{1}{2} \kappa (\log J)^2 \)
- Synthetic data from fine (200 \( \times \) 200) mesh, contaminated \( SNR = 5 \times 10^3 \)
- \( \text{dim}(\Psi) = N = 25000 \), reduced-dimension \( k = 16 \)

**Figure:** Ground truth: Log of material parameter \( c_1 \)
Numerical Illustration

Example:

(a) Posterior along diagonal

(b) Posterior along diagonal

(c) Posterior mean

(d) Posterior mean

Figure: (Left) Without (Right) With updating $W$
Numerical Illustration

Example:

![Figure: Evolution of variational objective $\mathcal{F}$](image)

- **Fixed W**
- **Updated W**

$\mathcal{F}$
Numerical Illustration

Example:

(a) iteration 1  
(b) iteration 21  
(c) iteration 41

Figure: Evolution of most important (i.e. largest $< \theta_j^2 >$) basis vector in $W$
Variational Bayesian methods offer comparable accuracy and much greater efficiency as compared to sampling (MCMC/SMC) methods.

By approximating the log-evidence one can obtain automatic regularization and enable significant dimensionality reduction.

- **Adaptivity:**
  - incorporate data sequentially
  - utilize a hierarchy of forward models
  - experimental design i.e. determine measurement locations or excitations that will maximize information intake

- **Accuracy:**
  - *Mixture models*: Consider a mixture of $M$ reduced-representations

\[
\Psi|m = \mu_m + W_m \theta_m,
\]

\[
\to p(\Psi|\hat{u}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\Psi; \mu_m + W_m \mu_{\theta_m}, W_m S_{\theta_m} W_m^T)
\]

- this can capture *non-Gaussian* projections
- lead to greater dimensionality reduction
Conclusion & Extensions

- Variational Bayesian methods offer comparable accuracy and much greater efficiency as compared to sampling (MCMC/SMC) methods.
- By approximating the log-evidence one can obtain automatic regularization and enable significant dimensionality reduction.

Adaptivity:
- incorporate data sequentially
- utilize a hierarchy of forward models
- experimental design i.e. determine measurement locations or excitations that will maximize information intake

Accuracy:
- **Mixture models**: Consider a mixture of $M$ reduced-representations

$$\Psi|m = \mu_m + W_m \theta_m,$$

$$\rightarrow p(\Psi|\hat{u}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\Psi; \mu_m + W_m \mu_{\theta_m}, W_m S_{\theta_m} W_m^T)$$

- this can capture non-Gaussian projections
- lead to greater dimensionality reduction
Conclusion & Extensions

- Variational Bayesian methods offer comparable accuracy and much greater efficiency as compared to sampling (MCMC/SMC) methods.
- By approximating the log-evidence one can obtain automatic regularization and enable significant dimensionality reduction.

Adaptivity:
- incorporate data sequentially
- utilize a hierarchy of forward models
- experimental design i.e. determine measurement locations or excitations that will maximize information intake

Accuracy:
- **Mixture models**: Consider a mixture of $M$ reduced-representations

This can capture *non-Gaussian* projections and lead to greater dimensionality reduction.

\[
p(\Psi | \hat{u}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\Psi; \mu_m + W_m \theta_m, W_m S_{\theta_m} W_m^T)
\]