Variational Bayesian formulations with sparsity-enforcing priors for model calibration

Fachgebiet für
Kontinuums
Mechanik

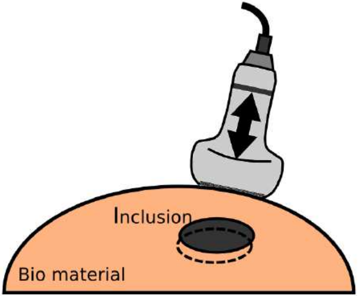
I. Franck, P.S. Koutsourelakis
Continuum Mechanics Group
Technical University of Munich
p.s.koutsourelakis@tum.de

WCCM XI
Barcelona, July 23 2014

Motivation

Can we use (continuum) models from solid mechanics to make/assist medical diagnosis?

$$\text{Model } \mathcal{M} \begin{cases} \text{Governing equation:} & \nabla \cdot (\boldsymbol{FS}) = 0, \quad \mathcal{B} \\ \text{Boundary conditions:} & \boldsymbol{u} = \boldsymbol{u}_0, \quad \partial\mathcal{B} \\ \text{Constitutive law:} & \boldsymbol{S} = \boldsymbol{S}(\boldsymbol{C}; \boldsymbol{\Psi}) \\ \text{(In-compressibility:} & J = 1) \end{cases}$$



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

$$\downarrow$$

$$\boldsymbol{\Psi} = ?$$

Probabilistic approach

Bayes' rule:

$$p(\underbrace{\boldsymbol{\Psi}}_{material\ par.} \mid \underbrace{\hat{\boldsymbol{u}}}_{data}, \underbrace{\mathcal{M}}_{model}) = \frac{\overbrace{p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi},\mathcal{M})}^{likelihood}\ \overbrace{p(\boldsymbol{\Psi}|\mathcal{M})}^{prior}}{\underbrace{p(\hat{\boldsymbol{u}}|\mathcal{M})}_{evidence}}$$

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

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

Probabilistic approach

Bayes' rule:

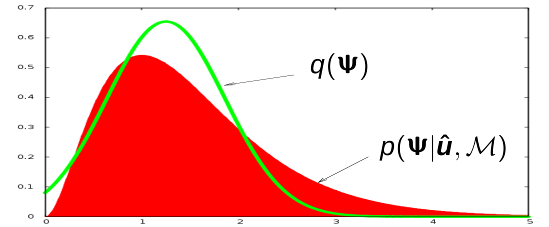
$$p(\underbrace{\boldsymbol{\Psi}}_{material\ par.} \mid \underbrace{\hat{\boldsymbol{u}}}_{data}, \underbrace{\mathcal{M}}_{model}) = \frac{\overbrace{p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi},\mathcal{M})}^{likelihood}\ \overbrace{p(\boldsymbol{\Psi}|\mathcal{M})}^{prior}}{\underbrace{p(\hat{\boldsymbol{u}}|\mathcal{M})}_{evidence}}$$

Challenges:

- computational efficiency
- regularization (i.e. prior specification)
- dimensionality reduction

Variational Bayes

Variational inference attempts to *approximate* the posterior $p(\boldsymbol{\Psi}|\hat{\boldsymbol{u}}, \mathcal{M})$ with a density $q^*(\boldsymbol{\Psi})$ (belonging to an appropriate family of distributions $\mathcal{Q}$) such that (Bishop 2006):



$$q^*(\boldsymbol{\Psi}) = \arg\min_{q \in \mathcal{Q}} KL(q(\boldsymbol{\Psi})||p(\boldsymbol{\Psi}|\hat{\boldsymbol{u}}, \mathcal{M})) = -\int q(\boldsymbol{\Psi}) \log \frac{p(\boldsymbol{\Psi}|\hat{\boldsymbol{u}}, \mathcal{M})}{q(\boldsymbol{\Psi})} \, d\boldsymbol{\Psi}$$

Variational Bayes

$$p(\underbrace{\boldsymbol{\Psi}}_{material\ par.} | \underbrace{\hat{\mathbf{u}}}_{data}, \underbrace{\mathcal{M}}_{model}) = \frac{\overbrace{p(\hat{\mathbf{u}}|\boldsymbol{\Psi},\mathcal{M})}^{likelihood}\ \overbrace{p(\boldsymbol{\Psi}|\mathcal{M})}^{prior}}{\underbrace{p(\hat{\mathbf{u}}|\mathcal{M})}_{evidence}}$$

- Minimizing the Kullback-Leibler divergence is equivalent to maximizing $\mathcal{F}(q,\mathcal{M})$:

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

where:

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

Variational Bayes

- If $< . >$ implies expectation with $q(\boldsymbol{\Psi})$:

$$\begin{aligned} \mathcal{F}(q, \mathcal{M}) &= \int q(\boldsymbol{\Psi}) \log \frac{p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi}, \mathcal{M})\ p(\boldsymbol{\Psi}|\mathcal{M})}{q(\boldsymbol{\Psi})}\, d\boldsymbol{\Psi} \\ &=< \log p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi}, \mathcal{M}) > + < \log p(\boldsymbol{\Psi}|\mathcal{M}) > - < \log q > \end{aligned}$$

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

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

where:

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

Variational Bayes

- If $< . >$ implies expectation with $q(\boldsymbol{\Psi})$:

$$\begin{aligned} \mathcal{F}(q, \mathcal{M}) &= \int q(\boldsymbol{\Psi}) \log \frac{p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi}, \mathcal{M})\ p(\boldsymbol{\Psi}|\mathcal{M})}{q(\boldsymbol{\Psi})}\ d\boldsymbol{\Psi} \\ &= \underbrace{< \log p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi}, \mathcal{M}) >}_{\text{difficult}} + \underbrace{< \log p(\boldsymbol{\Psi}|\mathcal{M}) > - < \log q >}_{\text{easy}} \end{aligned}$$

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

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

where:

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

Variational Bayes

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

$$\boldsymbol{u}(\boldsymbol{\Psi}) \approx \boldsymbol{u}(\boldsymbol{\Psi}_0) + \boldsymbol{G}(\boldsymbol{\Psi} - \boldsymbol{\Psi}_0)$$

- As a result:

$$\begin{aligned} \log p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi}, \mathcal{M}) \quad & = -\tfrac{\tau}{2}|\hat{\boldsymbol{u}} - \boldsymbol{u}(\boldsymbol{\Psi})|^2 \\ & = -\tfrac{\tau}{2}(|\boldsymbol{u}(\boldsymbol{\Psi}) - \boldsymbol{u}(\boldsymbol{\Psi}_0)|^2 - 2(\boldsymbol{u}(\boldsymbol{\Psi}) - \boldsymbol{u}(\boldsymbol{\Psi}_0))^T \boldsymbol{G}(\boldsymbol{\Psi} - \boldsymbol{\Psi}_0) \\ & +(\boldsymbol{\Psi} - \boldsymbol{\Psi}_0)^T \boldsymbol{G}^T \boldsymbol{G}(\boldsymbol{\Psi} - \boldsymbol{\Psi}_0)) \end{aligned}$$

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

Variational Bayes

Algorithm

$$\max_{\boldsymbol{\mu},\boldsymbol{S}} F(q,\mathcal{M}) =< \log p(\hat{\boldsymbol{u}}|\boldsymbol{\Psi},\mathcal{M}) > + < \log p(\boldsymbol{\Psi}|\mathcal{M}) > - < \log q >$$

0. Suppose a prior $p(\boldsymbol{\Psi}|\mathcal{M}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$. Initialize $q(\boldsymbol{\Psi}) \equiv \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{S})$
1. Set $\boldsymbol{\Psi}_0 = \boldsymbol{\mu}$ and linearize $u(\boldsymbol{\Psi}) \approx \boldsymbol{u}(\boldsymbol{\Psi}_0) + \boldsymbol{G}(\boldsymbol{\Psi} - \boldsymbol{\Psi}_0)$.
2. Update for $q(\boldsymbol{\Psi})$:

$$\boldsymbol{S}^{-1} = \tau \boldsymbol{G}^T \boldsymbol{G} + \boldsymbol{S}^{-1}$$
$$\boldsymbol{S}^{-1}\boldsymbol{\mu} = \tau \boldsymbol{G}^T(\hat{\boldsymbol{u}} - \boldsymbol{u}(\boldsymbol{\Psi}_0)) + \boldsymbol{S}_0^{-1}\boldsymbol{\mu}_0$$

3. Goto 1. until convergence

Variational Bayes

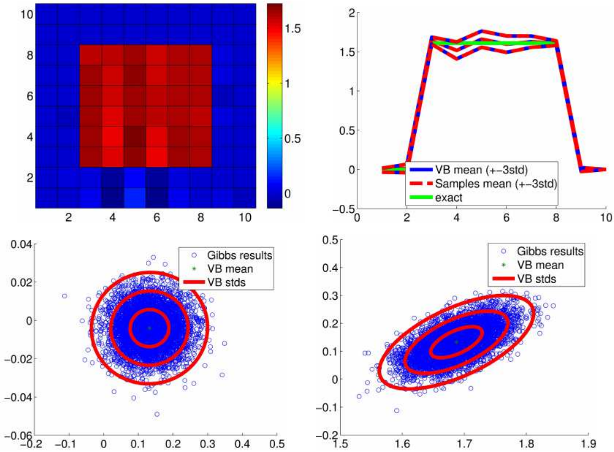


Figure: MCMC: 20,000 forward runs vs Variational Bayes: 50 forward runs

Regularization & Dimensionality reduction

- What should the prior be for an undetermined problem i.e. when data $\hat{\boldsymbol{u}} \in \mathbb{R}^n$ and unknowns $\boldsymbol{\Psi} \in \mathbb{R}^N, \ N >> n$:

  1) Smoothness-enforcing prior:

  $$p(\boldsymbol{\Psi}|\mathcal{M}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$$

  where the covariance $\boldsymbol{S}_0$ 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 Ψ_i which leads to (Bardsley 2013):

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

     - How big/small should the neighborhoods be?
     - Must also infer the hyperparameters (same or different for each jump)

Regularization & Dimensionality reduction

- What should the prior be for an undetermined problem i.e. when data $\hat{\boldsymbol{u}} \in \mathbb{R}^n$ and unknowns $\boldsymbol{\Psi} \in \mathbb{R}^N,\ N >> n$:

  1) Smoothness-enforcing prior:

  $$p(\boldsymbol{\Psi}|\mathcal{M}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$$

  where the covariance $\boldsymbol{S}_0$ 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 Ψ_i which leads to (Bardsley 2013):

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

     - How big/small should the neighborhoods be?
     - Must also infer the hyperparameters (same or different for each jump)

Regularization & Dimensionality reduction

- What should the prior be for an undetermined problem i.e. when data $\hat{\boldsymbol{u}} \in \mathbb{R}^n$ and unknowns $\boldsymbol{\Psi} \in \mathbb{R}^N, \ N >> n$:

  1) Smoothness-enforcing prior:

  $$p(\boldsymbol{\Psi}|\mathcal{M}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$$

  where the covariance $\boldsymbol{S}_0$ 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 Ψ_i which leads to (Bardsley 2013):

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

  - How big/small should the neighborhoods be?
  - Must also infer the hyperparameters (same or different for each jump).

Regularization & Dimensionality reduction

- What should the prior be for an undetermined problem i.e. when data $\hat{\boldsymbol{u}} \in \mathbb{R}^n$ and unknowns $\boldsymbol{\Psi} \in \mathbb{R}^N, \; N >> n$:

1) Smoothness-enforcing prior:

$$p(\boldsymbol{\Psi}|\mathcal{M}) \equiv \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{S}_0)$$

where the covariance $\boldsymbol{S}_0$ 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 Ψ_i which leads to (Bardsley 2013):

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

- How big/small should the neighborhoods be?
- Must also infer the hyperparameters (same or different for each jump).

Regularization & Dimensionality reduction

- Can one infer $\boldsymbol{\Psi} \in \mathbb{R}^N$ on a (much lower) dimensional subspace?

$$\underbrace{\boldsymbol{\Psi}}_{N\times 1} = \underbrace{\boldsymbol{\mu}}_{N\times 1} + \underbrace{\boldsymbol{W}}_{N\times k}\underbrace{\boldsymbol{\theta}}_{k\times 1}, \quad k << N$$

- The basis vectors $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k]$ should depend on the data and the model $\mathcal{M}$.
- Given data $\hat{\boldsymbol{u}}$ and a forward model $\mathcal{M}$, the best $(\boldsymbol{\mu}, \boldsymbol{W})$ should maximize the evidence:

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

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

$$\begin{aligned} p(\hat{\boldsymbol{u}}|\mathcal{M}) &\approx \mathcal{F}(q(\boldsymbol{\theta}),\ \boldsymbol{\mu}, \boldsymbol{W}) \\ &=< \log p(\hat{\boldsymbol{u}}|\boldsymbol{\theta},\ \boldsymbol{\mu}, \boldsymbol{W}) > + < \log p(\boldsymbol{\theta}|\mathcal{M}) > - < \log q(\boldsymbol{\theta}) > \\ &= - < \frac{\tau}{2}|\hat{\boldsymbol{u}} - \boldsymbol{u}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\theta})|^2 > + \ldots\ldots. \end{aligned}$$

where the expectation $< . >$ is with respect to the approximate posterior $q(\boldsymbol{\theta})$ of the reduced coordinates $\boldsymbol{\theta}$

Regularization & Dimensionality reduction

- Can one infer $\boldsymbol{\Psi} \in \mathbb{R}^N$ on a (much lower) dimensional subspace?

$$\underbrace{\boldsymbol{\Psi}}_{N\times 1} = \underbrace{\boldsymbol{\mu}}_{N\times 1} + \underbrace{\boldsymbol{W}}_{N\times k}\underbrace{\boldsymbol{\theta}}_{k\times 1}, \quad k << N$$

- The basis vectors $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_k]$ should depend on the data and the model $\mathcal{M}$.
- Given data $\hat{\boldsymbol{u}}$ and a forward model $\mathcal{M}$, the best $(\boldsymbol{\mu}, \boldsymbol{W})$ should maximize the evidence:

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

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

$$\begin{aligned} p(\hat{\boldsymbol{u}}|\mathcal{M}) \quad & \approx \mathcal{F}(q(\boldsymbol{\theta}),\ \boldsymbol{\mu}, \boldsymbol{W}) \\ & =< \log p(\hat{\boldsymbol{u}}|\boldsymbol{\theta},\ \boldsymbol{\mu}, \boldsymbol{W}) > + < \log p(\boldsymbol{\theta}|\mathcal{M}) > - < \log q(\boldsymbol{\theta}) > \\ & = - < \frac{\tau}{2}|\hat{\boldsymbol{u}} - \boldsymbol{u}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\theta})|^2 > + \ldots .... \end{aligned}$$

where the expectation $< . >$ is with respect to the approximate posterior $q(\boldsymbol{\theta})$ of the reduced coordinates $\boldsymbol{\theta}$

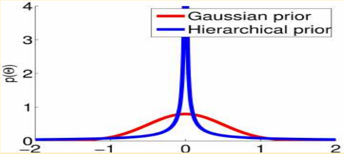
Regularization & Dimensionality reduction

$$\underbrace{\boldsymbol{\Psi}}_{N\times 1} = \underbrace{\boldsymbol{\mu}}_{N\times 1} + \underbrace{\boldsymbol{W}}_{N\times k}\underbrace{\boldsymbol{\theta}}_{k\times 1}, \quad k << N$$

How can one infer the effective dimensionality k?

- Hierarchical heavy-tailed prior:

$$p(\boldsymbol{w}_j|a_j) \equiv \mathcal{N}(0, a_j^{-1}\boldsymbol{I}_{N\times N})$$
$$p(a_j) \equiv Gamma(\alpha, \beta), \quad j = 1, \dots, k$$



- Automatic Relevance Determination priors (ARD, MacKay 1994)): $a_j \to \infty$ then $\boldsymbol{w}_j \to \boldsymbol{0}$ (i.e. basis vector j is inactive)
- Closely related to LASSO (Tibshirani 1996), Compressive Sensing (Candés et al 2006, Donoho et al 2006)

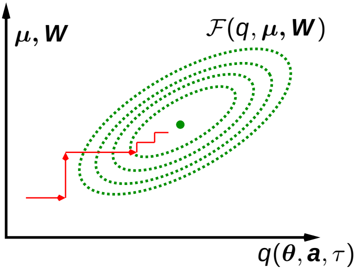
Variational Expectation-Maximization

$$\begin{aligned} \max \mathcal{F}(q(\boldsymbol{\theta}, \boldsymbol{a}, \tau), \boldsymbol{\mu}, \boldsymbol{W}) &= < \frac{n}{2} \log \tau >_{q(\tau)} - < \frac{\tau}{2} |\hat{\boldsymbol{u}} - \boldsymbol{u}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\theta})|^2 >_{q(\boldsymbol{\theta},\tau)} \quad \text{(likelihood)} \\ &+ < \log p(\boldsymbol{\theta}) >_{q(\boldsymbol{\theta})} + < logp(\boldsymbol{W}|\boldsymbol{a})p(\boldsymbol{a}) >_{q(\boldsymbol{a})} \quad \text{(priors)} \\ &- < \log q(\boldsymbol{\theta}, \boldsymbol{a}, \tau) >_{q(\boldsymbol{\theta},\boldsymbol{a},\tau)} \end{aligned}$$

- Assumption 1: Mean-field approximation $q(\boldsymbol{\theta}, \boldsymbol{a}, \tau) \approx q(\boldsymbol{\theta})\, q(\tau) q(\boldsymbol{a})$ (Wainwright 2008)
- Assumption 2: Linearize $u(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{\theta}) \approx \boldsymbol{u}(\boldsymbol{\mu}) + \boldsymbol{G}\boldsymbol{W}\boldsymbol{\theta}$

Algorithm $O(N)$:

0. Initialize $\boldsymbol{\mu}, \boldsymbol{W}$
1. Repeat until convergence:
   - Fix $\boldsymbol{\mu}, \boldsymbol{W}$ and update $q(\boldsymbol{\theta})\, q(\tau), q(\boldsymbol{a})$
   - Fix $\boldsymbol{W}, q(\boldsymbol{\theta})\, q(\tau), q(\boldsymbol{a})$ and update $\boldsymbol{\mu}$
   - Fix $\boldsymbol{\mu}, q(\boldsymbol{\theta})\, q(\tau), q(\boldsymbol{a})$ and update $\boldsymbol{W}$



Numerical Illustration

Example:

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

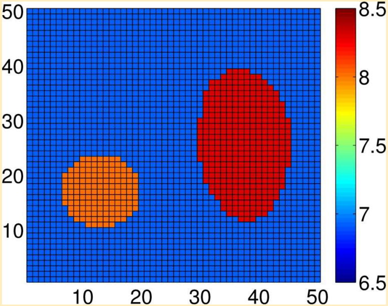
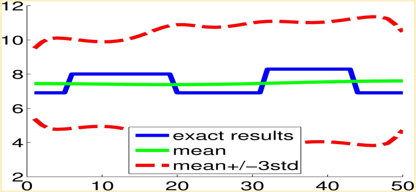


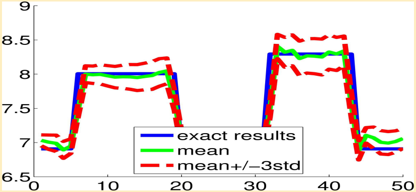
Figure: Ground truth: Log of material parameter c_1

Numerical Illustration

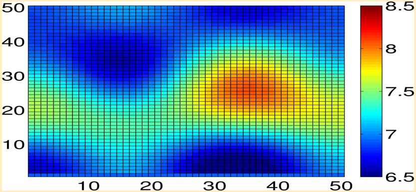
Example:



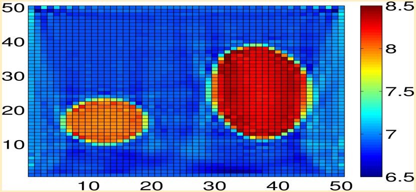
(a) Posterior along diagonal



(b) Posterior along diagonal



(c) Posterior mean



(d) Posterior mean

Figure: (Left) Without (Right) With updating ***W***

Numerical Illustration

Example:

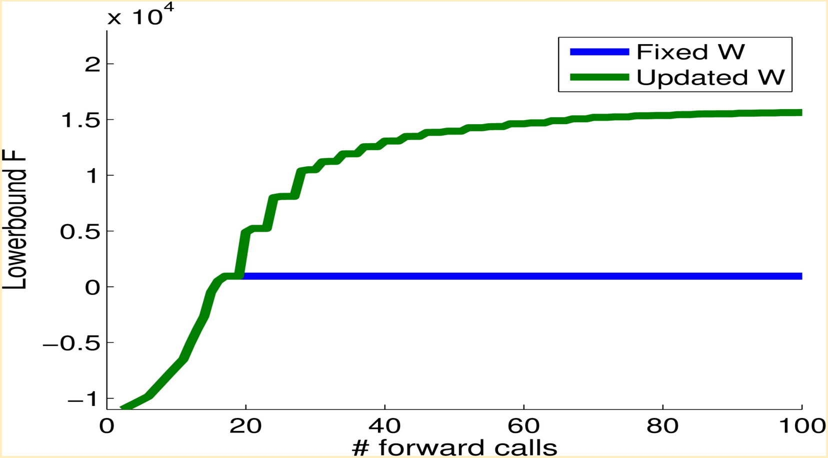
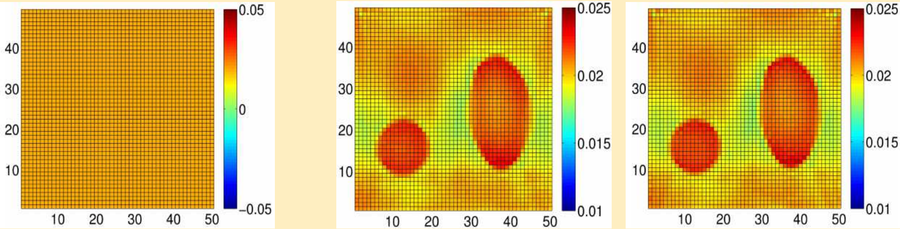


Figure: Evolution of variational objective $\mathcal{F}$

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 $\boldsymbol{W}$

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

$$\boldsymbol{\Psi}|m = \boldsymbol{\mu}_m + \boldsymbol{W}_m\boldsymbol{\theta}_m,$$
$$\rightarrow p(\boldsymbol{\Psi}|\hat{\boldsymbol{u}}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\boldsymbol{\Psi}; \boldsymbol{\mu}_m + \boldsymbol{W}_m\, \boldsymbol{\mu}_{\theta_m}, \boldsymbol{W}_m \boldsymbol{S}_{\theta_m} \boldsymbol{W}_m^T)$$

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

Conclusion & Extensions

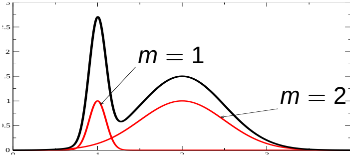
- Variational Bayesian methods offer comparable accuracy and much greater efficiency as compared to sampling (MCMC/SMC) methods
- By approximating the log-evidence one can obtain automatic regularization and enable significant dimensionality reduction.
- Adaptivity:
  - incorporate data sequentially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

$$\boldsymbol{\Psi}|m = \boldsymbol{\mu}_m + \boldsymbol{W}_m \boldsymbol{\theta}_m,$$
$$\rightarrow p(\boldsymbol{\Psi}|\hat{\boldsymbol{u}}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\boldsymbol{\Psi}; \boldsymbol{\mu}_m + \boldsymbol{W}_m \, \boldsymbol{\mu}_{\theta_m}, \boldsymbol{W}_m \boldsymbol{S}_{\theta_m} \boldsymbol{W}_m^T)$$

    - 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



$$\boldsymbol{\Psi}|m = \boldsymbol{\mu}_m + \boldsymbol{W}_m \boldsymbol{\theta}_m,$$
$$\rightarrow p(\boldsymbol{\Psi}|\hat{\boldsymbol{u}}) = \sum_{m=1}^{M} \pi_m \mathcal{N}(\boldsymbol{\Psi}; \boldsymbol{\mu}_m + \boldsymbol{W}_m \, \boldsymbol{\mu}_{\theta_m}, \boldsymbol{W}_m \boldsymbol{S}_{\theta_m} \boldsymbol{W}_m^T)$$

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