Lecture #3: Bayesian modeling and computation for inverse problems

Youssef Marzouk

Department of Aeronautics and Astronautics
Center for Computational Engineering
Statistics and Data Science Center

Massachusetts Institute of Technology
http://uqgroup.mit.edu
ymarz@mit.edu

19–22 March 2018

Agenda

Plan for the lectures:

1. Lectures 1–2: Bayesian inference and MCMC foundations
   - Bayesian modeling
   - MCMC algorithms and demos
2. Lectures 3–4: Bayesian approach to inverse problems
   - Elements of a Bayesian inverse problem formulation
   - Linear–Gaussian problems in detail
   - Surrogate modeling and likelihood approximations
   - Dimension reduction
3. Lecture 4+: Bayesian optimal experimental design *or some other topic TBD*

Inverse problems in the Bayesian perspective

How do **inverse problems** differ from generic **parameter estimation** problems?

Typical characteristics of inverse problems:

- Observations **indirectly** related to parameters
- Observations (perhaps) limited in number
- Observations are noisy
- Parameters are high dimensional (in principle, *functions*)

Inverse problems

Key building block: the **forward model**

- A (deterministic) operator G that maps parameters θ to predictions of the observations
- Enters the likelihood function $p(y|\theta)$, when combined with a suitable statistical model
- (Simplest) example:

$$y = G(\theta) + \epsilon, \ \epsilon \sim N(0, \Gamma_{\text{obs}})$$
$$\text{then } y|\theta \sim N\left(G(\theta), \Gamma_{\text{obs}}\right)$$

Here ϵ represents observational error and (crudely) error in the forward model

Inverse problems

Key building block: the **forward model**

- A (deterministic) operator G that maps parameters θ to predictions of the observations
- Enters the likelihood function $p(y|\theta)$, when combined with a suitable statistical model
- (Simplest) example:

$$y = G(\theta) + \epsilon, \;\; \epsilon \sim N(0, \Gamma_{\text{obs}})$$
$$\text{then } y|\theta \sim N\left(G(\theta), \Gamma_{\text{obs}}\right)$$

  Here ϵ represents observational error and (crudely) error in the forward model

Inverse problems

Why are inverse problems difficult?

Classically **ill-posed**:

- No solution may match the data **(existence)**
  - Linear case, $G \in \mathbb{R}^{m \times n}$: $\exists$ a non-trivial left nullspace $\text{Ker}(G^\top)$
- Many solutions may match the data **(uniqueness)**
  - Linear case, $G \in \mathbb{R}^{m \times n}$: $\exists$ a non-trivial nullspace $\text{Ker}(G)$
- Ill-conditioning or **instability**: small changes in data y can lead to large changes in (unregularized) estimates $\hat{\theta}(y)$
  - Linear case, $G \in \mathbb{R}^{m \times n}$: singular values $\sigma_i(G)$ decay rapidly to zero
  - Yields sensitivity to noise

Deterministic approach: **regularization**!

Inverse problems

Why are inverse problems difficult?

Classically **ill-posed**:

- No solution may match the data **(existence)**
  - Linear case, $G \in \mathbb{R}^{m \times n}$: $\exists$ a non-trivial left nullspace $\text{Ker}(G^\top)$
- Many solutions may match the data **(uniqueness)**
  - Linear case, $G \in \mathbb{R}^{m \times n}$: $\exists$ a non-trivial nullspace $\text{Ker}(G)$
- Ill-conditioning or **instability**: small changes in data y can lead to large changes in (unregularized) estimates $\hat{\theta}(y)$
  - Linear case, $G \in \mathbb{R}^{m \times n}$: singular values $\sigma_i(G)$ decay rapidly to zero
  - Yields sensitivity to noise

Deterministic approach: **regularization**!

Inverse problems

Classical **regularization approach** to inverse problems *(an example)*:

$$\hat{\theta}(y) = \arg\min \|y - G(\theta)\|^2_{\Gamma_{obs}} + \lambda \mathcal{R}(\theta)$$

- Without regularization term $\lambda\mathcal{R}$, and under an additive Gaussian noise assumption, this would be the maximum likelihood estimate: an ill-posed problem (may lack uniqueness and stability)!
- With regularization, can be interpreted as a penalized ML estimate
- **Enormous** literature on the design of suitable regularization functionals $\mathcal{R}$
  - Basic example: zeroth-order Tikhonov, $\mathcal{R}(\theta) = \|\theta\|_2^2$
- Also, many techniques for selecting regularization parameter λ
- Instead we take a Bayesian statistical perspective. . .

Inverse problems

Classical **regularization approach** to inverse problems *(an example)*:

$$\hat{\theta}(y) = \arg\min \|y - G(\theta)\|^2_{\Gamma_{obs}} + \lambda \mathcal{R}(\theta)$$

- Without regularization term $\lambda\mathcal{R}$, and under an additive Gaussian noise assumption, this would be the maximum likelihood estimate: an ill-posed problem (may lack uniqueness and stability)!
- With regularization, can be interpreted as a penalized ML estimate
- **Enormous** literature on the design of suitable regularization functionals $\mathcal{R}$
  - Basic example: zeroth-order Tikhonov, $\mathcal{R}(\theta) = \|\theta\|_2^2$
- Also, many techniques for selecting regularization parameter λ
- Instead we take a Bayesian statistical perspective...

Bayesian inverse problems

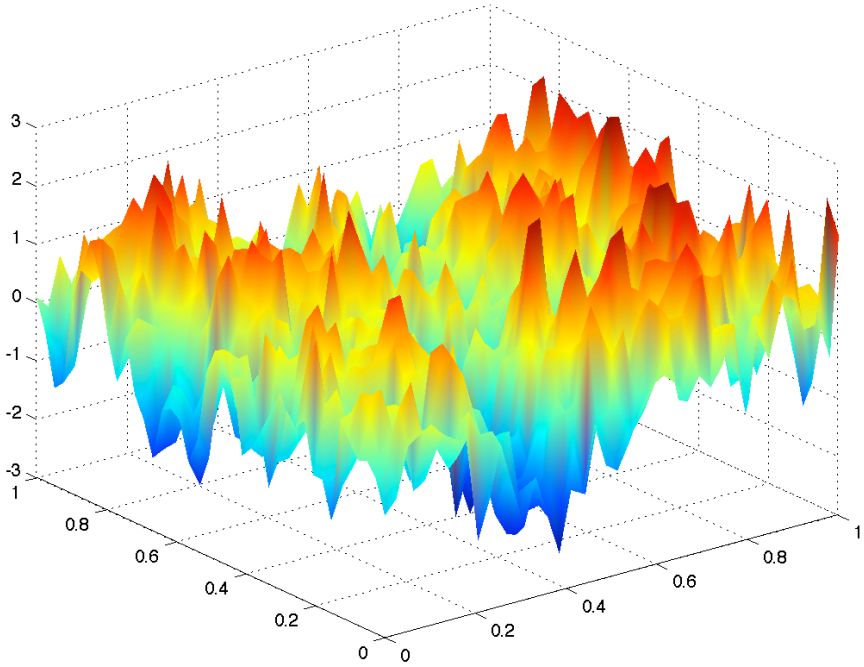
Prior distributions

- In inverse problems, prior information plays a key role. Broadly, priors serve as regularizers.
- Intuitive idea: assign lower probability to neighborhoods of θ that you don't expect to see, higher probability to neighborhoods of θ that you *do* expect to see
- Examples
  1. Gaussian processes with specified covariance kernel
  2. Gaussian Markov random fields
  3. Gaussian priors derived from differential operators
  4. Hierarchical priors
  5. Besov space priors
  6. Other non-Gaussian priors
  7. Higher-level representations (objects, marked point processes)

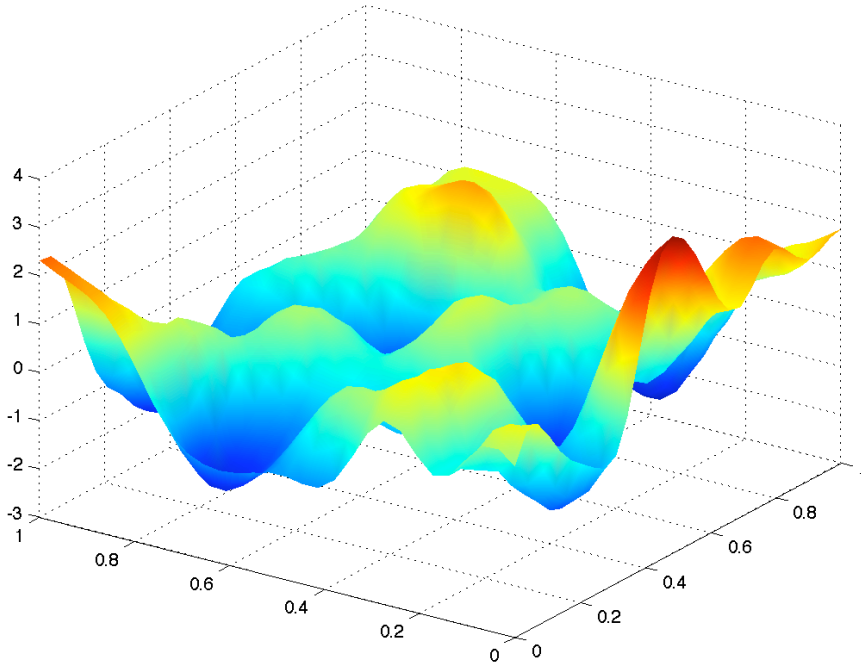
Gaussian process priors

- Key idea: any finite-dimensional distribution of the stochastic process $\theta(\mathbf{x}, \omega) : D \times \Omega \to \mathbb{R}$ is multivariate normal.
- In other words: $\theta(\mathbf{x}, \omega)$ is a collection of jointly Gaussian random variables, indexed by $\mathbf{x}$
- Specify via *mean function* and *covariance function*
  - $\mathbb{E}[\theta(\mathbf{x})] = \mu(\mathbf{x})$
  - $\mathbb{E}[(\theta(\mathbf{x}) - \mu)(\theta(\mathbf{x}') - \mu)] = C(\mathbf{x}, \mathbf{x}')$
- Smoothness of process is controlled by behavior of covariance function as $\mathbf{x}' \to \mathbf{x}$
- Common symmetries:
  - Stationarity: $C(\mathbf{x}, \mathbf{x}') = \widetilde{C}(\boldsymbol{\tau})$, where $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$
  - Isotropy: $C(\mathbf{x}, \mathbf{x}') = \bar{C}(\tau)$, where $\tau = \|\mathbf{x} - \mathbf{x}'\|$

Example: stationary Gaussian random fields



(exponential covariance kernel)



(Gaussian covariance kernel)

Both are $\theta(\mathbf{x}, \omega) : D \times \Omega \to \mathbb{R}$, with $D = [0, 1]^2$.

Gaussian Markov random fields

- Key idea: discretize space and specify a *sparse* inverse covariance ("precision") matrix $\mathbf{W}$

$$p(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\gamma \boldsymbol{\theta}^T \mathbf{W} \boldsymbol{\theta}\right)$$

  where γ controls scale
- Full conditionals $p(\theta_i|\theta_{\sim i})$ are available analytically and may simplify dramatically.
- Represent **conditional independence** structure via an undirected graphical model
- Example: $\mathbb{E}\left[\theta_i|\theta_{\sim i}\right]$ is just an average of site i's nearest neighbors

Priors through differential operators

- Key idea: return to infinite-dimensional setting; again penalize roughness in $\theta(\mathbf{x})$
- Stuart 2010: define the prior using fractional negative powers of the Laplacian $\mathcal{A} = -\Delta$:

$$\theta \sim \mathcal{N}\left(\theta_0, \beta \mathcal{A}^{-\alpha}\right)$$

- Sufficiently large α ($\alpha > d/2$), along with conditions on the likelihood, ensure that posterior measure is well defined

GPs, GMRFs, and SPDEs

In fact, all three "types" of Gaussian priors just described are closely connected.

- Linear (fractional) SPDE:
  $$\left(\kappa^2 - \Delta\right)^{\beta/2} \theta(\mathbf{x}) = \mathcal{W}(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^d,\ \beta = \nu + d/2,\ \kappa > 0, \nu > 0$$
- Then $\theta(\mathbf{x})$ is a Gaussian field with Matérn covariance:
  $$C(\mathbf{x}, \mathbf{x}') = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} (\kappa \|\mathbf{x} - \mathbf{x}'\|)^\nu K_\nu (\kappa \|\mathbf{x} - \mathbf{x}'\|)$$
- Covariance kernel is Green's function of differential operator
  $\left(\kappa^2 - \Delta\right)^{\beta} C(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}')$
- $\nu = 1/2$ equivalent to exponential covariance; $\nu \to \infty$ equivalent to squared exponential covariance
- Can construct a discrete GMRF that approximates the solution of SPDE (See Lindgren, Rue, Lindström JRSSB 2011.)

Hierarchical Gaussian priors

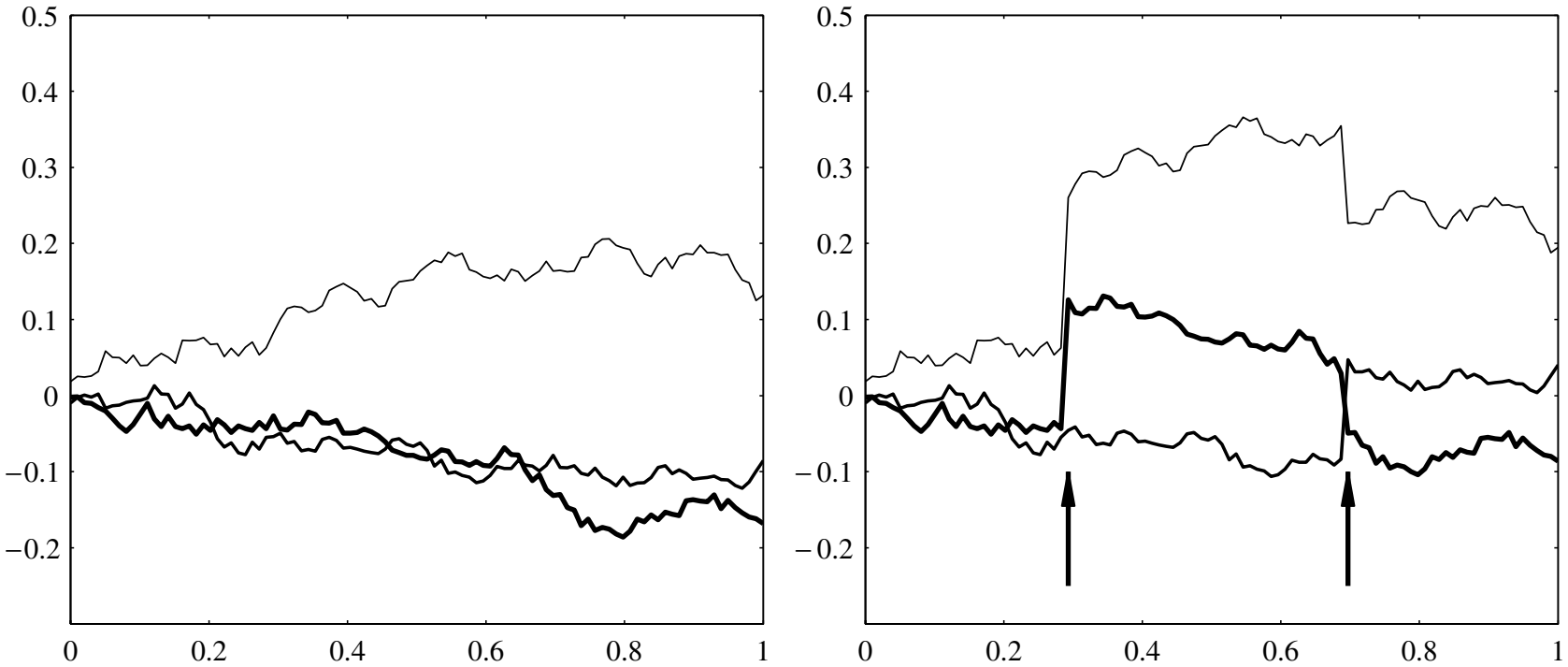


Figure 1. Three realization drawn from the prior (6) with constant variance $\theta_j = \theta_0$ (left) and from the corresponding prior where the variance is 100–fold at two points indicated by arrows (right).

Calvetti & Somersalo, *Inverse Problems* 24 (2008) 034013.

Hierarchical Gaussian priors

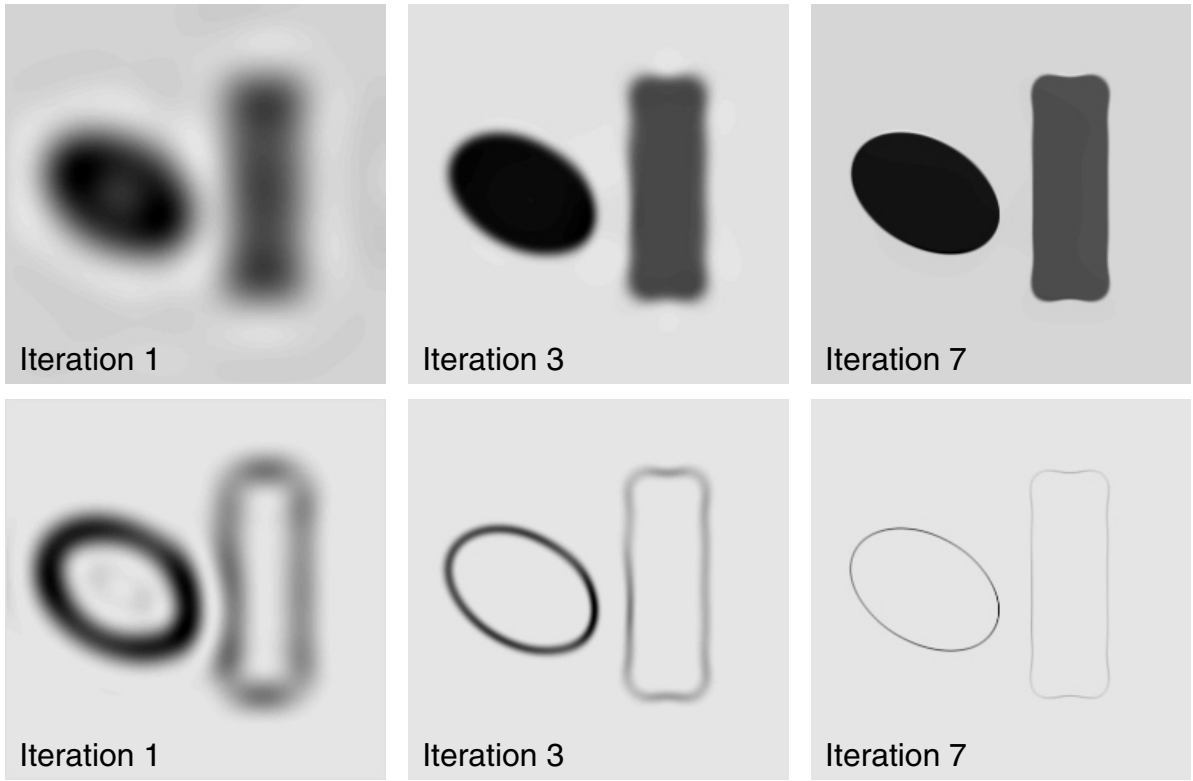


Figure 5. Approximation of the MAP estimate of the image (top row) and of the variance (bottom row) after 1, 3 and 5 iteration of the cyclic algorithm when using the CGLS method to compute the updated of the image at each iteration step

Calvetti & Somersalo, *Inverse Problems* 24 (2008) 034013.

Non-Gaussian priors

- Besov space $B^s_{pq}(\mathbb{T})$:

$$\theta(x) = c_0 + \sum_{j=0}^{\infty} \sum_{h=0}^{2^j-1} w_{j,h} \psi_{j,h}(x)$$

and

$$\|\theta\|_{B^s_{pq}(\mathbb{T})} := \left(|c_0|^q + \sum_{j=0}^{\infty} 2^{jq(s+\frac{1}{2}-\frac{1}{p})} \left(\sum_{h=0}^{2^j-1} |w_{j,h}|^p \right)^{q/p} \right)^{1/q} < \infty.$$

- Consider $p = q = s = 1$:

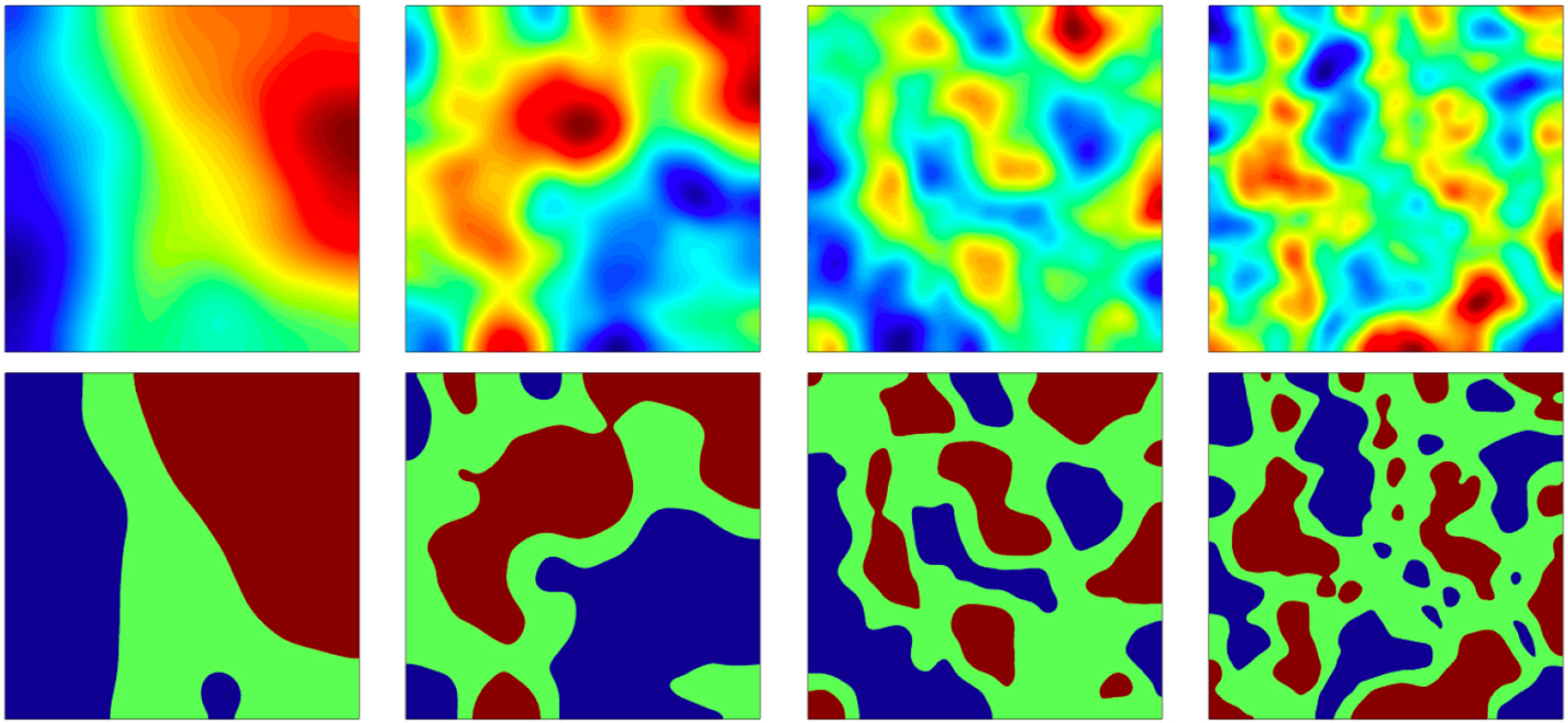
$$\|\theta\|_{B^1_{11}(\mathbb{T})} = |c_0| + \sum_{j=0}^{\infty} \sum_{h=0}^{2^j-1} 2^{j/2} |w_{j,h}|.$$

Then the distribution of θ is a *Besov prior* if αc_0 and $\alpha 2^{j/2} w_{j,h}$ are independent and Laplace(1).

- Loosely, $\pi(\theta) = \exp\left(-\alpha \|\theta\|_{B^1_{11}(\mathbb{T})}\right)$.

Non-Gaussian priors

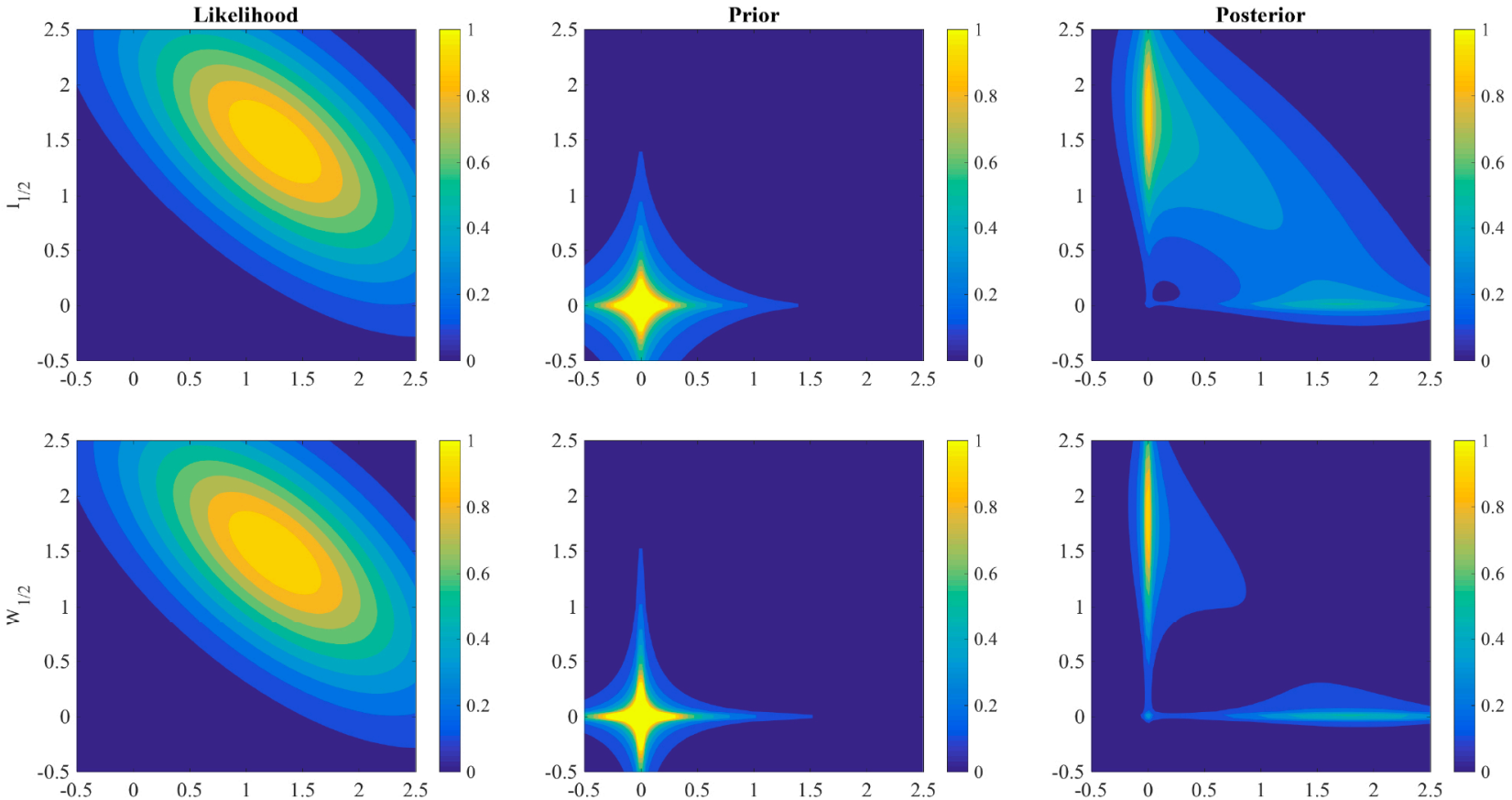
Level set representations



Dunlop, Iglesias, & Stuart, *Stat. Comp.* 27 (2017), 1555–1584.

Non-Gaussian priors

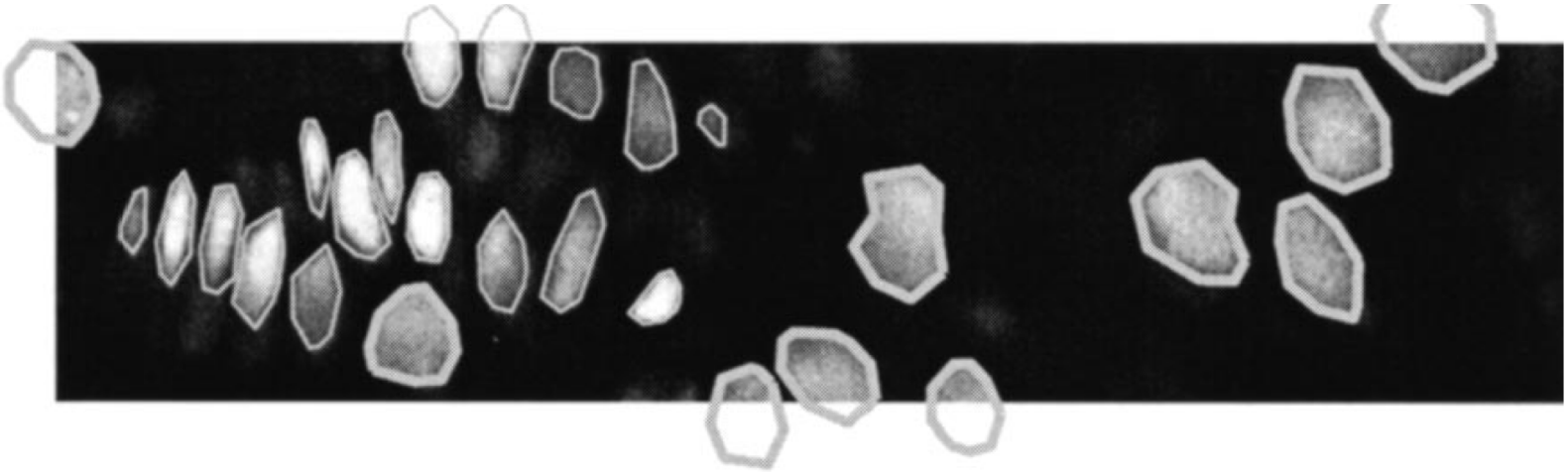
Heavy-tailed priors and sample sparsity



Hosseini, *SIAM JUQ* 5 (2017), 1024–1060.

Higher-level representations

Marked point processes, and more:



Rue & Hurn, *Biometrika* 86 (1999), 649–660.

Bayesian inference

Hierarchical modeling

- One of the key flexibilities of the Bayesian construction!
- Hierarchical modeling has important implications for the design of efficient MCMC samplers
- Examples:
  1. Unknown noise variance
  2. Unknown variance of a Gaussian process prior (cf. choosing the **regularization** parameter)
  3. Many more, as dictated by the physical models at hand

Example: prior variance hyperparameter in an inverse diffusion problem

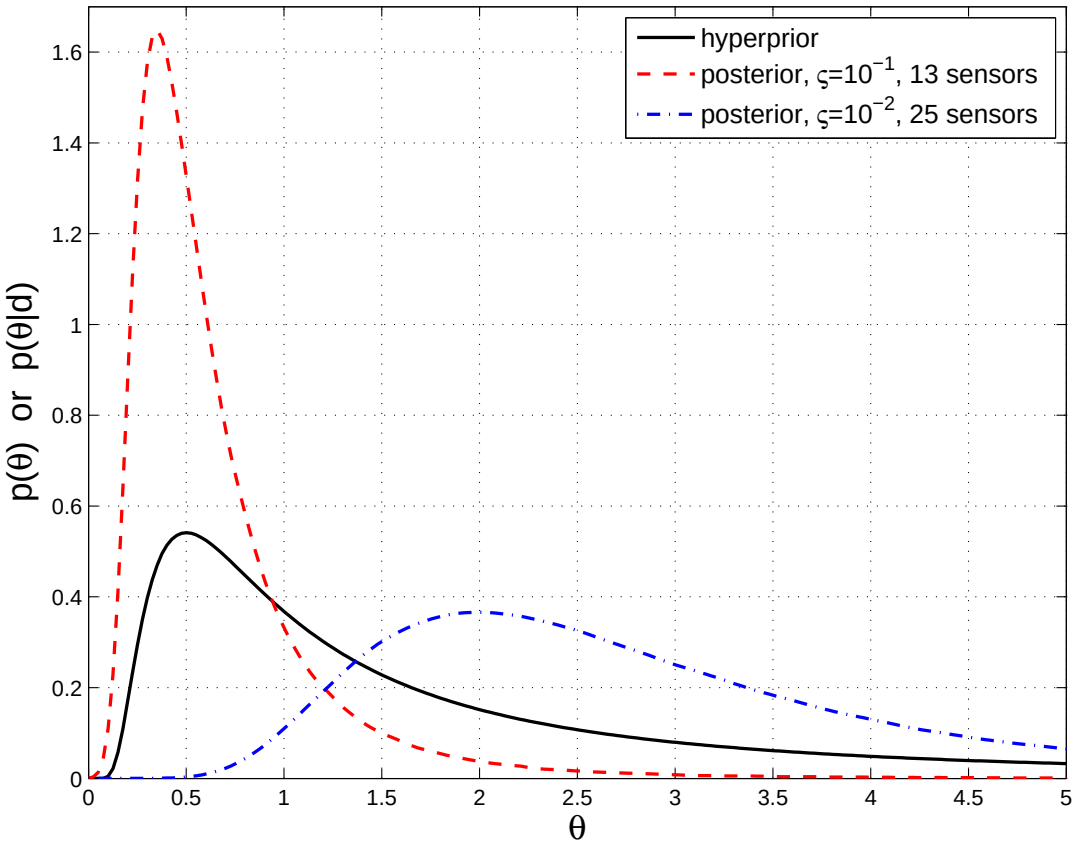


Figure: Posterior marginal density of the variance hyperparameter σ^2, versus quality of data (number and noise variance ς^2), contrasted with its prior density. "Regularization" $\lambda \propto \varsigma^2/\sigma^2$.

The linear Gaussian model

A key building-block problem:

- Parameters $\theta \in \mathbb{R}^n$, observations $y \in \mathbb{R}^m$
- Forward model $f(\theta) = G\theta$, where $G \in \mathbb{R}^{m \times n}$
- Additive noise yields observations: $y = G\theta + \epsilon$
- $\epsilon \sim N(0, \Gamma_{\text{obs}})$ and is independent of θ
- Endow θ with a Gaussian prior, $\theta \sim N(\mu_{\text{pr}}, \Gamma_{\text{pr}})$.

Posterior probability density

$$p(\theta|y) \propto p(y|\theta)p(\theta) = L(\theta)p(\theta) \propto \exp\left(-\frac{1}{2}(y - G\theta)^\top \Gamma_{\text{obs}}^{-1}(y - G\theta)\right)$$

$$\times \exp\left(-\frac{1}{2}(\theta - \mu_{\text{pr}})^\top \Gamma_{\text{pr}}^{-1}(\theta - \mu_{\text{pr}})\right)$$

$$\propto \exp\left(-\frac{1}{2}(\theta - \mu_{\text{pos}})^\top \Gamma_{\text{pos}}^{-1}(\theta - \mu_{\text{pos}})\right)$$

The linear Gaussian model

A key building-block problem:

- Parameters $\theta \in \mathbb{R}^n$, observations $y \in \mathbb{R}^m$
- Forward model $f(\theta) = G\theta$, where $G \in \mathbb{R}^{m \times n}$
- Additive noise yields observations: $y = G\theta + \epsilon$
- $\epsilon \sim N(0, \Gamma_{\text{obs}})$ and is independent of θ
- Endow θ with a Gaussian prior, $\theta \sim N(\mu_{\text{pr}}, \Gamma_{\text{pr}})$.

Posterior probability density

$$
\begin{aligned}
p(\theta|y) \quad &\propto \quad p(y|\theta)p(\theta) = L(\theta)p(\theta) \propto \exp\left(-\frac{1}{2}\left(y - G\theta\right)^\top \Gamma_{\text{obs}}^{-1}\left(y - G\theta\right)\right) \\
&\qquad \times \exp\left(-\frac{1}{2}(\theta - \mu_{\text{pr}})^\top \Gamma_{\text{pr}}^{-1}(\theta - \mu_{\text{pr}})\right) \\
&\propto \quad \exp\left(-\frac{1}{2}\left(\theta - \mu_{\text{pos}}\right)^\top \Gamma_{\text{pos}}^{-1}\left(\theta - \mu_{\text{pos}}\right)\right)
\end{aligned}
$$

The linear Gaussian model

- Posterior is again Gaussian:

$$\begin{aligned}\Gamma_{\text{pos}} &= \left(G^\top \Gamma_{\text{obs}}^{-1} G + \Gamma_{\text{pr}}^{-1}\right)^{-1} \\ &= \Gamma_{\text{pr}} - \Gamma_{\text{pr}} G^\top \left(G \Gamma_{\text{pr}} G^\top + \Gamma_{\text{obs}}\right)^{-1} G \Gamma_{\text{pr}} \\ &= (I - K G)\, \Gamma_{\text{pr}}\end{aligned}$$

$$\mu_{\text{pos}} = \Gamma_{\text{pos}} \left(G^\top \Gamma_{\text{obs}}^{-1} y + \Gamma_{\text{pr}}^{-1} \mu_{\text{pr}}\right)$$

- In the context of filtering, K is known as the (optimal) Kalman gain.
- $H := G^\top \Gamma_{\text{obs}}^{-1} G$ is the Hessian of the negative log-likelihood
- How does low rank of H affect the structure of the posterior? How does H interact with the prior?

Likelihood-informed directions

- Consider the Rayleigh ratio

$$\mathcal{R}(w) = \frac{w^\top H w}{w^\top \Gamma_{\text{pr}}^{-1} w}.$$

  When $\mathcal{R}(w)$ is large, likelihood dominates the prior in direction w.

- The ratio is maximized by solutions to the generalized eigenvalue problem

$$Hw = \lambda \Gamma_{\text{pr}}^{-1} w.$$

- The posterior covariance can be written as a negative update along these "likelihood-informed" directions, and approximations can be obtained using the r largest eigenvalues:

$$\Gamma_{\text{pos}} = \Gamma_{\text{pr}} - \sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top \approx \Gamma_{\text{pr}} - \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top$$

Likelihood-informed directions

- Consider the Rayleigh ratio

$$\mathcal{R}(w) = \frac{w^\top H w}{w^\top \Gamma_{\text{pr}}^{-1} w}.$$

  When $\mathcal{R}(w)$ is large, likelihood dominates the prior in direction w.

- The ratio is maximized by solutions to the generalized eigenvalue problem

$$Hw = \lambda \Gamma_{\text{pr}}^{-1} w.$$

- The posterior covariance can be written as a negative update along these "likelihood-informed" directions, and approximations can be obtained using the r largest eigenvalues:

$$\Gamma_{\text{pos}} = \Gamma_{\text{pr}} - \sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top \approx \Gamma_{\text{pr}} - \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top$$

Likelihood-informed directions

- Consider the Rayleigh ratio

$$\mathcal{R}(w) = \frac{w^\top H w}{w^\top \Gamma_{\text{pr}}^{-1} w}.$$

  When $\mathcal{R}(w)$ is large, likelihood dominates the prior in direction w.

- The ratio is maximized by solutions to the generalized eigenvalue problem

$$Hw = \lambda \Gamma_{\text{pr}}^{-1} w.$$

- The posterior covariance can be written as a negative update along these "likelihood-informed" directions, and approximations can be obtained using the r largest eigenvalues:

$$\Gamma_{\text{pos}} = \Gamma_{\text{pr}} - \sum_{i=1}^{n} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top \approx \Gamma_{\text{pr}} - \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top$$

Optimality results for $\widehat{\Gamma}_{\text{pos}}$

- The approximation

$$\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top$$

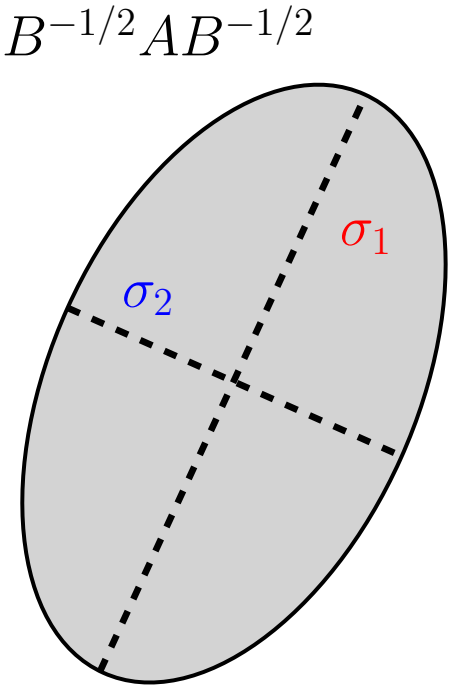
is **optimal** in a class of loss functions $L(\widehat{\Gamma}_{\text{pos}}, \Gamma_{\text{pos}})$ for approximations of form $\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top$, where $\text{rank}(K) \leq r$.
[Spantini *et al. SIAM J. Sci. Comp.* 2016]

A metric between covariance matrices

Fisher-Rao/Förstner metric

Let $A, B \succ 0$, and (σ_i) be the eigenvalues of (A, B). Then:

$$\begin{aligned} d^2_{\mathcal{F}}(A,B) &= \operatorname{tr}\left[\ln^2\left(B^{-\frac{1}{2}}AB^{-\frac{1}{2}}\right)\right] \\ &= \sum_i \ln^2(\sigma_i) \end{aligned}$$



- Compare curvatures: $\sup_u \frac{u^\top A u}{u^\top B u} = \sigma_1$
- Unique geodesic distance on Sym^+ satisfying invariances:
  $d_{\mathcal{F}}(A,B) = d_{\mathcal{F}}\left(A^{-1}, B^{-1}\right)$ $\qquad$ $d_{\mathcal{F}}(A,B) = d_{\mathcal{F}}\left(MAM^\top, MBM^\top\right)$
- Frobenius $d_{\mathrm{F}}(A,B) = \|A-B\|_{\mathrm{F}}$ does not share these properties

Remarks on the optimal approximation

$$\widehat{\Gamma}^*_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top, \qquad KK^\top = \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top$$

- $\widehat{\Gamma}^*_{\text{pos}}$ is the minimizer of $d_{\mathcal{F}}$ between Γ_{pos} and an element of $\mathcal{M}_r = \{\Gamma_{\text{pr}} - KK^\top : \text{rank}(K) \leq r\}$.
- $\widehat{\Gamma}^*_{\text{pos}}$ also minimizes the Hellinger distance and the Kullback–Leibler divergence between $\mathcal{N}(\mu_{\text{pos}}(y), \Gamma \in \mathcal{M}_r)$ and $\mathcal{N}(\mu_{\text{pos}}(y), \Gamma_{\text{pos}})$.
- These results can also be used to devise optimal approximations for the posterior mean (e.g., a low-rank matrix applied to the data y)
  - Minimize Bayes risk for squared-error loss weighted by the posterior precision

Remarks on the optimal approximation

$$\widehat{\Gamma}^*_{\text{pos}} = \Gamma_{\text{pr}} - KK^\top, \qquad KK^\top = \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i} w_i w_i^\top$$

- $\widehat{\Gamma}^*_{\text{pos}}$ is the minimizer of $d_{\mathcal{F}}$ between Γ_{pos} and an element of $\mathcal{M}_r = \{\Gamma_{\text{pr}} - KK^\top \,:\, \text{rank}(K) \leq r\}$.
- $\widehat{\Gamma}^*_{\text{pos}}$ also minimizes the Hellinger distance and the Kullback–Leibler divergence between $\mathcal{N}(\mu_{\text{pos}}(y), \Gamma \in \mathcal{M}_r)$ and $\mathcal{N}(\mu_{\text{pos}}(y), \Gamma_{\text{pos}})$.
- These results can also be used to devise optimal approximations for the posterior mean (e.g., a low-rank matrix applied to the data y)
  - Minimize Bayes risk for squared-error loss weighted by the posterior precision

Remarks on the optimal approximation

$$\widehat{\Gamma}^{*}_{\text{pos}} = \Gamma_{\text{pr}} - KK^{\top}, \qquad KK^{\top} = \sum_{i=1}^{r} \frac{\lambda_i}{1+\lambda_i} w_i w_i^{\top}$$

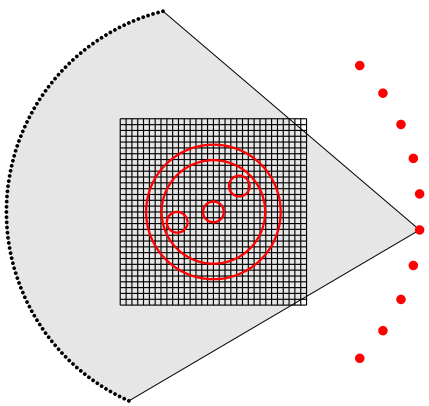
- The form of the optimal update is widely used (Flath *et al.* 2011)
- Compute with Lanczos, randomized SVD, etc.
- Directions $\widetilde{w}_i = \Gamma_{\text{pr}}^{-1} w_i$ *maximize* the **relative** difference between prior and posterior variance:

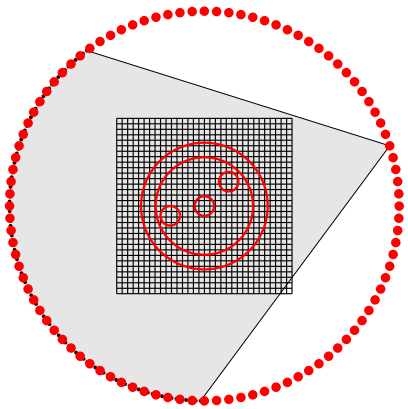
$$\frac{\mathbb{V}\text{ar}\left(\widetilde{w}_i^{\top} x\right) - \mathbb{V}\text{ar}\left(\widetilde{w}_i^{\top} x \,|\, y\right)}{\mathbb{V}\text{ar}\left(\widetilde{w}_i^{\top} x\right)} = \frac{\lambda_i}{1+\lambda_i}$$

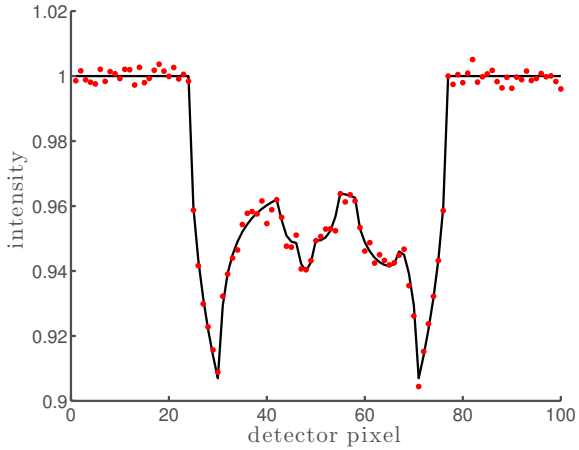
- Using the Frobenius norm as a loss would instead yield directions of greatest **absolute** difference between prior and posterior variance.

Example: computerized tomography

X-rays travel from sources to detectors through an object of interest. Intensities from the sources are measured at the detectors, and the goal is to reconstruct the density of the object.



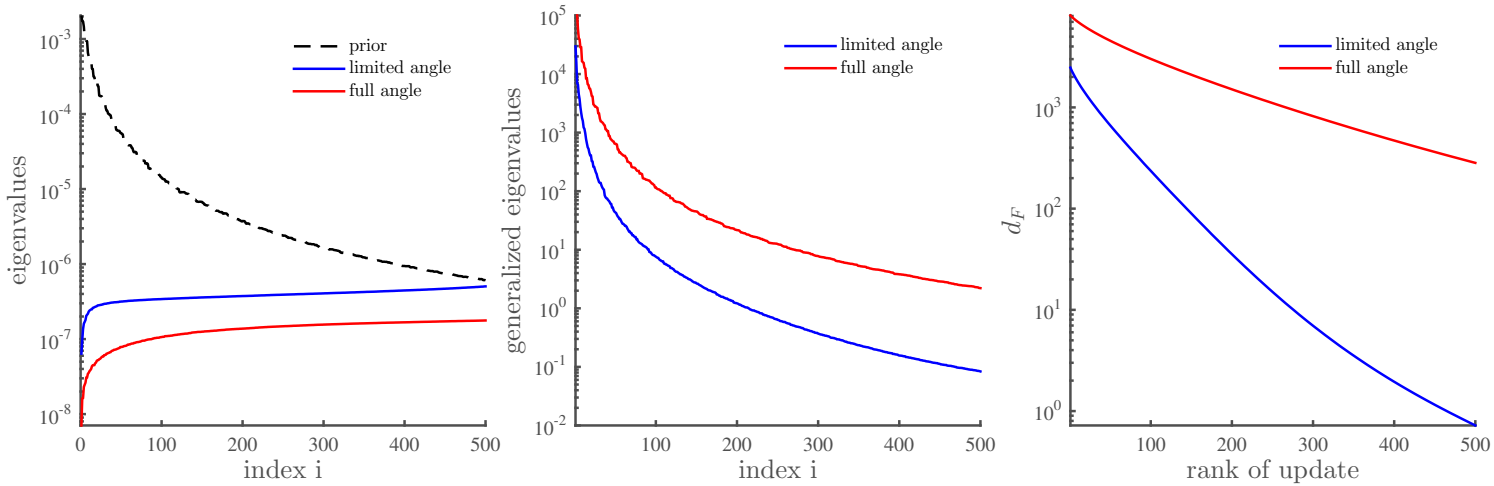




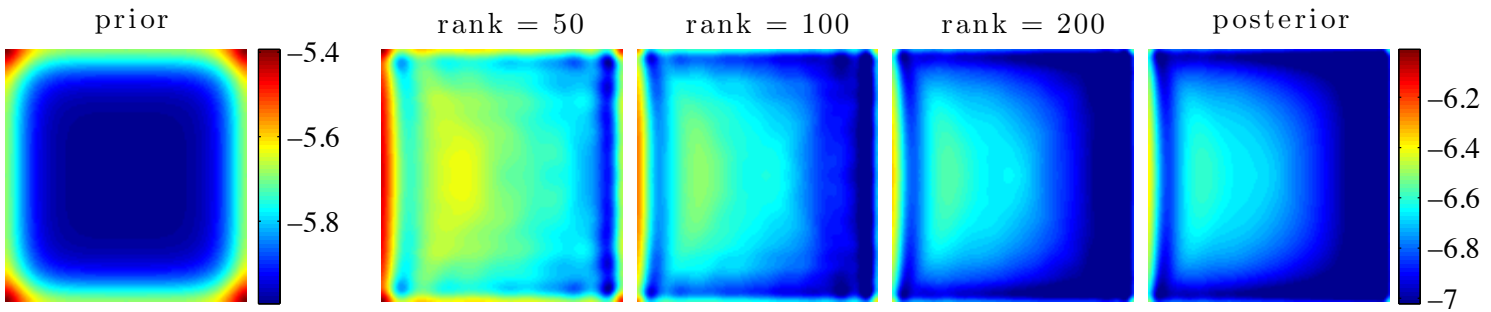
This synthetic example is motivated by real-time X-ray imaging of logs entering a sawmill, for automatic quality control (see `http://finnos.fi`)

Example: computerized tomography

Weaker data $\rightarrow$ faster decay of generalized eigenvalues $\rightarrow$ lower order approximations possible.



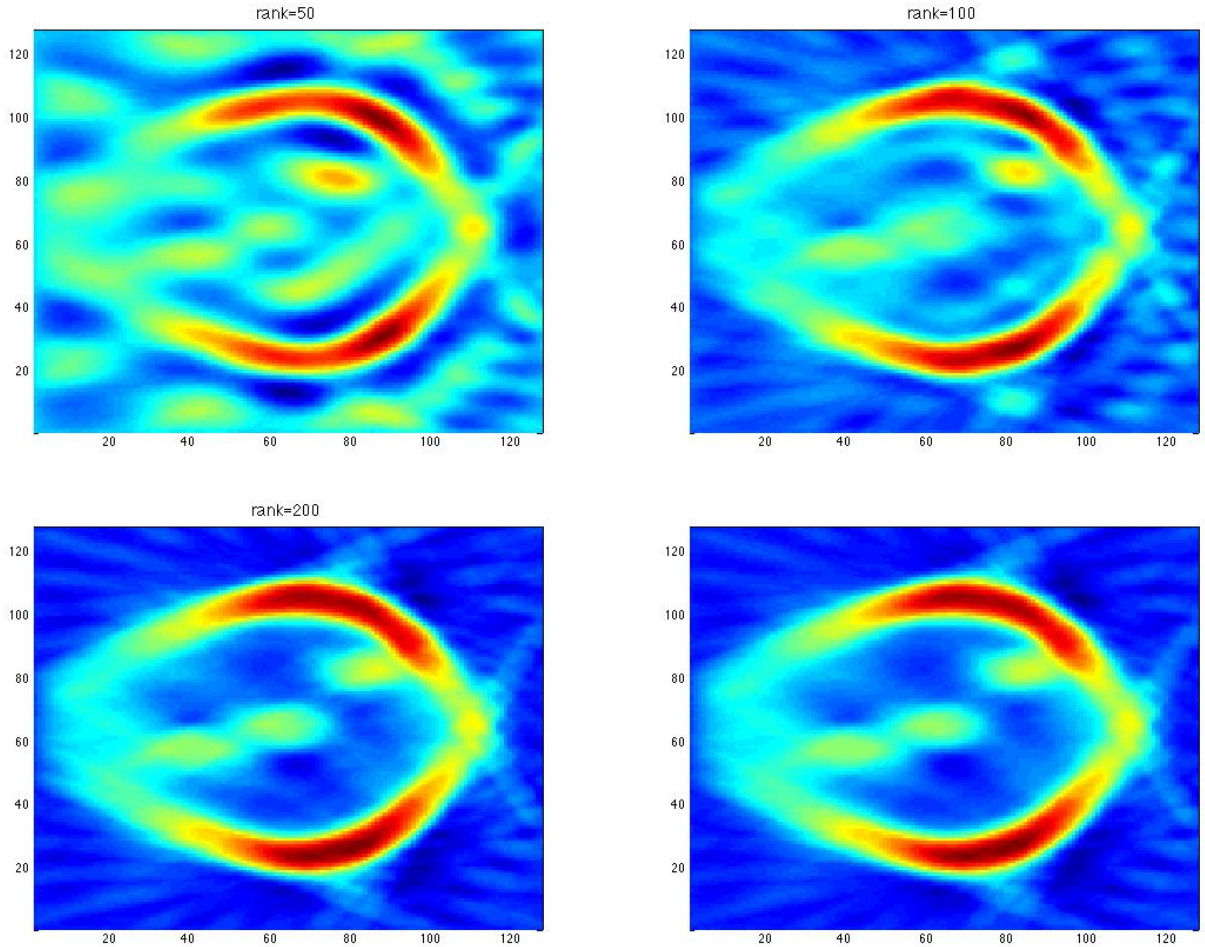
In the limited angle case, roughly $r = 200$ is enough to get a good approximation (with full angle $r \approx 800$ needed). **Variance fields:**



Example: computerized tomography

Approximation of the posterior mean:

$$\mu_{\text{pos}}(y) = \Gamma_{\text{pos}}\, G^\top \Gamma_{\text{obs}}^{-1}\, y \approx \sum_{i=1}^{r} \delta_i (1+\delta_i^2)^{-1} w_i v_i^\top\, y =: A_r\, y$$



Note: pre-computing A_r offline enables fast reconstructions for repeated data

Questions to answer

- How to simulate from or explore general *non-Gaussian* posterior distributions?
- How to make Bayesian inference computationally tractable when the forward model is expensive (e.g., a PDE) and the parameters are high- or infinite-dimensional?

MCMC in infinite dimensions

- Would like to construct a well-defined MCMC sampler for **functions** $u \in \mathcal{H}$.
- First, the posterior measure μ_y should be a well-defined probability measure on $\mathcal{H}$ (see Stuart *Acta Numerica* 2010). For simplicity, let the prior μ_0 be $\mathcal{N}(0, C)$.
- Now let q be the proposal distribution, and consider pair of measures

$$\nu(du, du') = q(u, du')\mu_y(du), \;\; \nu^{\perp}(du, du') = q(u', du)\mu_y(du');$$

- Then the MCMC acceptance probability is

$$\alpha(u_k, u') = \min\left\{1, \frac{d\nu^{\perp}}{d\nu}(u_k, u')\right\}$$

- To define a **valid** transition kernel, we need absolute continuity $\nu^{\perp} \ll \nu$; in turn, this places requirements on the proposal q

MCMC in infinite dimensions

- Would like to construct a well-defined MCMC sampler for **functions** $u \in \mathcal{H}$.
- First, the posterior measure μ_y should be a well-defined probability measure on $\mathcal{H}$ (see Stuart *Acta Numerica* 2010). For simplicity, let the prior μ_0 be $\mathcal{N}(0, C)$.

- Now let q be the proposal distribution, and consider pair of measures
$$\nu(du, du') = q(u, du')\mu_y(du), \;\; \nu^{\perp}(du, du') = q(u', du)\mu_y(du');$$
- Then the MCMC acceptance probability is
$$\alpha(u_k, u') = \min\left\{1, \frac{d\nu^{\perp}}{d\nu}(u_k, u')\right\}$$
- To define a **valid** transition kernel, we need absolute continuity $\nu^{\perp} \ll \nu$; in turn, this places requirements on the proposal q

MCMC in infinite dimensions (cont.)

- One way to produce a valid transition kernel is the preconditioned Crank-Nicolson (pCN) proposal (Cotter *et al.* 2013):

$$u' = (1-\beta^2)^{1/2} u_k + \beta \xi_k, \ \xi_k \sim \mathcal{N}(0, C), \ \beta \in (0,1).$$

- Practical impact: sampling efficiency does not degenerate as discretization of u is refined
- More sophisticated versions: combine pCN with **Hessian/geometry** information, e.g., DILI (dimension-independent likelihood-informed) proposals [Cui, Law, M 2016]
  - Approximations of the (local/linearized) posterior covariance as a **low-rank update** of the prior covariance are essential to scalability
  - Roughly: pCN in directions not informed by the data *(infinitely many)* + preconditioned MALA in the data-informed directions *(finite in number)*

Approximations in MCMC

Efficient sampling is great, but what if each posterior evaluation is very expensive?

Obvious answer: *approximate* the expensive part, e.g., the forward model.

This raises many interesting issues:

- What kind of approximation scheme to use? What properties of the forward model/likelihood are being exploited?
- When to construct the approximation (offline versus online) and what kind of accuracy to demand from it?
- What is the accuracy of the resulting posterior? Bias in posterior estimates? Can/should we correct for these?

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Construction schemes:

- Surrogates accurate over the prior (e.g., convergent in $L^2_{\pi_{\text{prior}}}$ sense) versus *posterior-focused* (and hence data-driven) surrogates
- Constructed offline or *online* during posterior sampling

Errors and correction:

- Convergence rate of the forward model approximation transfers to the posterior it induces [M & Xiu 2009; Cotter, Dashti, Stuart 2010]
- Can always correct using a *delayed-acceptance* scheme [Christen & Fox 2005], but at a price
- Recent work in *asymptotically exact* online approximations [Conrad, M, Pillai, Smith *JASA* 2016]...

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Construction schemes:

- Surrogates accurate over the prior (e.g., convergent in $L^2_{\pi_{\text{prior}}}$ sense) versus *posterior-focused* (and hence data-driven) surrogates
- Constructed offline or *online* during posterior sampling

Errors and correction:

- Convergence rate of the forward model approximation transfers to the posterior it induces [M & Xiu 2009; Cotter, Dashti, Stuart 2010]
- Can always correct using a *delayed-acceptance* scheme [Christen & Fox 2005], but at a price
- Recent work in *asymptotically exact* online approximations [Conrad, M, Pillai, Smith *JASA* 2016]...

Approximations in MCMC

Much work has been done on this topic.

Approximation schemes: coarse-grid PDE models, polynomial expansions, Gaussian process emulators, reduced-basis methods and reduced-order models, simplified physics, etc.

Construction schemes:

- Surrogates accurate over the prior (e.g., convergent in $L^2_{\pi_{\text{prior}}}$ sense) versus *posterior-focused* (and hence data-driven) surrogates
- Constructed offline or *online* during posterior sampling

Errors and correction:

- Convergence rate of the forward model approximation transfers to the posterior it induces [M & Xiu 2009; Cotter, Dashti, Stuart 2010]
- Can always correct using a *delayed-acceptance* scheme [Christen & Fox 2005], but at a price
- Recent work in *asymptotically exact* online approximations [Conrad, M, Pillai, Smith *JASA* 2016]...

Advanced posterior sampling for inverse problems (1)

Disclaimer: this is a hopelessly incomplete list!

- S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White, "MCMC methods for functions: modifying old algorithms to make them faster." *Statistical Science*, 28: 424–446, 2013.
- T. Cui, K. Law, and Y. Marzouk, "Dimension-independent likelihood-informed MCMC." *J. Comp. Phys.* 304: 1090–137, 2016.
- T. Cui, J. Martin, Y. Marzouk, A. Solonen, and A. Spantini, "Likelihood informed dimension reduction for nonlinear inverse problems." *Inverse Problems*, 30: 114015, 2015.
- M. Girolami and B. Calderhead, "Riemann manifold Langevin and Hamiltonian Monte Carlo methods." *J. Roy. Stat. Soc. B*, 73: 123–214, 2011.
- C. Ketelsen, R. Scheichl, A. Teckentrup, "A hierarchical multilevel Markov chain Monte Carlo algorithm with applications to uncertainty quantification in subsurface flow." arXiv:1303.7343, 2013.
- J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion." *SIAM J. Sci. Comp.* 34: A1460–A1487, 2012.

Advanced posterior sampling for inverse problems (2)

- T. A. Moselhy and Y. Marzouk, "Bayesian inference with optimal maps." *J. Comp. Phys.*, 231: 7815–7850, 2012.
- M. Parno, Y. Marzouk, "Transport map accelerated Markov chain Monte Carlo." Preprint, arXiv:1412.5492, 2016.
- N. Petra, J. Martin, G. Stadler, and O. Ghattas, "A computational framework for infinite-dimensional Bayesian inverse problems: Part II. Stochastic Newton MCMC with application to ice sheet inverse problems." *SIAM J. Sci. Comp.*, 36: A1525–A1555, 2014.
- C. Schillings, Ch. Schwab, "Sparse adaptive Smolyak quadratures for Bayesian inverse problems." *Inverse Problems* 29: 065011, 2013.
- A. Spantini, A. Solonen, T. Cui, J. Martin, L. Tenorio, Y. Marzouk, "Optimal low-rank approximations of Bayesian linear inverse problems" *SIAM J. Sci. Comp.* 37: A2451–A2487, 2015.

MCMC with surrogate modeling (1)

Disclaimer: this is a hopelessly incomplete list!

- J. A. Christen and C. Fox, "Markov chain Monte Carlo using an approximation." *J. Comp. Graph. Stat.* 14: 795–810, 2005.
- P. Conrad, Y. Marzouk, N. Pillai, and A. Smith, "Accelerating asymptotically exact MCMC for computationally intensive models via local approximations." *J. Amer. Stat. Assoc.* 111: 1591–160, 2016.
- P. Conrad, Y. Marzouk, N. Pillai, and A. Smith, "Parallel local approximation MCMC for expensive models," Preprint, arXiv:1607.02788, 2016.
- T. Cui, C. Fox, and M. J. O'Sullivan, "Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance Metropolis Hastings algorithm." *Water Resources Research* 47: W10521, 2011.
- T. Cui, Y. Marzouk, and K. E. Willcox, "Data-driven model reduction for the Bayesian solution of inverse problems." *Int. J. Num. Meth. Eng.*, 102: 966–990, 2015.
- T. Cui, Y. Marzouk, and K. E. Willcox, "Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction." *J. Comp. Phys.* 315: 363-387, 2016.

MCMC with surrogate modeling (2)

- V. Hoang, Ch. Schwab, A. Stuart, "Complexity analysis of accelerated MCMC methods for Bayesian inversion." *Inverse Problems* 29: 085010, 2013.
- J. Li and Y. Marzouk, "Adaptive construction of surrogates for the Bayesian solution of inverse problems." *SIAM J. Sci. Comp.*, 36: A1163–A1186, 2014.
- Y. Marzouk, H. Najm, L. Rahn, "Stochastic spectral methods for efficient Bayesian solution of inverse problems." *J. Comp. Phys.* 224: 560–586, 2007.
- Y. Marzouk, H. Najm. "Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems." *J. Comp. Phys.*, 228: 1862–1902, 2009.
- Y. Marzouk, D. Xiu. "A stochastic collocation approach to Bayesian inference in inverse problems." *Comm. Comp. Phys.*, 6(4): 826–847, 2009.