Caltech Division of Engineering & Applied Science

Robust MCMC Sampling with Non-Gaussian and Hierarchical Priors

IPAM, UCLA, November 14, 2017

Matt Dunlop

Victor Chen (Caltech)

Omiros Papaspiliopoulos (ICREA, UPF)

Andrew Stuart (Caltech)

Outline

1. Introduction
2. Series-Based Priors
3. Hierarchical Priors
4. Level Set Priors
5. Numerical Illustrations
6. Conclusions

The Inverse Problem

Problem Statement

Find u from y where $\mathscr{G} : X \to Y$, η is noise and

$$y = \mathscr{G}(u) + \eta.$$

- Problem can contain many degrees of uncertainty related to η and $\mathscr{G}$. Solution to problem should hence also contain uncertainty.
- Quantifying prior beliefs about the state u by a probability measure, Bayes' theorem tells us how to update this distribution given the data y, producing the posterior distribution.
- In the Bayesian approach, the solution to the problem is the posterior distribution.

The Posterior Distribution

- Assume, for simplicity, $Y = \mathbb{R}^J$ and the observational noise $\eta \sim N(0, \Gamma)$ is Gaussian. The likelihood of y given u is

$$\mathbb{P}(y|u) \propto \exp\left(-\frac{1}{2}\|y - \mathcal{G}(u)\|_\Gamma^2\right) =: \exp\left(-\Phi(u; y)\right).$$

- Quantify prior beliefs by a prior distribution $\mu_0 = \mathbb{P}(u)$ on X.
- Posterior distribution $\mu = \mathbb{P}(u|y)$ on X is given by Bayes' theorem:

$$\mu(\mathrm{d}u) = \frac{1}{Z} \exp\left(-\Phi(u; y)\right) \mu_0(\mathrm{d}u).$$

  See e.g. (Stuart, 2010; Sullivan, 2017).
- We first assume μ_0 is Gaussian, then move to more general priors.

Robust Sampling

- Expectations under μ can be estimated numerically using samples from μ. These may be generated though MCMC methods.
- Samples from MCMC chains are correlated, and strong correlations lead to poor expectation estimates. With straightforward MCMC methods, these correlations will become stronger as the dimension of state space N is increased.
- Preferable to have MCMC chains with convergence rate bounded independently of dimension, e.g. if μ_N is an approximation to μ,

  $$d(P_N^k(u_0, \cdot), \mu_N) \leq C(1 - \rho_N)^k$$

  with $\rho_N \geq \rho_* > 0$ for all N.
- General idea: if the chain is ergodic in infinite dimensions, the same should hold for finite-dimensional approximations.

Robust Sampling Gaussian Prior

- If the prior $\mu_0 = N(0, C)$ is Gaussian, posterior can be sampled with MCMC in a dimension robust manner using, for example, the pCN algorithm (Beskos et al., 2008; Hairer et al., 2014):

pCN Algorithm

1. Set $n = 0$ and choose $\beta \in (0, 1]$. Initial State $u^{(n)} \in X$.
2. Propose new state $\hat{u}^{(n)} = (1-\beta^2)^{\frac{1}{2}} u^{(n)} + \beta\zeta^{(n)}, \quad \zeta^{(n)} \sim N(0, C)$
3. Set $u^{(n+1)} = \hat{u}^{(n)}$ with probability

$$\alpha(u^{(n)} \to \hat{u}^{(n)}) = \min\left\{1, \exp\left(\Phi(u^{(n)}; y) - \Phi(\hat{u}^{(n)}; y)\right)\right\}$$

   or else set $u^{(n+1)} = u^{(n)}$.
4. Set $n \mapsto n + 1$ and go to 2.

- Dimension robust geometric methods, such as ∞-MALA, ∞-HMC, are also available, which take into account derivative information.

Robust Sampling Non-Gaussian Priors

- Non-Gaussian priors may be preferred over Gaussians to allow for, for example,
  - Piecewise constant fields/sparsity,
  - Heavier tails,
  - Different scaling properties,
  - Complex non-stationary behavior.
- For non-Gaussian priors, dimension robust properties can be obtained by choosing a prior-reversible proposal kernel such that the resulting MCMC kernel has a uniformly positive spectral gap (Vollmer, 2015).
- We consider non-Gaussian random variables that may be written as non-linear transformations of Gaussians, allowing us to take advantage of known dimension robust methods for Gaussian priors.

Robust Sampling General Idea

- Suppose that we may write $\mu_0 = T^\sharp \nu_0$ for some probability measure ν_0 and transformation map $T : \Xi \to X$. Then $\xi \sim \nu_0$ implies that $T(\xi) \sim \mu_0$.
- Consider the two distributions

$$\mu(\mathrm{d}u) = \frac{1}{Z}\exp\big(-\Phi(u; y)\big)\mu_0(\mathrm{d}u),$$
$$\nu(\mathrm{d}u) = \frac{1}{Z}\exp\big(-\Phi(T(\xi); y)\big)\nu_0(\mathrm{d}\xi).$$

Then $\mu = T^\sharp \nu$.

- If we can sample ν in a dimension robust manner, we can therefore sample μ in a dimension robust manner.

Outline

1. Introduction
2. Series-Based Priors
3. Hierarchical Priors
4. Level Set Priors
5. Numerical Illustrations
6. Conclusions

Series-Based Priors

Let

- $\{\varphi_j\}_{j\geq 1}$ a deterministic X-valued sequence,
- $\{\rho_j\}_{j\geq 1}$ be a deterministic $\mathbb{R}$-valued sequence,
- $\{\zeta_j\}_{j\geq 1}$ an independent random $\mathbb{R}$-valued sequence,
- $m \in X$,

and define μ_0 to be the law of the random variable

$$u = m + \sum_{j=1}^{\infty} \rho_j \zeta_j \varphi_j.$$

Such priors have been considered in Bayesian inversion in cases where, for example, $\{\zeta_j\}$ is uniform (Schwab, Stuart, 2012), Besov (Lassas et al., 2009) and stable (Sullivan, 2017).

Series-Based Priors

We can find a Gaussian measure ν_0 and map T such that $\mu_0 = T^\sharp \nu_0$ as follows:

1. Write down a method for sampling ζ_j.
2. Using an inverse CDF type method, rewrite this method in terms of a (possibly multivariate) Gaussian random variable ξ_j, so that $\xi_j \overset{d}{=} \Lambda_j(\xi_j)$.
3. Define

$$T(\xi) = m + \sum_{j=1}^{\infty} \rho_j \Lambda_j(\xi_j)\varphi_j$$

and ν_0 the joint distribution of $\{\xi_j\}_{j\geq 1}$.

Series-Based Priors Examples

Example (Uniform)

- We want $\zeta_j \sim U(-1, 1)$.
- Define $\nu_0 = N(0,1)^\infty$ and $\Lambda_j(z) = 2\varphi(z) - 1$, where φ is the standard normal CDF.

Example (Besov)

- We want $\zeta_j \sim \pi_q$, where $\pi_q(z) \propto \exp\left(-\frac{1}{2}|z|^q\right)$.
- Define $\nu_0 = N(0,1)^\infty$ and

$$\Lambda_j(z) = 2^{1/q}\mathrm{sgn}(z)\left(\gamma_{1/q}^{-1}(\varphi(|z|) - 1)\right)^{1/q}$$

where $\gamma_{1/q}$ is the normalized lower incomplete gamma function.

Numerical Example I Dimensional Robustness

- We consider a 2D regression problem with a Besov prior.
- We compare the average acceptance rate of the whitened pCN algorithm with those of three different Random Walk Metropolis algorithms, as the dimension of the state space is increased.

RWM Algorithm

1. Set $n = 0$ and choose $\beta > 0$. Initial State $u^{(n)} \in X$.
2. Propose new state $\hat{u}^{(n)} = u^{(n)} + \beta\zeta^{(n)}, \quad \zeta^{(n)} \sim Q(\zeta)$
3. Set $u^{(n+1)} = \hat{u}^{(n)}$ with probability

$$\alpha(u^{(n)} \to \hat{u}^{(n)}) = \min\left\{1, \exp\left(\Phi(u^{(n)}; y) - \Phi(\hat{u}^{(n)}; y)\right) \frac{\mu_0(\hat{u}^{(n)})}{\mu_0(u^{(n)})}\right\}$$

   or else set $u^{(n+1)} = u^{(n)}$.
4. Set $n \mapsto n + 1$ and go to 2.

Numerical Example I Dimensional Robustness

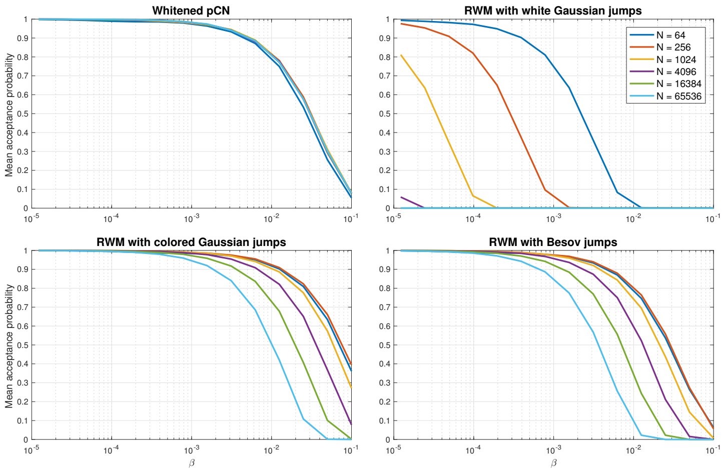


Figure: Mean acceptance rates for different MCMC algorithms. Curves are shown for different state space dimensions N.

Numerical Example II Existing Methods

- Vollmer (2015) provides methodology for robust sampling with a uniform prior, by reflecting random walk proposals at the boundary of $[-1, 1]^N$.
- We compare these methods, using both uniform and Gaussian reflected random walk proposals, with the reparametrized pCN method described above.
- As ν has a Gaussian prior, we have additional methodology available: we compare also with a dimension robust version of MALA, which makes likelihood informed proposals (Beskos et al, 2017).

Numerical Example II Existing Methods

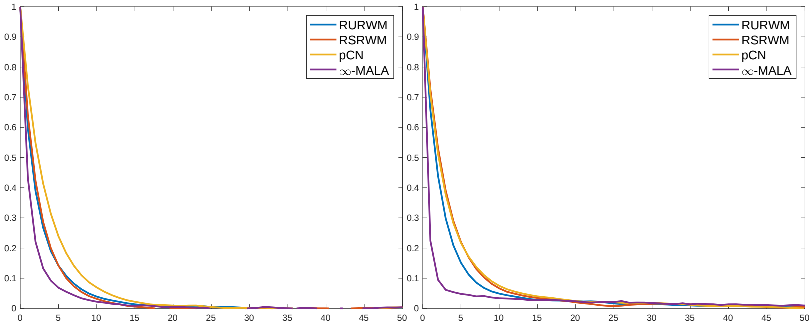


Figure: Autocorrelations for different sampling methods for a linear deblurring problem, with 8/32 observations (left/right).

Outline

1. Introduction
2. Series-Based Priors
3. Hierarchical Priors
4. Level Set Priors
5. Numerical Illustrations
6. Conclusions

Hierarchical Priors

- A family of Gaussian distributions will often have a number of parameters associated with it controlling sample properties.
- For example, Whittle-Matérn distributions with parameters $\theta = (\nu, \tau, \sigma)$, which have covariance function

$$c_\theta(x, y) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)}(\tau|x-y|)^\nu K_\nu(\tau|x-y|).$$

  - ν controls smoothness – draws from Gaussian fields with this covariance have ν fractional Sobolev and Hölder derivatives.
  - τ is an inverse length-scale.
  - σ is an amplitude scale.
- These parameters may not be known a priori, and may be treated hierarchically as hyperparameters in the problem.

Hierarchical Priors

- Placing a hyperprior π_0 on the hyperparameters θ, the posterior $\mu(\mathrm{d}u, \mathrm{d}\theta)$ may be sampled using a Metropolis-within-Gibbs method, alternating the following two steps:

Metropolis-within-Gibbs

1. Update $u^{(n)} \longmapsto u^{(n+1)}$ using pCN (etc) method for the conditional distribution $u|(\theta^{(n)}, y)$.
2. Update $\theta^{(n)} \longmapsto \theta^{(n+1)}$ using an MCMC method that samples $\theta|(u^{(n+1)}, y)$ in stationarity.

- Problem: For different choices of θ, the Gaussian distributions $u|\theta$ are typically singular. Moves in the hyperparameters are hence never accepted in the infinite-dimensional limit.

Hierarchical Priors

- We use the same methodology for robust sampling as for the series-based priors, finding an appropriate measure ν_0 and map T.
- Suppose $u|\theta \sim N(0, C(\theta))$. If $\xi \sim N(0, I)$ is white noise, then

$$C(\theta)^{1/2}\xi \sim N(0, C(\theta)).$$

  We hence define $\nu_0 = N(0, I) \otimes \pi_0$ and $T(\xi, \theta) = C(\theta)^{1/2}\xi$.
- ξ and θ are independent under ν_0, and so singularity issues from Metropolis-within-Gibbs disappear when sampling ν.
- Referred to as *non-centered* parameterization in literature (Papaspiliopoulos et al., 2007)

Hierarchical Priors Centered vs Non-Centered

- Current state θ, proposed state $\theta' = \theta + \varepsilon\eta$.
- Centered acceptance rate for $\theta|u, y$:

$$1 \wedge \sqrt{\frac{\det C(\theta)}{\det C(\theta')}} \exp\left(\frac{1}{2}\|C(\theta)^{-1/2}u\|^2 - \frac{1}{2}\|C(\theta')^{-1/2}u\|^2\right)\frac{\pi_0(\theta')}{\pi_0(\theta)}.$$

- Non-Centered acceptance rate for $\theta|\xi, y$:

$$1 \wedge \exp\left(\Phi(T(\xi,\theta);y) - \Phi(T(\xi,\theta');y)\right)\frac{\pi_0(\theta')}{\pi_0(\theta)}.$$

Hierarchical Priors

Example (Whittle-Matérn Distributions)

- If u is a Gaussian random field on $\mathbb{R}^d$ with covariance function $c_\theta(\cdot,\cdot)$ above, then it is equal in law to the solution of the SPDE

$$(\tau^2 I - \triangle)^{\nu/2+d/4} u = \beta(\nu)\sigma\tau^\nu \xi$$

  where $\xi \sim N(0, I)$. (Lindgren *et al.*, 2011)

- The unknown field in the inverse problem can thus either be treated as u or ξ, related via the mapping

$$u = T(\xi, \theta) = \beta(\nu)\sigma\tau^\nu (\tau^2 I - \triangle)^{-\nu/2-d/4} \xi.$$

- We set $\nu_0 = N(0, I) \otimes \pi_0$ and define T as above.

Outline

1. Introduction
2. Series-Based Priors
3. Hierarchical Priors
4. Level Set Priors
5. Numerical Illustrations
6. Conclusions

Level Set Method

- It is often of interest to recover a piecewise constant function, such as in inverse interface or classification problems.
- We can construct a piecewise constant function by thresholding a continuous function at a number of levels.
- Define $F : C^0(D) \to L^\infty(D)$ by

$$F(\xi)(x) = \begin{cases} \kappa_1 & \xi(x) \in (-\infty, -1) \\ \kappa_2 & \xi(x) \in [-1, 1) \\ \kappa_3 & \xi(x) \in [1, \infty). \end{cases}$$

  Then if $\xi \sim N(0, C)$ is continuous Gaussian random field, $u = F(\xi)$ is a piecewise constant random field.
- We have a natural method for sampling when a prior $\mu_0 = F^\sharp N(0, C)$ is used.

Level Set Method

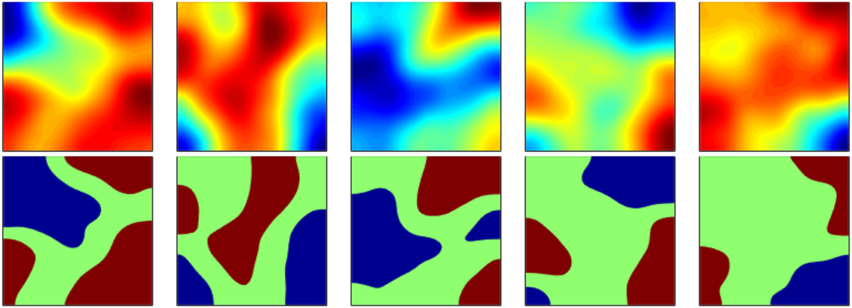


Figure: Examples of continuous fields $\xi \sim N(0, C)$ (top) and the thresholded fields $u = F(\xi)$ (bottom).

Level Set Method

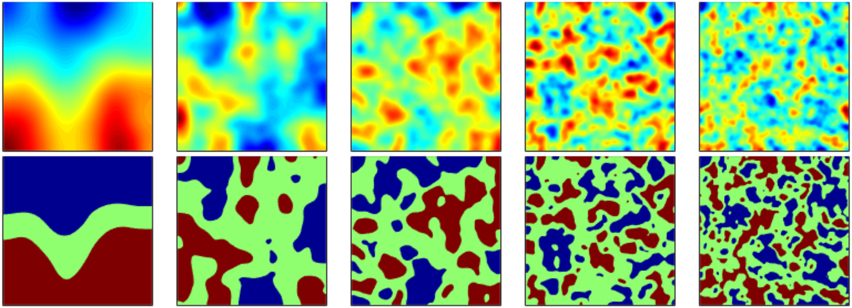


Figure: Examples of continuous fields $\xi \sim N(0, C(\tau))$ for various τ (top) and the thresholded fields $u = F(\xi)$ (bottom).

- The thresholding function $F : C^0(D;\mathbb{R}) \to L^\infty(D;\mathbb{R})$ has the disadvantage that an ordering of layers/classes is asserted. This can be overcome by instead using a vector level set method.
- Let k denote the number of layers/classes. Define $S : C^0(D;\mathbb{R}^k) \to L^\infty(D;\mathbb{R}^k)$ by

$$S(\xi)(x) = e_{r(x;\xi)}, \quad r(x;\xi) = \underset{j=1,\dots,k}{\operatorname{argmax}}\, \xi_j(x), \tag{1}$$

  where $\{e_r\}_{r=1}^k$ denotes the standard basis for $\mathbb{R}^k$.
- Samples from $\mu_0 = S^\sharp N(0,C)^k$ then take values in the set $\{e_r\}_{r=1}^k$ for each $x \in D$.

Outline

1. Introduction
2. Series-Based Priors
3. Hierarchical Priors
4. Level Set Priors
5. Numerical Illustrations
6. Conclusions

Numerical Example III Groundwater Flow

- We consider the case where the forward map represents the mapping from permeability to pressure in the steady state Darcy flow model.
- Let $D = (0,1)^2$. Given $u \in L^\infty(D)$, denote $p(x) = p(x; u)$ the solution to

$$\begin{cases} -\nabla \cdot (u \nabla p) = f & x \in D \\ p = 0 & x \in \partial D \end{cases}$$

- Define $\mathscr{G} : L^\infty(D) \to \mathbb{R}^K$ by $\mathscr{G}(u) = (p(x_1; u), \dots, p(x_J; u))$.
- We look at the Bayesian inverse problem of recovering u from a noisy measurement of $\mathscr{G}(u)$, using non-hierarchical and hierarchical level set priors.

Numerical Example III Groundwater Flow

- We consider the case where the forward map represents the mapping from permeability to pressure in the steady state Darcy flow model, observed at 36 points.

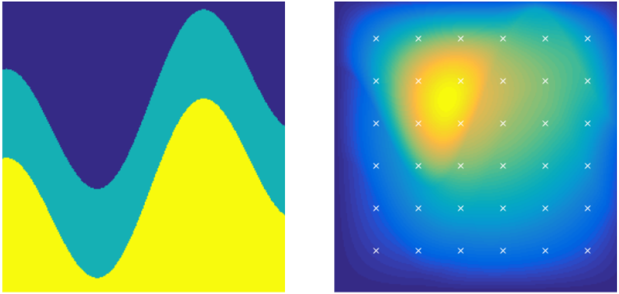


Figure: (Left) True permeability $u^\dagger$, (Right) True pressure and observed data y.

Numerical Example III Groundwater Flow

- We consider the case where the forward map represents the mapping from permeability to pressure in the steady state Darcy flow model, observed at 36 points.

Figure: (Left) True permeability $u^{\dagger}$, (Right) Posterior mean $F(\mathbb{E}(\xi))$.

Numerical Example III Groundwater Flow

- We consider the case where the forward map represents the mapping from permeability to pressure in the steady state Darcy flow model, observed at 36 points.

Figure: (Left) True permeability $u^\dagger$, (Right) Posterior mean $F\Big(T\big(\mathbb{E}(\xi), \mathbb{E}(\alpha), \mathbb{E}(\tau)\big)\Big)$.

Numerical Example IV Anisotropic Length Scale I

- Consider the infinite hierarchy $\{u_n\}_{n\in\mathbb{N}}$ of the form

$$u_{n+1} \,|\, u_n \sim N(0, C(u_n))$$

- We can alternatively write this in the form

$$u_{n+1} = L(u_n)\xi_n, \quad \xi_n \sim N(0, I) \text{ i.i.d}$$

  where $L(u)L(u)^* = C(u)$.

- We choose $L(u)$ such that $C(u)$ is a Whittle-Matern-type covariance with spatially varying inverse length-scale $\tau(u)$.
- Ergodicity of this process means that it is sufficient for inference to terminate the hierarchy after a small number of levels.

Numerical Example IV Anisotropic Length Scale I

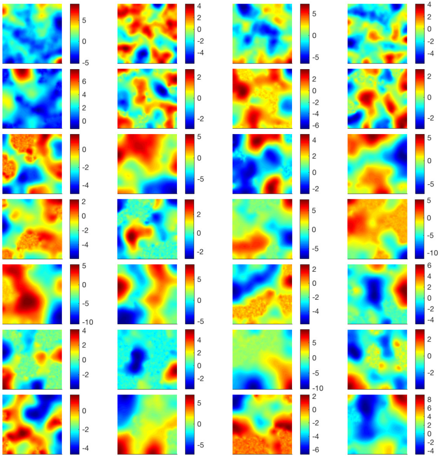
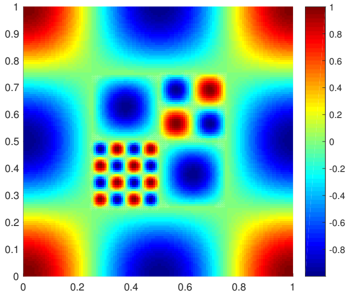


Figure: Four independent samples from the anisotropic Gaussian hierarchy prior. The first seven levels shown for each chain.

Numerical Example IV Anisotropic Length Scale I



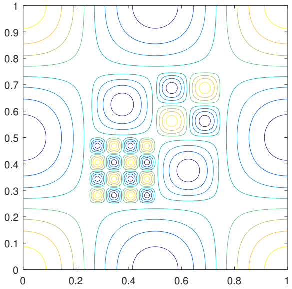


Figure: The true field $u^{\dagger}$, which posesses multiple length-scales. It is observed on a grid of J points.

Numerical Example IV Anisotropic Length Scale I

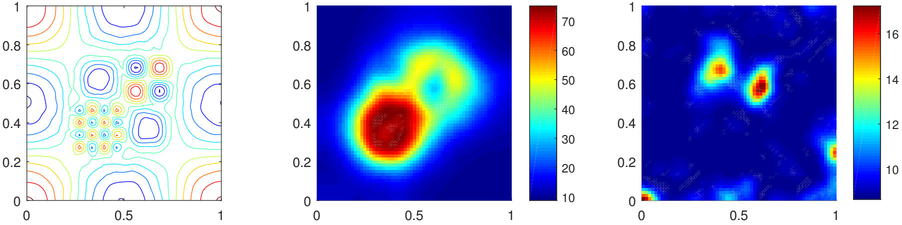


Figure: (Left) $\mathbb{E}(u_3)$, (Middle, Right) The length-scale fields $\mathbb{E}\big(F(u_2)\big)$, $\mathbb{E}\big(F(u_1)\big)$. Here $J = 2^{10}$.

Numerical Example IV Anisotropic Length Scale I

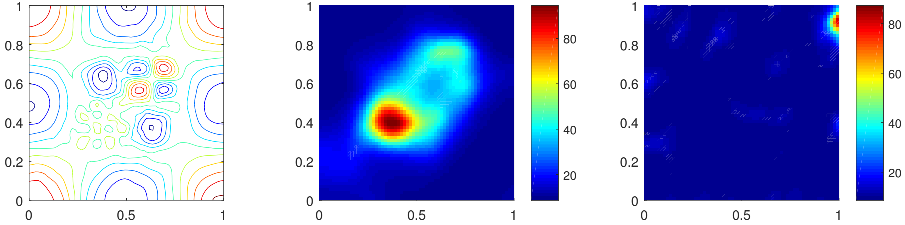


Figure: (Left) $\mathbb{E}(u_3)$, (Middle, Right) The length-scale fields $\mathbb{E}\big(F(u_2)\big)$, $\mathbb{E}\big(F(u_1)\big)$. Here $J = 2^8$.

Numerical Example V Anisotropic Length Scale II

- We now consider the case where the forward map is the mapping from conductivity to boundary voltage measuremeants in the complete electrode model for EIT (Somersalo et al, 1992).
- This is similar to the groundwater flow problem, except boundary measurements are made instead of interior measurements, for a collection of source terms.
- The prior is taken to be a 2 layer hierarchy of the form considered above, except each continuous field is thresholded to be piecewise constant.

Numerical Example V Anisotropic Length Scale II

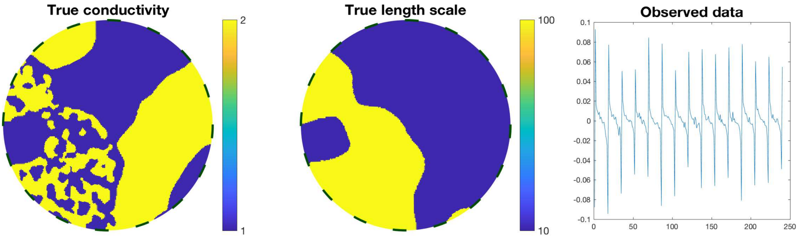


Figure: (Left) True conductivity field (Middle) True length-scale field (Right) Vector of observed boundary voltages.

Numerical Example V Anisotropic Length Scale II

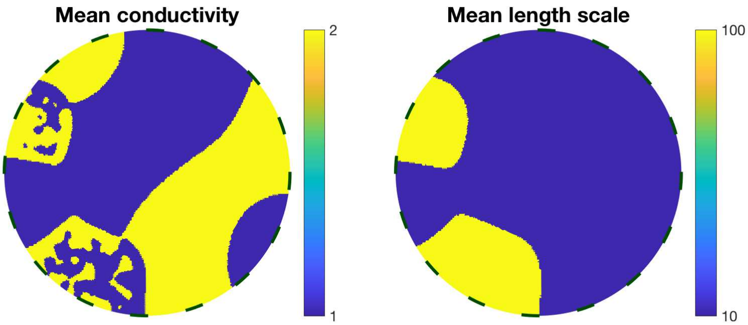


Figure: (Left) Pushforward of posterior mean conductivity (Right) pushforward of posterior mean length scale.

Numerical Example V Anisotropic Length Scale II

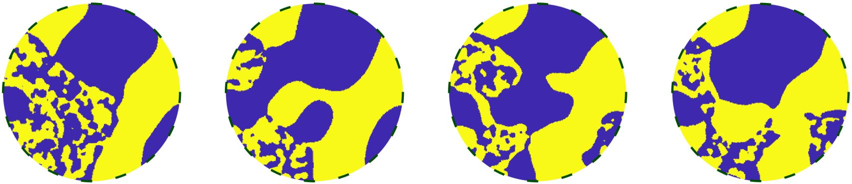


Figure: Examples of samples of the conductivity field under the posterior.

Numerical Example VI Graph-Based Learning

- Assume we have a collection of data points $\{x_j\}_{j\in Z} \subset \mathbb{R}^d$ that we wish to divide into classes.
- We have class labels $\{y_j\}_{j\in Z'}$ for a subset of points $\{x_j\}_{j\in Z'}$, $Z' \subset Z$, and wish to propagate these to all points.
- A priori clustering information can be obtained from the spectral properties of a graph Laplacian $L_N \in \mathbb{R}^{N\times N}$ associated with the data:

$$L_N = D - W, \quad W_{ij} = \eta(|x_i - x_j|), \quad D_{ii} = \sum_{j=1}^{N} W_{ij}$$

where η is a similarity function.

- If there are K clusters and η can distinguish between them perfectly, the first K eigenvectors of L_N can determine these clusters perfectly.

Numerical Example VI Graph-Based Learning

- We define the prior $\nu_0 = N(0, C(\theta))^K$ with $C(\theta) = P_M(L + \tau^2)^{-\alpha} P_M^*$, $\theta = (\alpha, \tau, M)$.
- A model for the labels is given,

$$y_j = S(u(x_j)) + \eta_j, \quad j \in Z',$$

where $S : \mathbb{R}^K \to \mathbb{R}^K$ is a multiclass thresholding function. This defines the likelihood, and hence posterior distribution.

- This problem is high- but finite-dimensional: it is not directly an approximation of a problem on Hilbert space. However, with appropriate assumptions and scaling, it does have a continuum limit as $N \to \infty$. (Garcia Trillos, Slepčev, 2014)

Numerical Example VI Graph-Based Learning

- We consider the case where the data points are elements of the MNIST dataset: 28×28 pixel images of handwritten digits $0, 1, \ldots, 9$.
- We are hierarchical about subsets of the parameters (α, τ, M), and see how classification accuracy is affected by this choice.
- A sample $u^{(k)} \sim N(0, C(\theta))$ can be expressed in the eigenbasis $\{\lambda_j, q_j\}$ of L_N as

$$u^{(k)} = \sum_{j=1}^{M} (\lambda_j + \tau^2)^{-\alpha/2} \xi_j q_j, \quad \xi_j \sim N(0, 1) \text{ i.i.d.}.$$

- We restrict to the digits 3, 4, 5, 9 here to speed up computations, and take ≈ 2000 of each digit so that $|Z| \approx 8000$. We provide labels for a random subset of $|Z'| = 80$ digits.

Numerical Example VI Graph-Based Learning

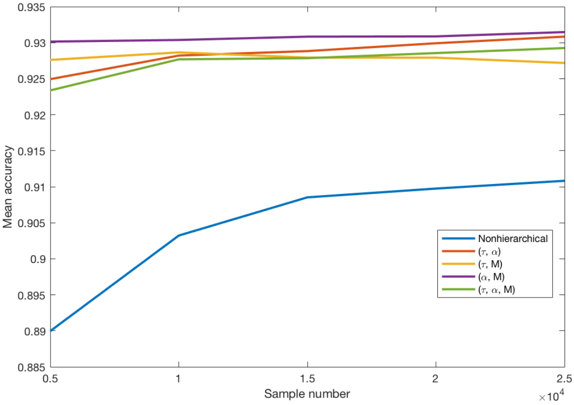


Figure: Mean classification accuracy of 40 trials against sample number, for MNIST digits 3,4,5,9.

- We now focus on a test set of 10000 digits, with approximately 1000 of each digit 0, 1, . . . , 9, being hierarchical about (α, M).
- From samples we can look at uncertainty in addition to accuracy.
- We introduce the measure of uncertainty associated with a data point x_j as

$$U(x_j) = \sqrt{\frac{k}{2k-1}} \min_{r=1,\dots,k} \left\| \mathbb{E}(Su)(x_j) - e_r \right\|_2$$

Remark: $(Su)(x)$ lies in the set $\{e_r\}_{r=1}^{k}$ for each fixed u, x, but the mean $\mathbb{E}(Su)(x)$ in general will not.

Numerical Example VI Graph-Based Learning

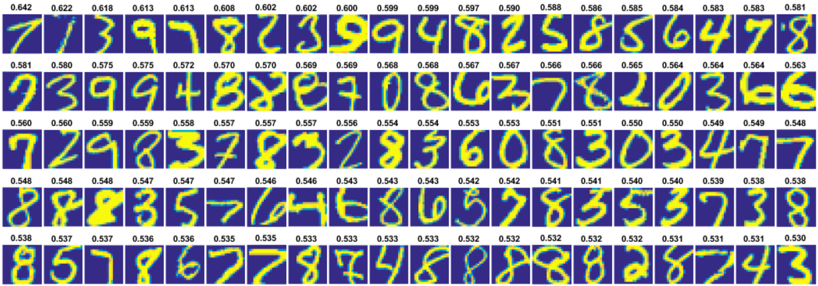


Figure: 100 most uncertain digits, 200 labels.

Mean uncertainty: 10%.

Numerical Example VI Graph-Based Learning

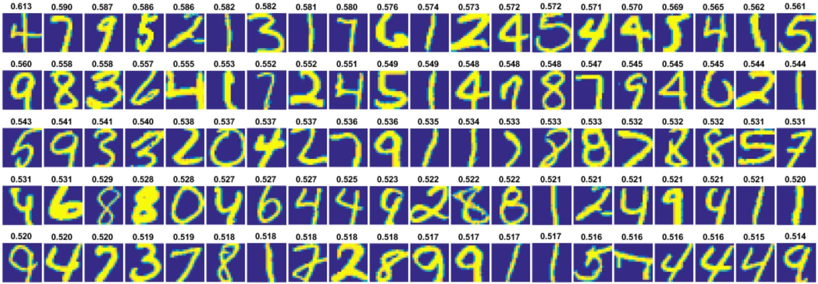


Figure: 100 most uncertain digits, 200+100 labels.

Mean uncertainty: 6.7%.

Numerical Example VI Graph-Based Learning

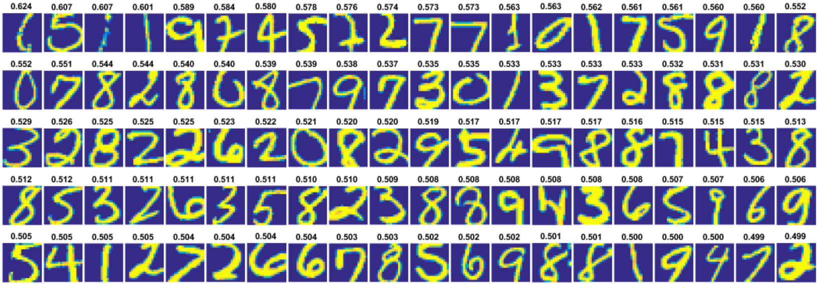


Figure: 100 most uncertain digits, 200+200 labels.

Mean uncertainty: 4.9%.

Outline

1. Introduction
2. Series-Based Priors
3. Hierarchical Priors
4. Level Set Priors
5. Numerical Illustrations
6. Conclusions

Conclusions

- Non-Gaussian priors can be more appropriate to use than Gaussians in certain contexts, and can provide extra flexibility via use of hyperparameters.
- By writing a non-Gaussian prior as the pushforward of a Gaussian under some non-linear map, existing dimension robust MCMC sampling methods for Gaussian priors can be used.
- For hierarchical priors, the transformation can also chosen such that the priors on the field and hyperparameters are independent (non-centering), avoiding issues associated with measure singularity in the Gibbs sampler.

References

[1] A. M. Stuart. *Inverse problems: a Bayesian perspective*. Acta Numerica, 2010.

[2] M. M. Dunlop, M. Girolami, A. M. Stuart and A. L. Teckentrup. *How deep is a deep Gaussian process?* In Preparation.

[3] V. Chen, M. M. Dunlop, O. Papaspiliopoulos and A. M. Stuart. *Robust MCMC Sampling in High Dimensions with Non-Gaussian and Hierarchical Priors*. In Preparation.

[4] M. M. Dunlop, M. A. Iglesias and A. M Stuart. *Hierarchical Bayesian level set inversion*. Statistics and Computing, 2016.

[5] Z. Wang, J. M. Bardsley, A. Solonen, T. Cui and Y. M. Marzouk. *Bayesian inverse problems with ℓ_1 priors: a randomize-then-optimize approach*. SIAM Journal on Scientific Computing, 2017.