Hierarchical Methods for Bayesian Inverse Problems

Optimization and Inversion under Uncertainty, RICAM, November 12th, 2019

Matt Dunlop (Courant)
Tapio Helin (Lappeenranta)
Andrew Stuart (Caltech)

Outline

Outline

The Inverse Problem

Problem Statement

Find u from y where $A : X \to Y$, η is noise and

$$y = Au + \eta.$$

- Problem can contain many degrees of uncertainty related to η and A. 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 Definition

- 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

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

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

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

  See e.g. (Stuart, 2010; Sullivan, 2017).
- We will generally assume throughout that $\mu_0(\cdot|\theta)$ is Gaussian.

The Posterior Distribution Hierarchical Definition

- A family of Gaussian distributions $\{\mu_0(\cdot|\theta)\}_{\theta\in\Theta}$ will often have a number of parameters associated with it controlling sample properties.
- These parameters may not be known a priori, and may be treated hierarchically as hyperparameters in the problem.
- We now have a prior $\mathbb{P}(u, \theta)$ on the pair (u, θ), which we assume factors as

$$\mu_0(\mathrm{d}u, \mathrm{d}\theta) = \mu_0(\mathrm{d}u|\theta)\rho_0(\mathrm{d}\theta).$$

- Posterior distribution $\mu = \mathbb{P}(u, \theta|y)$ on $X \times \Theta$ is given by Bayes' theorem:

$$\mu(\mathrm{d}u, \mathrm{d}\theta) = \frac{1}{Z}\exp\left(-\Phi(u; y)\right)\mu_0(\mathrm{d}u|\theta)\rho_0(\mathrm{d}\theta).$$

The Posterior Distribution Inference

- What do we do with a probability distribution for a solution?
  - Obtain point estimates of unknown u, for example the mode or mean.
  - Find point estimates for quantities of interest $g(u)$ of the unknown.
  - Find uncertainty estimates, credible sets, etc for the above quantities.
- MAP Estimation: The mode can often be calculated quickly using optimization techniques, although this does not provide uncertainty information.
- Sampling: The other estimates typically require (approximate) samples from from the posterior to estimate expectations.

Approaches to the above vary based on whether we discretize the state space before applying Bayesian methodology, or work directly on function space before discretizing.

Outline

Outline

Hierarchical Priors Whittle Matérn

- Consider Whittle-Matérn distributions with parameters $\theta = (\sigma, \nu, \sigma) \in \mathbb{R}^3_+$, which have covariance function

$$c(x, x'|\theta) = \sigma^2 \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{|x - x'|}{\ell} \right)^{\nu} K_\nu \left(\frac{|x - x'|}{\ell} \right).$$

  - σ is an amplitude scale.
  - ν controls smoothness.
  - ℓ is a length-scale.

- A sample $v \sim \mathrm{GP}(0, c(x, x'|\theta))$ on $\mathbb{R}^d$ can be characterized as the solution to the SPDE (Lindgren et al, 2011)

$$(\ell^{-2} I - \Delta)^{\nu/2 + d/4} v = \sigma \beta(\nu) \ell^{-\nu} \xi.$$

- Choosing $\mu_0(\cdot|\theta)$ to be the law of the solution to this SPDE, and ρ_0 any measure supported on a subset of $\mathbb{R}^3_+$, gives an example of a hierarchical Gaussian prior.
- Note that the SPDE makes sense even when ℓ is not scalar – we could allow the length scale to vary in space, in which case the solution will formally have length-scale $\ell(x)$ at each point x.
- Choose instead ρ_0 to be supported on a subset of $R^2_+ \times C^0$. (Roininen et al, 2019)
- Alternatively iterate to have a multi-layer hierarchy (deep Gaussian process) for additional flexibility, so ρ_0 is supported on a subset of $\mathbb{R}^2_+ \times (C^0)^L$. (Dunlop et al, 2018)

Hierarchical Priors Sparsity promoting

Other examples may be defined on a set of discrete points in $\mathbb{R}^d$ rather than a continuum domain. For example, let $u = \{u_1, \ldots, u_N\} \subseteq \mathbb{R}$.

- Let $C(\theta) = \mathrm{diag}(\sqrt{\theta_j}) \in \mathbb{R}^N$, $\mu_0(\cdot|\theta) = N(0, C(\theta))$, and let ρ_0 be supported on $\mathbb{R}^N_+$.
- Choosing ρ_0 such that $\theta_j \sim \mathrm{Gamma}(\theta_j^*, \beta)$ i.i.d. leads to sparsity in MAP estimates, and can approximate ℓ^1 regularization for small enough β.
- Introduced by (Calvetti, Somersalo, 2007) and since analyzed in numerous papers.

Note that in continuum setting the covariance operator $C(\theta)$ is a multiplication operator, and so not compact on L^2. Samples from this distribution will hence not belong to L^2 almost-surely – instead, they will be distribution-valued.

Hierarchical Priors Other examples

- Given a set of input points $\mathbf{x} = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$, could choose to look at their empirical covariance ($\rightarrow$ PCA-basis) or a graph Laplacian based on them ($\rightarrow$ diagonalize). Both give sequence of eigenvalues and eigenvectors $\{\lambda_j, \varphi_j\}_{j=1}^N$.
  - Consider the prior $\mu_0(\cdot|\theta)$ on $\{u : \mathbf{x} \rightarrow \mathbb{R}\} \cong \mathbb{R}^N$ given by the law of the sum (Bertozzi et al, 2018)

$$\sum_{j=1}^{M} f(\lambda_j; \theta')\xi\varphi_j, \quad \xi_j \sim N(0,1), \quad \theta = (M, \theta').$$

- Alternatively, could consider general Gibbs-type priors (Zhou et al, 1997)

$$\mu_0(\mathrm{d}u|\theta) = \frac{1}{Z(\theta)} \exp(-\theta V(u))\,\mathrm{d}u, \quad Z(\theta) = \int \exp(-\theta V(u))\,\mathrm{d}u$$

Outline

Inference Gibbs Sampling

- We typically can't sample $\mathbb{P}(u, \theta|y)$ directly, however depending on model and choice of prior, we may be able to sample $\mathbb{P}(u|\theta, y)$ and $\mathbb{P}(\theta|u, y)$ directly (conjugate priors).
- Example: with a linear Gaussian data model, $\mu_0(\cdot|\theta) = N(0, \theta^{-1}C_0)$ and $\rho_0(\cdot)$ a gamma distribution, we have (Bernardo, Smith, 2009)

$$\mathbb{P}(u|\theta, y) = N(m(\theta, y), C(\theta, y)), \quad \mathbb{P}(\theta|u, y) = \text{Gamma}(\alpha(u), \beta(u))$$

- The Gibbs sampler forms a Markov chain $\{u^{(k)}, \theta^{(k)}\}$ by alternating the steps
  1. $u^{(k+1)} \sim \mathbb{P}(u|\theta^{(k)}, y)$
  2. $\theta^{(k+1)} \sim \mathbb{P}(\theta|u^{(k+1)}, y)$.

Inference Metropolis-within-Gibbs Sampling

- Depending on the model and choice of prior, we may be unable to sample either or both of the distributions $\mathbb{P}(u|\theta, y)$ and $\mathbb{P}(\theta|u, y)$ directly.
- When this is the case, we can use the Metropolis-within-Gibbs algorithm to sample, alternating the steps
  1. Update $u^{(k)} \mapsto u^{(k+1)}$ with MCMC method targeting $\mathbb{P}(u|\theta^{(k)}, y)$
  2. Update $\theta^{(k)} \mapsto \theta^{(k+1)}$ with MCMC method targeting $\mathbb{P}(\theta|u^{(k+1)}, y)$.
- Samples are typically more correlated than using direct Gibbs sampler.
- Since u is often high-dimensional, it can be useful to use a dimension-robust MCMC sampler for the u-update, such as pCN, ∞-(m)MALA, ∞-(m)HMC, etc; see (Beskos et al, 2017) for a review.
- The high-dimensionality of u can also cause problems with the θ-update due to measure singularity – this will be discussed later.

Inference Empirical Bayes

- Instead of maintaining uncertainty estimates on the state u and hyperparameters θ, it may be more computationally feasible to marginalize out one of them, and optimize the resulting density.
- Given an initial estimate for $\theta^{(0)}$, we may choose to alternate the steps
  1. (Expectation) Estimate $J^{(k)}(\theta; y) := \mathbb{E}^{u|\theta^{(k)},y}(\mu_0(u|\theta)\rho_0(\theta)/\mu_0(u|\theta^{(k)}))$ using samples from $\mathbb{P}(u|\theta^{(k)}, y)$.
  2. (Maximization) Find $\theta^{(k+1)} \in \operatorname*{argmax}_\theta J^{(k)}(\theta; y)$.
- The samples for the expectation step may be generated using a robust MCMC method such as those from previous slide.
- In the following section, we show that this approach leads to consistent estimators for θ in the limit of good data, under certain assumptions.

Inference MAP Estimation

- The empirical Bayesian method maximizes the marginal density on $u|\theta$, and so may be viewed as a compromise between the mean and mode of the joint posterior.
- Instead, one may choose to optimize over both the state and hyperparameter to provide the MAP estimate for the pair (u, θ); in practice this is much cheaper to compute than the above methods.
- It has direct relations to classical variational methods for inversion.
- It depends on the parameterization of the model and prior; in the following section we show that the estimators for θ may either be consistent or inconsistent based on the choice of parameterization.

Outline

MAP Estimation

- Assume $\dim(X) = N < \infty$, so we may view $u \in X$ as $u \in \mathbb{R}^N$.
- Define the prior $\mu_0(\cdot|\theta) = N(0, C(\theta))$ for some covariance matrix $C(\theta) \in \mathbb{R}^{N \times N}$.
- $\mu(\mathrm{d}u) = \pi(u)\mathrm{d}u$ has a Lebesgue density:

$$\pi(u) \propto \exp\left(-\Phi(u; y) - \frac{1}{2}\langle u, C(\theta)^{-1}u\rangle\right).$$

- u_* is a mode (MAP estimator) for μ if and only if

$$u_* \in \operatorname*{argmin}_{u \in \mathbb{R}^N} \Phi(u; y) + \frac{1}{2}\langle u, C(\theta)^{-1}u\rangle.$$

- Definitions of MAP estimators are available when $\dim(X) = \infty$ so that no Lebesgue density exists, see e.g. (Dashti et al, 2013), (Helin, Burger, 2015), (Lie, Sullivan, 2017).

MAP Estimation Linear Gaussian Setting

- When the forward map $A : X \to Y$ is linear, and the observational noise η and prior distribution $\mu_0(\cdot|\theta)$ are Gaussian, this optimization problem may be solved analytically:

$$u_* = C(\theta)A^*(\Gamma + AC(\theta)A^*)^{-1}y$$

- In fact this is also the posterior mean, since the posterior remains Gaussian in this setting.
- Our prior will only be conditionally Gaussian, so our posterior will not be Gaussian. However, this expression will be very useful for analysis of hierarchical MAP estimates later.

MAP Estimation Parameterization Dependence

- Key issue with MAP estimation: estimates depend on the choice of parameterization/coordinate system.
- Suppose that we have a smooth bijective map $T : X \to X$, and write $u = T(\xi)$ for some new coordinates ξ. Then the posterior density in terms of ξ is given by

$$\bar{\pi}(\xi) = \pi(T(\xi)) \times |\det(\nabla T(\xi))|.$$

  That is, for any bounded measurable $f : X \to \mathbb{R}$, we have

$$\int_X f(u)\pi(u)\,\mathrm{d}u = \int_X f(T(\xi))\bar{\pi}(\xi)\,\mathrm{d}\xi.$$

- This Jacobian determinant can completely change the location of modes (unless T is linear): in general, $u_* \neq T(\xi_*)$.

Hierarchical Priors

- We now have a prior $\mathbb{P}(u, \theta)$ on the pair (u, θ), which we assume factors as

$$\mu_0(\mathrm{d}u, \mathrm{d}\theta) = \mu_0(\mathrm{d}u|\theta)\rho_0(\mathrm{d}\theta)$$

- In the same setup as above, the posterior density $\pi(u, \theta)$ is then given by

$$\pi(u, \theta) \propto \exp\left(-\Phi(u; y) - \frac{1}{2}\langle u, C(\theta)^{-1}u\rangle - \frac{1}{2}\log\det C(\theta) + \log\rho_0(\theta)\right).$$

- Note the appearance of the log-determinant term, since the normalization constant of the conditionally Gaussian prior depends on θ.

Hierarchical MAP Estimation

- There may be no 'natural' choice of hyperparameter parameterization – for example should we work with length scale or inverse length scale when using Matérn priors? This will affect MAP estimates.
- More generally, we can choose a different parameterization for the pair (u, θ), so that we aren't necessarily directly trying to infer the field u of interest.
- The hierarchical posterior still makes sense when $\dim(X) = \infty$; we just cannot write down its Lebesgue density.
- However, in infinite dimensions the choice of parameterization can affect whether MAP estimation is even well-defined as an optimization problem!
- For example, what happens to the log-determinant term above when $\dim(X) \to \infty$, since $C(\theta)$ is a compact operator in infinite dimensions?

Hierarchical MAP Estimation

- Even though the optimization problem may not be well-defined in infinite dimensions, we can look at the behaviour of optimizers of sequences of finite-dimensional approximations.

Does the choice of parameterization affect the consistency of estimates as the dimensions of the data and state spaces go to infinity?

Centred Parameterization

- In the above, we have worked with the natural *centred* hierarchical parameterization: the likelihood does not depend explicitly on the hyperparameter θ.
- Such a choice of parameterization can be bad for *sampling* in high dimensions, due to measure singularity issues. However in finite dimensions we can still write down and optimize the negative log posterior in order to perform MAP estimation.
- The objective we wish to minimize is given by

$$I_C(u, \theta) = \frac{1}{2}\|Au - y\|_\Gamma^2 + \frac{1}{2}\langle u, C(\theta)^{-1}u\rangle + \frac{1}{2}\log\det C(\theta) + \log\rho_0(\theta).$$

Centred Parameterization

- We choose first to optimize over u, and then optimize over θ.
- Fix θ, and denote $u(\theta)$ a minimizer of $I_C(\cdot, \theta)$.
- In the linear setting we work in, we know this minimizer is unique and is given by

$$u(\theta) = C(\theta)A^*(\Gamma + AC(\theta)A^*)^{-1}y.$$

- We now define our reduced objective functional

$$J_C(\theta) := I_C(u(\theta), \theta)$$

and note that if θ_* minimizes J_C, then $(u(\theta_*), \theta_*)$ minimizes I_C.

Non-Centred Parameterization

- We now consider a different parameterization that arises when considering dimension-robust MCMC algorithms (Papaspiliopoulos et al., 2007), (Chen et al. 2019).
- Suppose $u|\theta \sim N(0, C(\theta))$. If $\xi \sim N(0, I)$ is white, then $C(\theta)^{1/2}\xi \sim N(0, C(\theta))$.
- We could hence parameterize the prior/posterior in terms of (ξ, θ) instead of (u, θ) – note that ξ and θ are independent under the prior.
- Define the map $T(\xi, \theta) = (C(\theta)^{1/2}\xi, \theta)$, the measure $\nu_0 = N(0, I)$, and the measure

$$\nu(\mathrm{d}\xi, \mathrm{d}\theta) = \frac{1}{Z}\exp(-\Phi(T(\xi, \theta); y))\nu_0(\mathrm{d}\xi)\rho_0(\mathrm{d}\theta).$$

  Then $\mu = T^{\#}\nu$.

Non-Centred Parameterization

- MAP estimation for ν is well-defined in infinite dimensions, due to the independence of ν_0 and ρ_0.
- MAP estimates are given by minimizers of the Onsager-Machlup functional

$$I_{NC}(\xi, \theta) = \frac{1}{2}\|AC(\theta)^{1/2}\xi - y\|_\Gamma^2 + \frac{1}{2}\langle \xi, \xi\rangle + \log \rho_0(\theta).$$

- Again we proceed by optimizing over ξ first,

$$\xi(\theta) = C(\theta)^{1/2}A^*(\Gamma + AC(\theta)A^*)^{-1}y,$$

and defining

$$J_{NC}(\theta) = I_{NC}(\xi(\theta), \theta).$$

Empirical Bayes

- A final approach we consider, a compromise between finding the mean and the MAP, is the empirical Bayesian approach. Here, the state u is marginalized out to leave an optimization problem just for the hyperparameter.
- We rewrite the model in non-centred coordinates:

$$y = AC(\theta)^{1/2}\xi + \eta, \quad \eta \sim N(0, \Gamma),$$

  with prior $\xi \sim N(0, I)$. From this it can be seen that $\mathbb{P}(y|\theta) = N(0, \Gamma + AC(\theta)A^*)$.
- We can hence apply Bayes' theorem to see that $\mathbb{P}(\theta|y) \propto \exp(-J_E(\theta))$, where

$$J_E(\theta) = \frac{1}{2}\|y\|^2_{\Gamma+AC(\theta)A^*} + \frac{1}{2}\log\det(\Gamma + AC(\theta)A^*) - \log\rho_0(\theta).$$

Outline

Outline

Consistency of Estimates

Consistency of the posterior can refer to a number of different concepts, all notions of convergence as the quality of data increases.

Non-hierarchical:

- Consistency of MAP estimators - does u_{MAP} converge to the true state?
- Consistency of full posterior - does $\mathbb{P}(u|y)$ converge to a Dirac measure at the true state? In what sense and at what rate?

Hierarchical:

- Consistency of MAP estimators - do u_{MAP}, θ_{MAP} converge to the true state and hyperparameter?
- Consistency of full posterior - given a method for choosing θ, does $\mathbb{P}(u|\theta, y)$ converge to the a Dirac measure at the true state? In what sense and at what rate?

A Motivating Example

- Consider the stationary Ornstein-Uhlenbeck process $\{u_t\}$ on $(0, 1)$ with variance σ^2 and length scale $\ell = \sigma^2/\beta$:

$$\mathrm{d}u_t = -\beta/\sigma^2 u_t \,\mathrm{d}t + \sqrt{2\beta}\,\mathrm{d}W_t, \quad u_0 \sim N(0, \sigma^2).$$

  Given observations of $\{u_{t_i}\}$ of a path u_t at infinitely many time points $\{t_i\} \subset (0, 1)$, can we determine both σ and β?

- How do we use the observations of u_t to construct estimators for σ and ℓ?
- β can be recovered by approximating the quadratic variation of u_t at any time t. However, it is known that β and σ^2 cannot be *jointly* recovered (van Zanten, 2001).
- Important note: the law of $\{u_t\}$ is equivalent to that of $\{\sqrt{2\beta}W_t\}$ for any choice of σ^2, by Girsanov's theorem.

Outline

Model Assumptions

In order to analyse the behaviour of these minimizers, we work in the simplified setup where the forward map A is linear, and A^*A is simultaneously diagonalizable with the family of covariance operators:

Assumptions I

(i) The map A^*A and family of prior covariance operators $\{C(\theta)\}_{\theta\in\Theta}$ are strictly positive and simultaneously diagonalizable with orthonormal eigenbasis $\{\varphi_j\}$, and we have

$$A^*A\varphi_j = a_j^2\varphi_j, \quad C(\theta)\varphi_j = \mu_j(\theta)\varphi_j \quad \text{for all } j \in \mathbb{N}, \theta \in \Theta.$$

(ii) The noise covariance $\Gamma = \gamma^2 I$ is white.

Model Definition

- Let $\theta^\dagger \in \Theta$ denote the true hyperparameter. We assume that the true state $u^\dagger \sim N(0, C(\theta^\dagger))$. Given a noise level γ, we define the data $y^\gamma \in Y$ by

$$y^\gamma = Au^\dagger + \gamma\eta, \quad \eta \sim N(0, I).$$

- Choosing the orthonormal basis $\{\psi_j\}$ for Y, where $\psi_j = A\varphi_j/a_j$, we have

$$y_j^\gamma := \langle y^\gamma, \psi_j \rangle \overset{d}{=} a_j u_j^\dagger + \gamma\eta_j, \quad \eta_j \overset{i.i.d}{\sim} N(0,1), \quad j \in \mathbb{N}.$$

  where $u_j^\dagger := \langle u^\dagger, \varphi_j \rangle$.

- We consider the sequence of problems from taking the first N of these observations:

$$y_j^\gamma \overset{d}{=} a_j u_j^\dagger + \gamma\eta_j, \quad \eta_j \overset{i.i.d}{\sim} N(0,1), \quad j = 1, \ldots, N.$$

Finite-Dimensional Problems

- The negative log-likelihood of these first N observations given u takes the form

$$\Phi_\gamma(u; y^\gamma_{1:N}) = \frac{1}{2\gamma^2} \sum_{j=1}^{N} |a_j u_j - y^\gamma_j|^2$$

  where $u_j := \langle u, \varphi_j \rangle$.
- The posterior on u_j for $j > N$ is hence uninformed by the data, and so remains the same as the prior; we therefore take the conditional prior for the Nth problem to be the projection of $N(0, C(\theta))$ onto the span of the first N eigenfunctions $\{\varphi_j\}$.
- We denote by $J_C^{N,\gamma}$, $J_{NC}^{N,\gamma}$ and $J_E^{N,\gamma}$ the functionals J_C, J_{NC} and J_E respectively constructed for these finite dimensional problems.

Objective Functions

Proposition

Define $s_j^\gamma(\theta) = a_j^2 \mu_j(\theta) + \gamma^2$. Then we have

$$J_C^{N,\gamma}(\theta) \propto \frac{1}{2N} \sum_{j=1}^{N} \left[\frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \log \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} \right] - \frac{1}{N} \log \rho_0(\theta),$$

$$J_{NC}^{N,\gamma}(\theta) \propto \frac{1}{2N} \sum_{j=1}^{N} \frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \frac{1}{N} \log \rho_0(\theta),$$

$$J_E^{N,\gamma}(\theta) \propto \frac{1}{2N} \sum_{j=1}^{N} \left[\frac{(y_j^\gamma)^2}{s_j^\gamma(\theta)} - \log \frac{s_j^\gamma(\theta^\dagger)}{s_j^\gamma(\theta)} \right] - \frac{1}{N} \log \rho_0(\theta).$$

Model Assumptions

Assumptions II

(i) $\Theta \subseteq \mathbb{R}^k$ is compact.

Model Assumptions

Assumptions II

(i) $\Theta \subseteq \mathbb{R}^k$ is compact.

(ii) $g(\theta, \theta^\dagger) := \lim_{j \to \infty} \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}$ exists for all $\theta \in \Theta$, and the map $\theta \mapsto g(\theta, \theta^\dagger) - \log g(\theta, \theta^\dagger)$ is lower semicontinuous.

Model Assumptions

Assumptions II

(i) $\Theta \subseteq \mathbb{R}^k$ is compact.

(ii) $g(\theta, \theta^\dagger) := \lim_{j \to \infty} \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}$ exists for all $\theta \in \Theta$, and the map $\theta \mapsto g(\theta, \theta^\dagger) - \log g(\theta, \theta^\dagger)$ is lower semicontinuous.

(iii) If $g(\theta, \theta^\dagger) = 1$, then $\theta = \theta^\dagger$.

Model Assumptions

Assumptions II

(i) $\Theta \subseteq \mathbb{R}^k$ is compact.

(ii) $g(\theta, \theta^\dagger) := \lim_{j \to \infty} \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}$ exists for all $\theta \in \Theta$, and the map $\theta \mapsto g(\theta, \theta^\dagger) - \log g(\theta, \theta^\dagger)$ is lower semicontinuous.

(iii) If $g(\theta, \theta^\dagger) = 1$, then $\theta = \theta^\dagger$.

(iv) γ_N is chosen such that $\min_{j=1,\dots,N} a_j^2 \mu_j(\theta)/\gamma_N^2 \to \infty$ as $N \to \infty$ for all $\theta \in \Theta$.

Model Assumptions

Assumptions II

(i) $\Theta \subseteq \mathbb{R}^k$ is compact.

(ii) $g(\theta, \theta^\dagger) := \lim_{j\to\infty} \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)}$ exists for all $\theta \in \Theta$, and the map $\theta \mapsto g(\theta, \theta^\dagger) - \log g(\theta, \theta^\dagger)$ is lower semicontinuous.

(iii) If $g(\theta, \theta^\dagger) = 1$, then $\theta = \theta^\dagger$.

(iv) γ_N is chosen such that $\min_{j=1,\dots,N} a_j^2 \mu_j(\theta)/\gamma_N^2 \to \infty$ as $N \to \infty$ for all $\theta \in \Theta$.

(v) The maps $\theta \mapsto \log \mu_j(\theta)$ are Lipschitz on Θ for each $j \in \mathbb{N}$, with Lipschitz constants uniformly bounded in j.

(vi) The maps $\theta \mapsto s_j^{\gamma_N}(\theta^\dagger)/s_j^{\gamma_N}(\theta)$ are Lipschitz on Θ for each $j = 1, \dots, N$, $N \in \mathbb{N}$, with Lipschitz constants uniformly bounded in j, N.

(vii) The map $\theta \mapsto \log \rho_0(\theta)$ is Lipschitz on Θ.

Main Theorem

Theorem

Let Assumptions I, II hold, and let $\{\theta_{\mathrm{C}}^N\}$, $\{\theta_{\mathrm{E}}^N\}$, $\{\theta_{\mathrm{NC}}^N\}$ *denote sequences of minimizers over* Θ *of* $\{\mathrm{J}_{\mathrm{C}}^{N,\gamma_N}\}$, $\{\mathrm{J}_{\mathrm{E}}^{N,\gamma_N}\}$, $\{\mathrm{J}_{\mathrm{NC}}^{N,\gamma_N}\}$ *respectively.*

(i) $\theta_{\mathrm{C}}^N, \theta_{\mathrm{E}}^N \to \theta^\dagger$ *in probability as* $N \to \infty$.

(ii) *Assume further that* $g(\cdot, \theta^\dagger)$ *has a unique minimizer* θ_*. *Then* $\theta_{\mathrm{NC}}^N \to \theta_*$ *in probability as* $N \to \infty$.

- An important implication of this result is that the hyperparameters can only be determined up to measure equivalence: by the Feldman–Hájek theorem, the measures $N(0, C(\theta^\dagger))$ and $N(0, C(\theta))$ are equivalent if and only if

$$\sum_{j=1}^{\infty} \left(\frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} - 1 \right)^2 < \infty.$$

Main Theorem Remark

- (Knapik et al, 2016) also studied consistency of empirical Bayesian estimators for diagonal inverse problems, but confined to a single hyperparameter describing the regularity of the Gaussian prior.
- However, in their setting they obtain stronger results, showing convergence rates of the empirical estimator (in probability), and they use this to deduce that the empirical posterior on u contracts around the ground truth at an optimal rate.
- A key difference is that we assume data to be generated according to $\mathbb{P}(y|\theta^\dagger)$, whereas they consider it to be generated according to $P(y|u^\dagger)$.
- The regularity of $u^\dagger$ is a almost-sure property of the prior, allowing for the latter to be used in that case, unlike for distributional properties like length-scale.

Outline

Application Whittle-Matérn Distributions

- We consider the recovery of the hyperparameters $\theta = (\sigma, \ell)$ of a Matérn field, with fixed regularity ν.
- Using the SPDE representation of Matérn fields (Lindgren et al, 2011), on a domain $D \subseteq \mathbb{R}^d$ we may write down the eigenvalues of the corresponding covariance operator:

$$\mu_j(\theta) = \kappa(\nu)\sigma^2\ell^{-2\nu}(\ell^{-2} + \lambda_j)^{-\nu-d/2}$$

  for some constant $\kappa(\nu)$, where $\{\lambda_j\}$ are the eigenvalues of the Laplacian on D.
- We may then calculate

$$g(\theta, \theta^\dagger) = \lim_{j\to\infty} \frac{\mu_j(\theta^\dagger)}{\mu_j(\theta)} = \left(\frac{\sigma^\dagger}{\sigma}\right)^2 \left(\frac{\ell}{\ell^\dagger}\right)^{2\nu}$$

  which equals 1 whenever $\sigma\ell^{-\nu} = \sigma^\dagger(\ell^\dagger)^{-\nu}$.

Application Whittle-Matérn Distributions

- To apply the theorem we need that $g(\cdot, \theta^\dagger)$ has a unique minimizer. We hence rewrite in terms of $\beta = \sigma \ell^{-\nu}$ and try to infer β.
- This gives

$$g(\theta, \theta^\dagger) = \left(\frac{\beta^\dagger}{\beta} \right)^2,$$

  and so with σ fixed we can infer the parameter β. Compare with the original OU example.
- If we assume $a_j \propto j^{-a}$, $\gamma_N \propto N^{-w}$, then Assumptions II (iv) is equivalent to

$$w > a + \frac{\nu}{d} + \frac{1}{2}.$$

Application Automatic Relevance Determination

- With ARD priors, hyperparameters are used to learn the most important coordinates of the input.
- Given an isotropic covariance function $c(x, x') = h(\|x - x'\|)$, $x, x' \in \mathbb{R}^d$, define

$$c(x, x'|\theta) = h\left(\left[\sum_{k=1}^{d}\left(\frac{x_k - x'_k}{\theta_k}\right)^2\right]^{1/2}\right)$$

- If θ_k is inferred to be large, the kth coordinate is deemed to be less relevant.
- ARD versions of Whittle-Matérn priors may be obtained by replacing the Laplacian Δ with the anisotropic Laplacian $\Delta_\theta = \sum_{k=1}^{d} \theta_k^2 \partial_{x_k}^2$ in their SPDE representation.
- Proof of Assumptions II holding is almost the same as previous example.

Outline

Numerical Illustration Problem Setup

- We consider a linear inverse source problem on $L^2(0,1)$, and generate the true state from Matérn prior with $\nu^\dagger = 3/2$ and $\sigma^\dagger = \ell^\dagger = 1$.
- The forward map is explicity given by

$$\langle Au, \varphi_j \rangle = \begin{cases} j^{-2}\langle u, \varphi_j \rangle & j \geq 1 \\ 0 & j = 0 \end{cases}$$

  where $\{\varphi_j\}$ is the cosine Fourier basis for $L^2(0,1)$.
- We therefore have $a_j \asymp j^{-2}$.
- We consider recovery of σ, ℓ or both.

Numerical Illustration Convergence of Errors

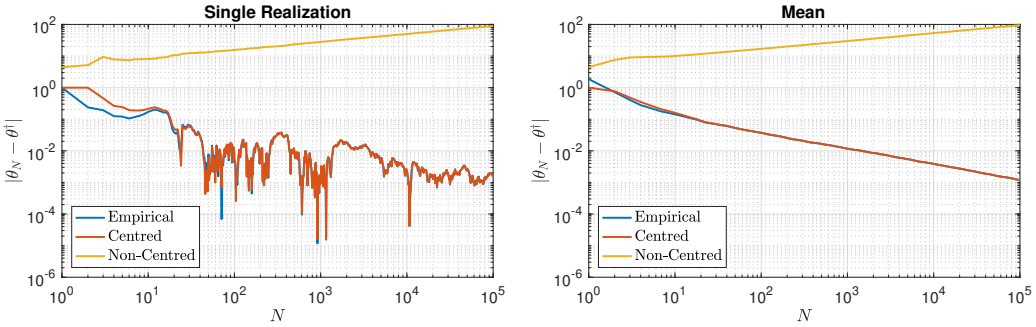

Figure: Errors between minimizers of the functionals versus state/data space dimension for (left) one realizations of the data and (right) averaged over 1000 such realizations.

Numerical Illustration Curve of Equivalent Measures

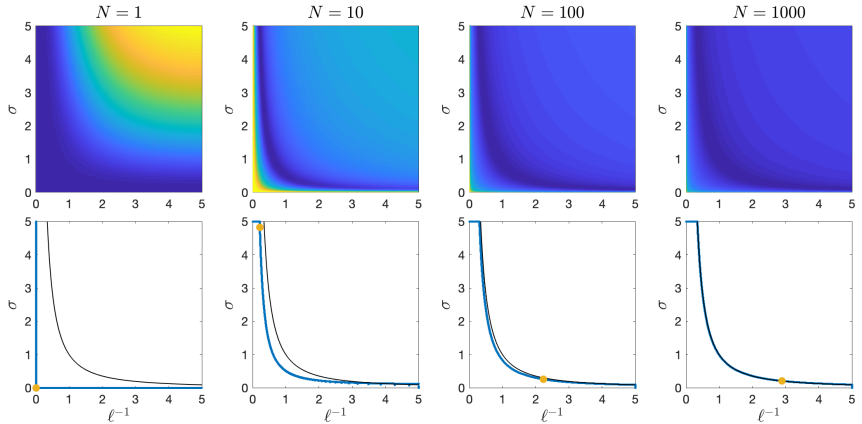

Figure: Objective functions and their minimizers on the pair (σ, ℓ^{-1}), as the state/data space dimension is increased. Red curve shows $\sigma/\ell = \sigma^\dagger/\ell^\dagger$

Numerical Illustration Sharpness of Results

- Recall, for a Whittle-Matern prior, when the algebraic decays $a_j \propto j^{-a}$, $\gamma_N \propto N^{-w}$ are assumed, the condition
$$w > a + \frac{\nu}{d} + \frac{1}{2}$$
is equivalent to one of the assumptions for the theory.
- We check numerically whether this condition is sharp for the above problem.
- In our setup, this condition reduces to $w > 4$.
- We consider the choices $w = 3.5, 4, 4.5$ for both centred and empirical methods, and look at the errors as the state/data space dimension is increased.

Numerical Illustration Sharpness of Results

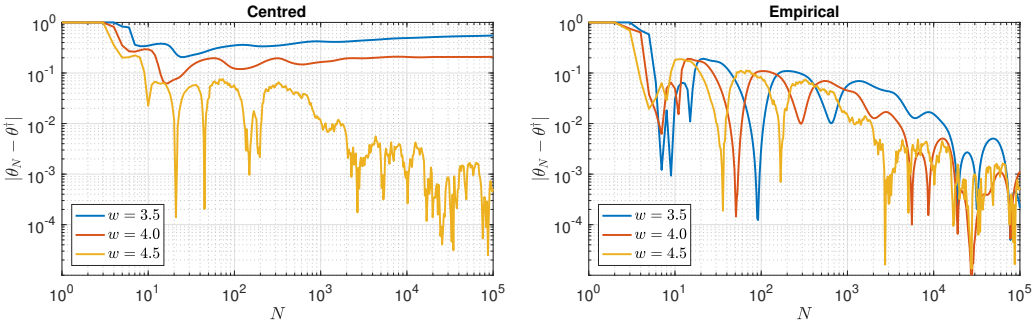

Figure: Errors between minimizers of the functionals versus state/data space dimension for different noise decay rates.

Outline

Conclusions

- For MAP estimation, centred parameterizations are preferable to noncentred parameterizations when a goal of the inference is recovery of the hyperparameters θ.
- The relative merits of centring and noncentring in this context differ from what is found for sampling methods such as MCMC.
- We provide conditions on the data model and prior distribution that lead to theorems describing the recovery, or lack of recovery, of the true hyperparameters in the simultaneous large data/small noise limit.
- The theory also holds for empirical Bayesian estimation of hyperparameters; numerically this appears to be more robust than using the centred parameterization.
- Hyperparameter recovery holds only up to measure equivalence, in a certain sense.

References I

[1] M. M. Dunlop, T. Helin and A. M. Stuart. *Hyperparameter Estimation in Bayesian MAP Estimation: Parameterizations and Consistency*. In revision, 2019.

[2] B. T. Knapik, B. T. Szabó, A. W. van der Vaart, and J. H. van Zanten. *Bayes procedures for adaptive inference in inverse problems for the white noise model*. Probability Theory and Related Fields, 2016.

[3] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[4] V. Chen, M. M. Dunlop, O. Papaspiliopoulos and A. M. Stuart. *Robust MCMC Sampling in High Dimensions with Non-Gaussian and Hierarchical Priors*. Submitted, 2018.

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

[6] M. Dashti, K. J. H. Law, A. M. Stuart and J. Voss. *MAP Estimators and Their Consistency in Bayesian Nonparametric Inverse Problems*. Inverse Problems, 2013.

References II

[7] F. Lindgren, H. Rue and J. Lindström. *An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach*. JRSSB, 2011.

[8] S. Agapiou, J. M. Bardsley, O. Papaspiliopoulos and A.M. Stuart. *Analysis of the Gibbs sampler for hierarchical inverse problems*. SIAM JUQ, 2014.

[9] D. Calvetti and E. Somersalo. *An introduction to Bayesian scientific computing: ten lectures on subjective computing*. Springer Science & Business Media, 2017.

[10] L. Roininen, M. Girolami, S. Lasanen and M. Markkanen. *Hyperpriors for Matérn fields with applications in Bayesian inversion*. Inverse Problems & Imaging, 2019.