UQ and Inverse Problems

Oliver Ernst

Short Course: **An Introduction to Uncertainty Quantification**

DTU Compute
Technical University of Denmark
November 24–25, 2016

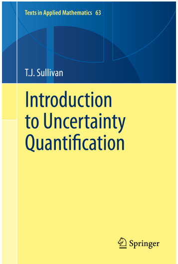
Mathematik!
TU Chemnitz

Uncertainty Quantification

Uncertainty Quantification

. . . but what is UQ? It is, roughly put, the coming together of probability theory and statistical practice with 'the real world'.

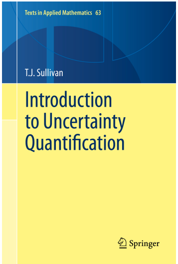
T. Sullivan, 2015



Uncertainty Quantification

. . . but what is UQ? It is, roughly put, the coming together of probability theory and statistical practice with 'the real world'.

T. Sullivan, 2015



Statisticians can't compute and numerical analysts can't handle data.

Unnamed colleague, 2016

Uncertainty Quantification

There are known knowns;
there are things we know we know.

We also know there are known unknowns;
that is to say, we know there are some things we do not know.

But there are also unknown unknowns – the ones we don't know we don't know.

Donald Rumsfeld, U.S. Secretary of Defense
DoD News Briefing; Feb. 12, 2002

- There are branches of science where certainty is unattainable.
- Examples
  - quantum physics
    Heisenberg uncertainty principle
  - geosciences
    properties of the subsurface, weather, climate
  - engineering
    manufacturing variations, impurities, sub-scale effects, structural safety, reliability analysis
  - finance
    price/interest fluctuations, Knightian uncertainty, ambiguity
- For many (routine) computational problems the effects of uncertainty outweigh other error sources (roundoff, discretisation);
  in others, it is at least an error worth considering.
- We now have the computer hardware, algorithms and data acquisition technology to address UQ computationally.

- We model uncertainty with probability. (There are alternatives.)
- Inasmuch as the associated phenomena are modeled by differential equations, the uncertain quantities enter as data.

Two fundamental problems:

(1) Given the probability law of the data (inputs), compute that of the outputs (solution, quantity of interest (QoI)).

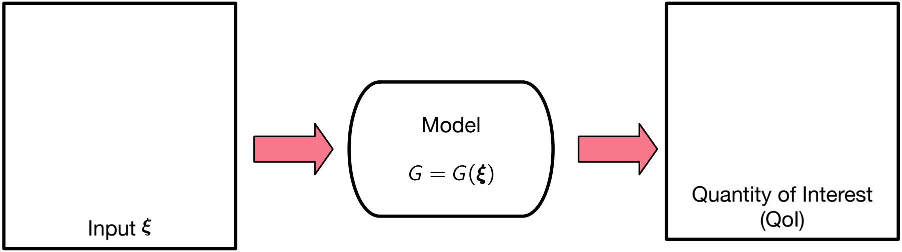
This is known as uncertainty propagation.

(2) How do we obtain the probability law of the inputs?
Merge models with observations.

This is an inverse problem for a probability measure.

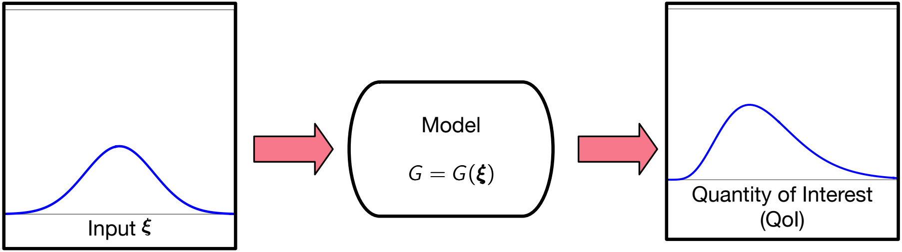
Uncertainty Quantification

Uncertainty propagation vs. Bayesian inversion



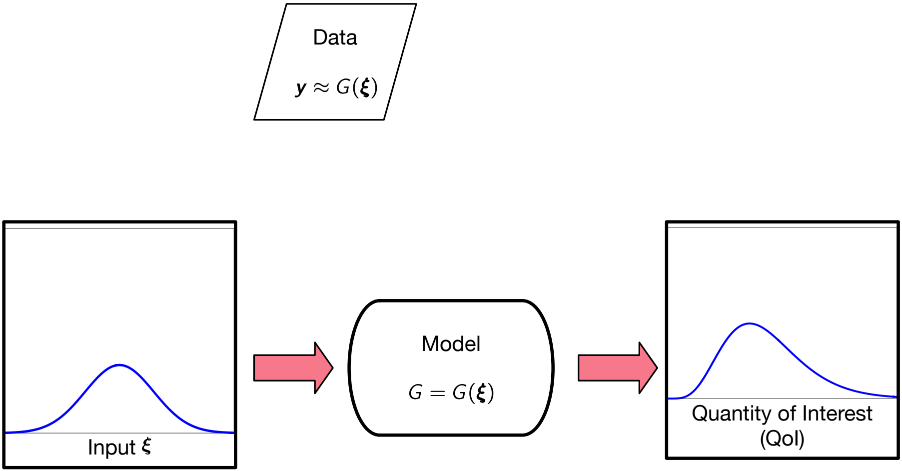
Uncertainty Quantification

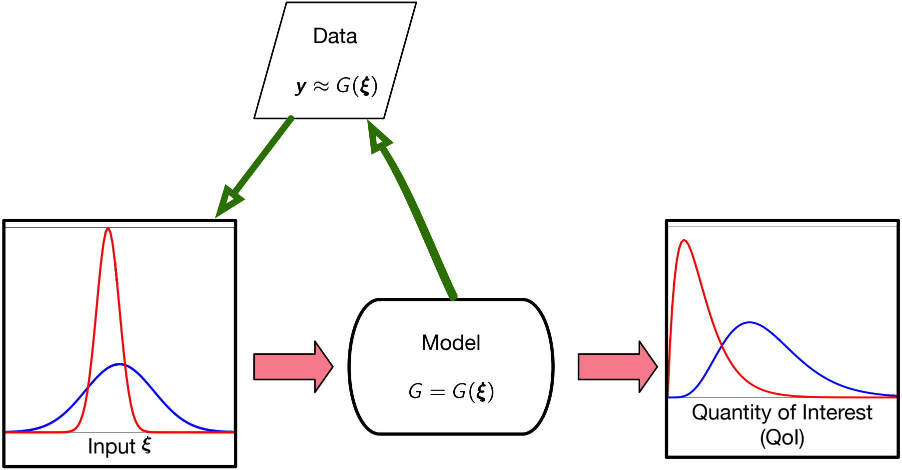
Uncertainty propagation vs. Bayesian inversion

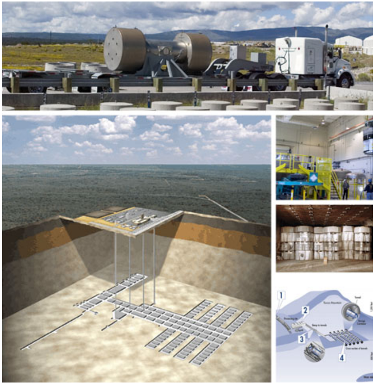


Uncertainty Quantification

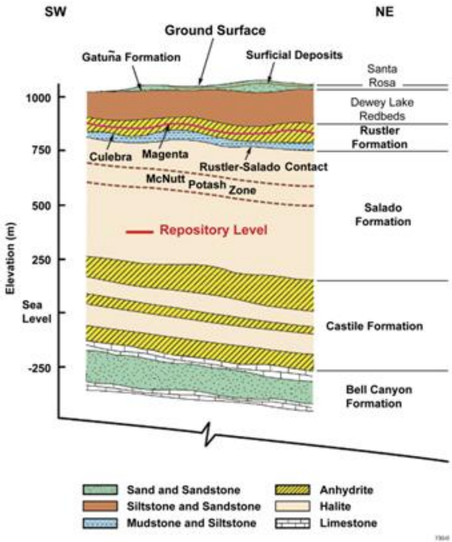
Uncertainty propagation vs. Bayesian inversion



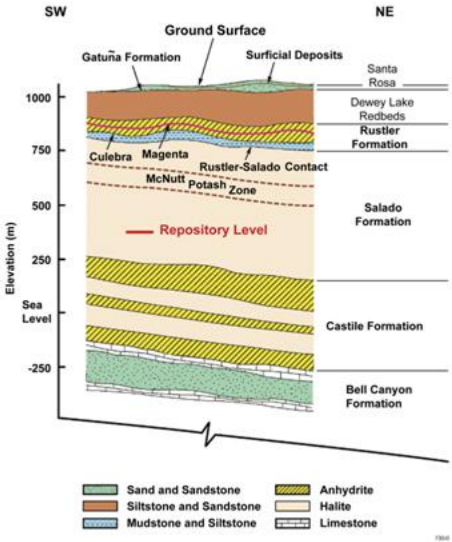

Data
$\boldsymbol{y} \approx G(\boldsymbol{\xi})$
Input ξ
Model
$G = G(\boldsymbol{\xi})$
Quantity of Interest
(QoI)



- Waste Isolation Pilot Plant (WIPP) Carlsbad, NM



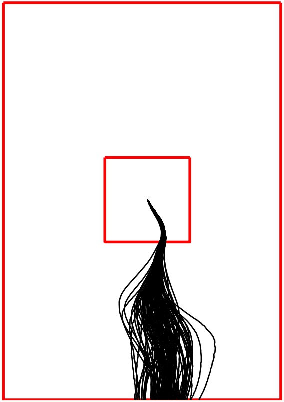
- Waste Isolation Pilot Plant (WIPP) Carlsbad, NM
- Groundwater transport of radionuclides



- Waste Isolation Pilot Plant (WIPP) Carlsbad, NM
- Groundwater transport of radionuclides
- Uncertainty in hydraulic conductivity

Groundwater Flow Problem at WIPP

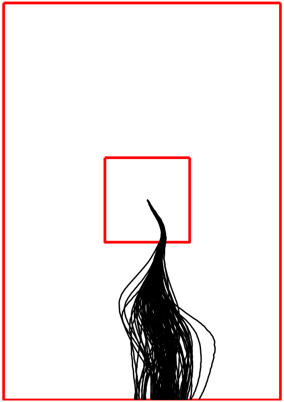
Setting



- Waste Isolation Pilot Plant (WIPP) Carlsbad, NM
- Groundwater transport of radionuclides
- Uncertainty in hydraulic conductivity
- Quantity of interest (QoI): contaminant travel time

Groundwater Flow Problem at WIPP

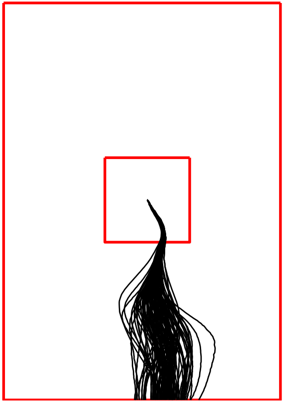
Setting



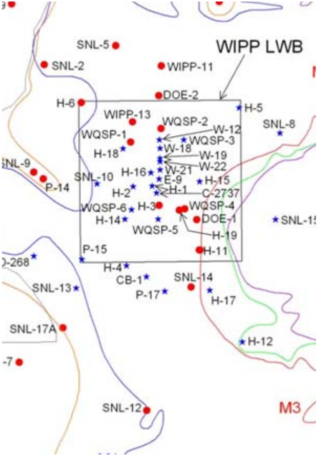
- Waste Isolation Pilot Plant (WIPP) Carlsbad, NM
- Groundwater transport of radionuclides
- Uncertainty in hydraulic conductivity
- Quantity of interest (QoI): contaminant travel time
- Certification: requires travel time $> 10^4$ years in case of breach.

Groundwater Flow Problem at WIPP

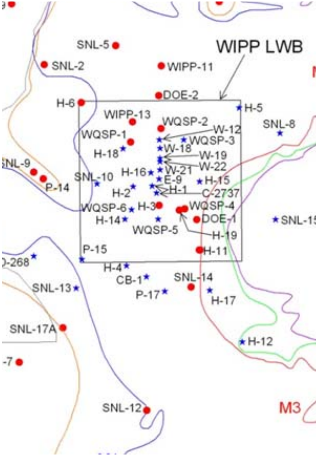
Setting



- Waste Isolation Pilot Plant (WIPP) Carlsbad, NM
- Groundwater transport of radionuclides
- Uncertainty in hydraulic conductivity
- Quantity of interest (QoI): contaminant travel time
- Certification: requires travel time $> 10^4$ years in case of breach.
- **Approach:**
  - Model uncertainty (lack of knowledge) probabilistically.
  - Merge stochastic model with direct and indirect observations.
  - Determine probability law of travel time.



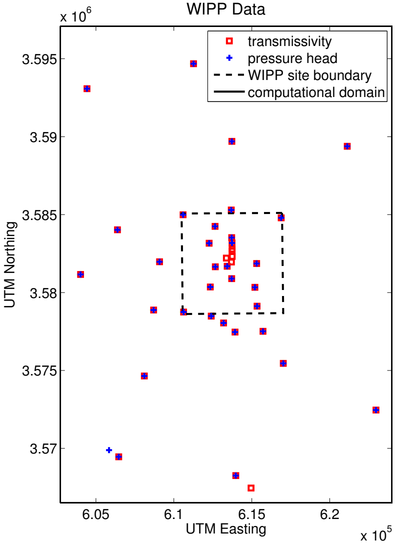
- Data publically available: Sandia TRs 1990s, certification/recertification documents 1996–2014



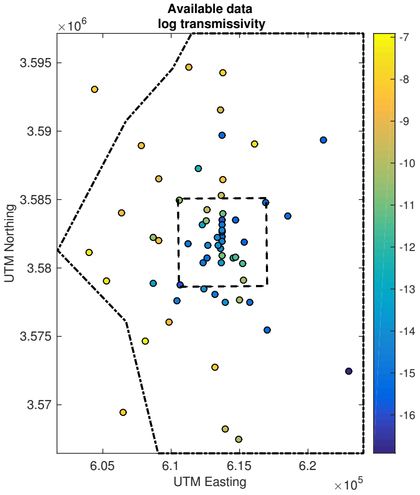
- Data publically available: Sandia TRs 1990s, certification/recertification documents 1996–2014
- Observations of direct (transmissivity) and indirect (hydraulic head) data.

Groundwater Flow Problem at WIPP

Data



- Data publically available: Sandia TRs 1990s, certification/recertification documents 1996–2014
- Observations of direct (transmissivity) and indirect (hydraulic head) data.
- WIPP-2 test data [Caufman et al., 1990], [La Venue et al., 1990]; 38 measurements of head and transmissivity



- Data publically available: Sandia TRs 1990s, certification/recertification documents 1996–2014
- Observations of direct (transmissivity) and indirect (hydraulic head) data.
- WIPP-2 test data [Caufman et al., 1990], [La Venue et al., 1990]; 38 measurements of head and transmissivity
- CRA-2014 data: measurements of head (44) and transmissivity (62).

Particle transport by groundwater modeled as ODE:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{u}(\boldsymbol{x}(t)), \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

Groundwater flux $\boldsymbol{u}$ given by Darcy's law

$$\boldsymbol{u}(\boldsymbol{x}) = -a(\boldsymbol{x})\nabla p(\boldsymbol{x})$$

relating hydraulic conductivity a and hydraulic head (pressure) p.

Mass conservation yields elliptic PDE for p:

$$-\nabla \cdot (a(\boldsymbol{x})\, \nabla p(\boldsymbol{x})) = 0 \qquad \text{on } D. \tag{PDE}$$

Boundary conditions on $\partial D = \Gamma_N \cup \Gamma_D$

$$\partial_{\boldsymbol{n}} p\big|_{\Gamma_N} = 0, \qquad p\big|_{\Gamma_D} = g.$$

Particle transport by groundwater modeled as ODE:

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{u}(\boldsymbol{x}(t)), \qquad \boldsymbol{x}(0) = \boldsymbol{x}_0.$$

Groundwater flux $\boldsymbol{u}$ given by Darcy's law

$$\boldsymbol{u}(\boldsymbol{x}) = -a(\boldsymbol{x})\nabla p(\boldsymbol{x})$$

relating hydraulic conductivity a and hydraulic head (pressure) p.

Mass conservation yields elliptic PDE for p:

$$-\nabla \cdot (a(\boldsymbol{x})\, \nabla p(\boldsymbol{x})) = 0 \qquad \text{on } D. \tag{PDE}$$

Boundary conditions on $\partial D = \Gamma_N \cup \Gamma_D$

$$\partial_{\boldsymbol{n}} p\big|_{\Gamma_N} = 0, \qquad p\big|_{\Gamma_D} = g.$$

Conductivity a and Dirichlet data g unknown – have to be estimated by data.

Main interest is in flux $\boldsymbol{u}$, therefore mixed (weak) form of (PDE) is solved:

Find $(\boldsymbol{u}, p) \in H_0(\mathrm{div}; D) \times L^2(D)$ such that

$$\langle a^{-1}\boldsymbol{u}, \boldsymbol{v}\rangle - \langle \nabla\cdot \boldsymbol{v}, p\rangle = \ell(\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in H_0(\mathrm{div}; D), \qquad \text{(PDE-mixed-a)}$$

$$\langle \nabla\cdot \boldsymbol{u}, q\rangle = 0 \qquad \forall q \in L^2(D), \qquad \text{(PDE-mixed-b)}$$

where $\langle\cdot,\cdot\rangle$ denotes $L^2(D)$-inner product, $\ell(\boldsymbol{v}) = -\int_{\Gamma_D} g\, \boldsymbol{v}\cdot\vec{\boldsymbol{n}}\, \mathrm{d}s$ and

$$H_0(\mathrm{div}; D) = \big\{ \boldsymbol{v} \in \boldsymbol{L}^2(D) : \nabla\cdot \boldsymbol{v} \in L^2(D), \\ \langle \boldsymbol{v}, \nabla v\rangle + \langle \nabla\cdot \boldsymbol{v}, v\rangle = 0 \quad \forall v \in H^1_{0,\Gamma_D}(D)\big\}.$$

Main interest is in flux $\boldsymbol{u}$, therefore mixed (weak) form of (PDE) is solved:

Find $(\boldsymbol{u}, p) \in H_0(\mathrm{div}; D) \times L^2(D)$ such that

$$\langle a^{-1}\boldsymbol{u}, \boldsymbol{v}\rangle - \langle \nabla\cdot \boldsymbol{v}, p\rangle = \ell(\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in H_0(\mathrm{div}; D), \qquad \text{(PDE-mixed-a)}$$

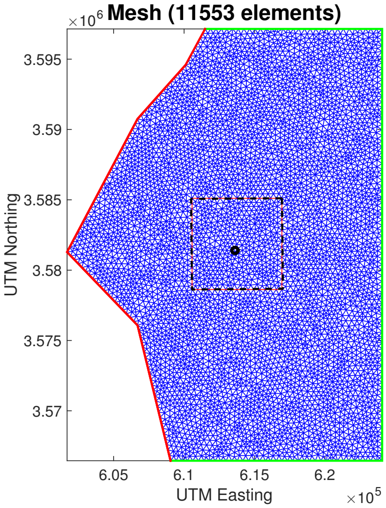
$$\langle \nabla\cdot \boldsymbol{u}, q\rangle = 0 \qquad \forall q \in L^2(D), \qquad \text{(PDE-mixed-b)}$$

where $\langle\cdot,\cdot\rangle$ denotes $L^2(D)$-inner product, $\ell(\boldsymbol{v}) = -\int_{\Gamma_D} g\, \boldsymbol{v}\cdot\vec{\boldsymbol{n}}\, ds$ and

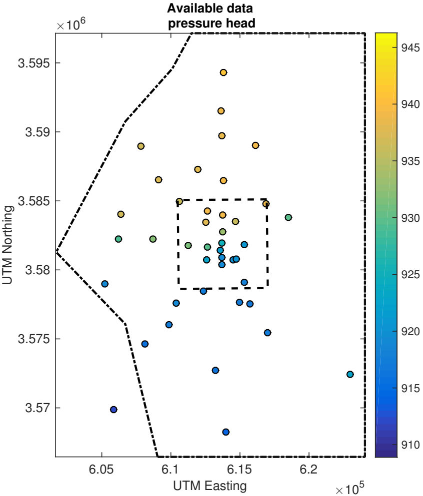
$$H_0(\mathrm{div}; D) = \big\{\boldsymbol{v} \in \boldsymbol{L}^2(D) : \nabla\cdot \boldsymbol{v} \in L^2(D), \\ \langle \boldsymbol{v}, \nabla v\rangle + \langle \nabla\cdot \boldsymbol{v}, v\rangle = 0 \quad \forall v \in H^1_{0,\Gamma_D}(D)\big\}.$$

- 2D model, as Culebra only 7.75m thick over area 20km $\times$ 30km.
- FE discretization of (PDE-mixed) using lowest-order Raviart-Thomas elements ($\boldsymbol{u}$) / piecewise constants (p).
- Flow divergence-free, thus $\boldsymbol{u}_h$ pcw. constant, particle tracking trivial.

Computational domain with Neumann and Dirichlet boundary Γ_N and Γ_D, resp.



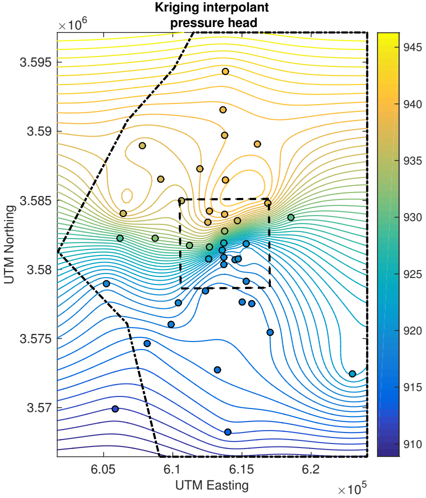
Head data (CRA 2014) ...



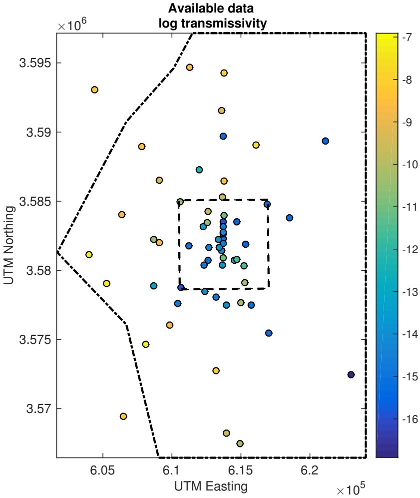
Groundwater Flow Problem at WIPP

Deterministic Calculation

Head data (CRA 2014) . . . and its geostatistical interpolant.



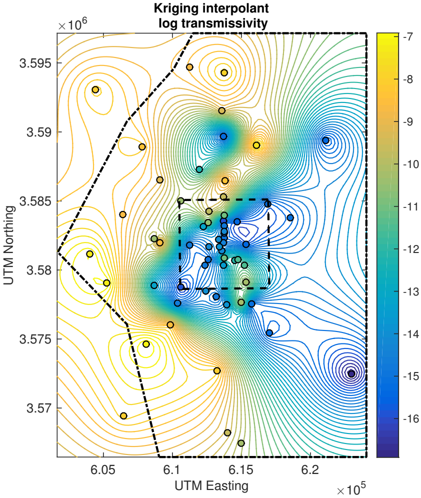
Log transmissivity data (CRA 2014) ...



Groundwater Flow Problem at WIPP

Deterministic Calculation

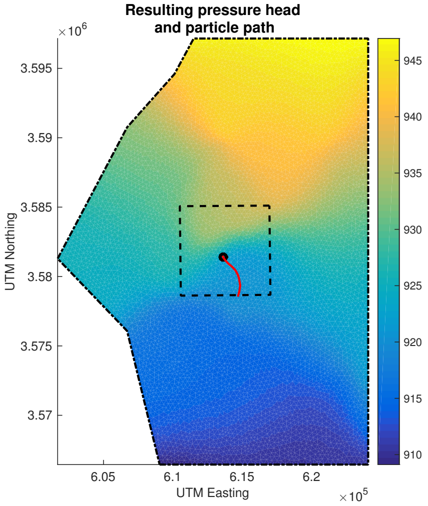
Log transmissivity data (CRA 2014) . . . and its geostatistical interpolant.



Groundwater Flow Problem at WIPP

Deterministic Calculation

Particle travel time given these estimates for a and g: $\boxed{18,424 \text{ years}}$.



Groundwater Flow Problem at WIPP

Re-evaluate

- What if true transmissivity a differs from best estimate?

  Have estimated a on a domain of 20 km $\times$ 30 km based on 62 data points!

Groundwater Flow Problem at WIPP

Re-evaluate

- What if true transmissivity a differs from best estimate?

  Have estimated a on a domain of 20 km $\times$ 30 km based on 62 data points!
- Need to quantify the effects of remaining uncertainty regarding a and g.

- What if true transmissivity a differs from best estimate?

  Have estimated a on a domain of 20 km $\times$ 30 km based on 62 data points!
- Need to quantify the effects of remaining uncertainty regarding a and g.

Alternative:

- Model a as random field $a : D \times \Omega \to \mathbb{R}$ w.r.t. probability space $(\Omega, \mathfrak{A}, \mathbf{P})$.
- In other words, $a(\boldsymbol{x}, \cdot)$ is a random variable for each $\boldsymbol{x} \in D$ where the randomness describes our uncertainty about $a(\boldsymbol{x})$.
- Therefore also p and $\boldsymbol{u}$ become random fields where now (PDE-mixed) holds $\mathbf{P}$-a.s.
- Also QoI particle travel time becomes a random variable and we aim to compute its distribution function.

Basic assumption: $\kappa := \log a$ and g are (stationary) Gaussian random fields (GRF).

Our different approaches so far:

- **Variant 1** (done): Consider g as deterministic,
  construct random field model for a from observations of a (geostatistics)
- **Variant 2** (done): In addition to Variant 1, incorporate measurements of p into model for a (Bayesian inversion)
- **Variant 3** (in progress): In addition to Variant 2, model also g as random field incorporating observational data.

[Cliffe, Ernst, Sprungk, Ullmann & van den Boogart, 2016],
[Ernst & Sprungk, 2016]

Plan of Lectures

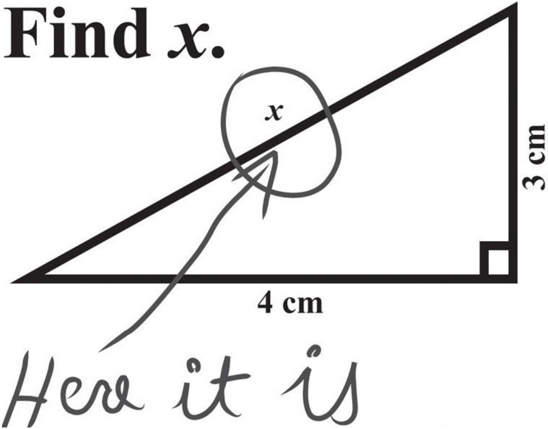
1. Inverse Problems
2. Bayesian Inference
3. Sampling from the Posterior
4. An Inverse Problem for Groundwater Flow at WIPP

Contents

1. Inverse Problems
2. Bayesian Inference
3. Sampling from the Posterior
4. An Inverse Problem for Groundwater Flow at WIPP

Contents

1 Inverse Problems
1.1 Introduction
1.2 Ill-Conditioned Linear Problems, Regularization
1.3 Infinite-Dimensional Problems
1.4 Outlook


Find x.
x
3 cm
4 cm
Here it is

- Inverse problems are concerned with determining a quantity x which is not directly observable.

- Inverse problems are concerned with determining a quantity x which is not directly observable.
- Available partial knowledge about x: its image y under a mapping F:

$$F(x) = y.$$

- Inverse problems are concerned with determining a quantity x which is not directly observable.
- Available partial knowledge about x: its image y under a mapping F:

$$F(x) = y.$$

F is called the forward map.

- Inverse problems are concerned with determining a quantity x which is not directly observable.
- Available partial knowledge about x: its image y under a mapping F:

$$F(x) = y.$$

F is called the forward map.
y is called the data, observation or measurement.

- Inverse problems are concerned with determining a quantity x which is not directly observable.
- Available partial knowledge about x: its image y under a mapping F:

$$F(x) = y.$$

F is called the forward map.
y is called the data, observation or measurement.
x is called the solution, unknown, parameter or reconstruction.

- Inverse problems are concerned with determining a quantity x which is not directly observable.
- Available partial knowledge about x: its image y under a mapping F:

$$F(x) = y.$$

F is called the forward map.
y is called the data, observation or measurement.
x is called the solution, unknown, parameter or reconstruction.
Determining y given x is called the direct of forward problem.

- Inverse problems are concerned with determining a quantity x which is not directly observable.
- Available partial knowledge about x: its image y under a mapping F:

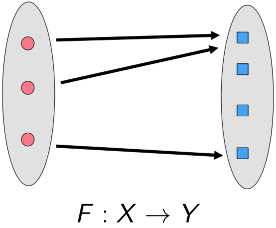
$$F(x) = y.$$

  F is called the forward map.
  y is called the data, observation or measurement.
  x is called the solution, unknown, parameter or reconstruction.
  Determining y given x is called the direct of forward problem.

- Other common characterizations of inverse problems:

  Determining a cause from its effect.
  Reconstructing an object from partial observations.
  Constructing a geometrical body from its projections.
  Concluding the input from knowing the output.

The forward map at the set level:

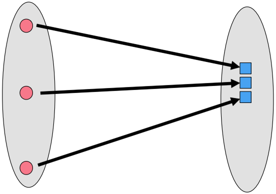


Properties:

- well-definedness
- surjectivity
- injectivity
- (solvability)

- For bijective F, inverse mapping F^{-1} exists, no (qualitative) difference between forward and inverse problem at set-theoretic level.
- F not surjective: restrict problem formulation to $F(X) \subsetneq Y$.
- F not injective: additional information, constraints.

The real challenge posed by inverse problems is topological:



$F : X \to Y$

- F "almost not injective", i.e.,
- F^{-1} is not continuous, even when it exists.
- Small changes in y correspond to large changes in x.

- Well-posed problems in the sense of [Hadamard, 1923] possess unique solutions x which depend continuously on the data y.
- Last point crucial since, in applications, data always contaminated by noise due to measurement error, discretization error, floating point error or uncertainty.
- In this sense inverse problems are ill-posed problems.

Consider linear forward map

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \quad \boldsymbol{x} \mapsto \boldsymbol{y} = F(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \qquad \boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon \end{bmatrix}, \quad \epsilon > 0,$$

with data vectors

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{(unperturbed)}, \qquad \boldsymbol{y}^\delta = \begin{bmatrix} 1 \\ \delta \end{bmatrix}, \ \delta > 0 \text{ (perturbed)}.$$

Then

$$\boldsymbol{x} := \boldsymbol{A}^{-1}\boldsymbol{y} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \boldsymbol{x}^\delta := \boldsymbol{A}^{-1}\boldsymbol{y}^\delta = \begin{bmatrix} 1 \\ \delta/\epsilon \end{bmatrix},$$

giving

$$\frac{\|\boldsymbol{x} - \boldsymbol{x}^\delta\|_p}{\|\boldsymbol{y} - \boldsymbol{y}^\delta\|_p} = \frac{\delta/\epsilon}{\delta} = \frac{1}{\epsilon}, \qquad \text{e.g. for } p = 1, 2, \infty.$$

For $0 < \epsilon \ll 1$ noise level strongly amplified in inversion process.

Same applies to any square diagonal matrix

$$\boldsymbol{A} = \begin{bmatrix} \epsilon_1 & & & \\ & \epsilon_2 & & \\ & & \ddots & \\ & & & \epsilon_n \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad \epsilon_1 \ge \cdots \ge \epsilon_n > 0.$$

For data and noise vectors $\boldsymbol{y}, \boldsymbol{0} \neq \boldsymbol{\delta} \in \mathbb{R}^n$, $\boldsymbol{y}^\delta := \boldsymbol{y} + \boldsymbol{\delta}$, we obtain

$$\frac{\|\boldsymbol{x} - \boldsymbol{x}^\delta\|_2}{\|\boldsymbol{y} - \boldsymbol{y}^\delta\|_2} = \frac{\|\boldsymbol{A}^{-1}\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}\|_2} \ge \frac{1/\epsilon_n \|\boldsymbol{\delta}\|_2}{\|\boldsymbol{\delta}\|_2} = \frac{1}{\epsilon_n}.$$

- Such forward mappings are called ill-conditioned.
- Since $\boldsymbol{A}^{-1}$ is continuous, the inverse problem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is, strictly speaking, still well-posed.
- True ill-posedness resides in mappings between infinite-dimensional spaces.

Contents

1 Inverse Problems
1.1 Introduction
1.2 Ill-Conditioned Linear Problems, Regularization
1.3 Infinite-Dimensional Problems
1.4 Outlook

Every matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has a singular value decomposition (SVD)

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \qquad \boldsymbol{U} \in \mathbb{R}^{m \times m}, \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_m, \quad \boldsymbol{V} \in \mathbb{R}^{n \times n}, \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_n,$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}, \ \boldsymbol{\Sigma}_r = \operatorname{diag}(\sigma_1, \dots, \sigma_r), \ \sigma_1 \geq \dots \geq \sigma_r > 0,$$

$$r = \operatorname{rank}(\boldsymbol{A}), \ 0 \leq r \leq \min\{m, n\}.$$

Every matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has a singular value decomposition (SVD)

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \qquad \boldsymbol{U} \in \mathbb{R}^{m\times m}, \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_m, \quad \boldsymbol{V} \in \mathbb{R}^{n\times n}, \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_n,$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}, \ \boldsymbol{\Sigma}_r = \operatorname{diag}(\sigma_1, \dots, \sigma_r), \ \sigma_1 \geq \cdots \geq \sigma_r > 0,$$

$$r = \operatorname{rank}(\boldsymbol{A}), \ 0 \leq r \leq \min\{m, n\}.$$

The equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is therefore equivalent with

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top\boldsymbol{x} = \boldsymbol{y} \quad \Leftrightarrow \quad \boldsymbol{\Sigma}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{y}}, \quad \tilde{\boldsymbol{x}} := \boldsymbol{V}^\top\boldsymbol{x}, \quad \tilde{\boldsymbol{y}} := \boldsymbol{U}^\top\boldsymbol{y}.$$

Every matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ has a singular value decomposition (SVD)

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \qquad \boldsymbol{U} \in \mathbb{R}^{m \times m}, \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}_m, \quad \boldsymbol{V} \in \mathbb{R}^{n \times n}, \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_n,$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}, \ \boldsymbol{\Sigma}_r = \operatorname{diag}(\sigma_1, \dots, \sigma_r), \ \sigma_1 \geq \cdots \geq \sigma_r > 0,$$

$$r = \operatorname{rank}(\boldsymbol{A}), \ 0 \leq r \leq \min\{m, n\}.$$

The equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$ is therefore equivalent with

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{x} = \boldsymbol{y} \quad \Leftrightarrow \quad \boldsymbol{\Sigma}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{y}}, \quad \tilde{\boldsymbol{x}} := \boldsymbol{V}^\top \boldsymbol{x}, \quad \tilde{\boldsymbol{y}} := \boldsymbol{U}^\top \boldsymbol{y}.$$

For $\boldsymbol{y}^\delta := \boldsymbol{y} + \boldsymbol{\delta}$ and $\boldsymbol{A}\boldsymbol{x}^\delta = \boldsymbol{y}^\delta$ we have

$$\frac{\|\boldsymbol{x} - \boldsymbol{x}^\delta\|_2}{\|\boldsymbol{y} - \boldsymbol{y}^\delta\|_2} = \frac{\|\tilde{\boldsymbol{x}} - \tilde{\boldsymbol{x}}^\delta\|_2}{\|\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{y}}^\delta\|_2} \geq \frac{1}{\sigma_r}.$$

Writing the transformed equation as

$$\begin{bmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{x}}_r \\ \tilde{\boldsymbol{x}}_\perp \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{y}}_r \\ \tilde{\boldsymbol{y}}_\perp \end{bmatrix}, \qquad \tilde{\boldsymbol{x}} = \begin{bmatrix} \tilde{\boldsymbol{x}}_r \\ \tilde{\boldsymbol{x}}_\perp \end{bmatrix}, \quad \tilde{\boldsymbol{y}} = \begin{bmatrix} \tilde{\boldsymbol{y}}_r \\ \tilde{\boldsymbol{y}}_\perp \end{bmatrix},$$

we observe

- no solution unless $\tilde{\boldsymbol{y}}_\perp = \boldsymbol{0} \Leftrightarrow \boldsymbol{y} \in \operatorname{range}(\boldsymbol{A})$,
- solution block $\tilde{\boldsymbol{x}}_\perp$ arbitrary, since $\tilde{\boldsymbol{x}}_\perp \in \operatorname{null}(\boldsymbol{A})$.

Writing the transformed equation as

$$\begin{bmatrix} \boldsymbol{\Sigma}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{x}}_r \\ \tilde{\boldsymbol{x}}_\perp \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{y}}_r \\ \tilde{\boldsymbol{y}}_\perp \end{bmatrix}, \qquad \tilde{\boldsymbol{x}} = \begin{bmatrix} \tilde{\boldsymbol{x}}_r \\ \tilde{\boldsymbol{x}}_\perp \end{bmatrix}, \quad \tilde{\boldsymbol{y}} = \begin{bmatrix} \tilde{\boldsymbol{y}}_r \\ \tilde{\boldsymbol{y}}_\perp \end{bmatrix},$$

we observe

- no solution unless $\tilde{\boldsymbol{y}}_\perp = \mathbf{0} \Leftrightarrow \boldsymbol{y} \in \operatorname{range}(\boldsymbol{A})$,
- solution block $\tilde{\boldsymbol{x}}_\perp$ arbitrary, since $\tilde{\boldsymbol{x}}_\perp \in \operatorname{null}(\boldsymbol{A})$.

Generalized solution $\boldsymbol{x}^\dagger := \boldsymbol{V} \begin{bmatrix} \boldsymbol{\Sigma}_r^{-1} \tilde{\boldsymbol{y}}_r \\ \mathbf{0} \end{bmatrix}$ (least squares solution, LS-solution)

- minimizes $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2$ among all $\boldsymbol{x} \in \mathbb{R}^n$,
- has the smallest 2-norm among all $\boldsymbol{x} \in \mathbb{R}^n$ that satisfy $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$,
- can be written in terms of the (Moore-Penrose-)Pseudoinverse $\boldsymbol{A}^\dagger$ as

$$\boldsymbol{x}^\dagger = \boldsymbol{A}^\dagger \boldsymbol{y}, \quad \boldsymbol{A}^\dagger := \boldsymbol{V} \boldsymbol{\Sigma}^\dagger \boldsymbol{U}^\top, \quad \boldsymbol{\Sigma}^\dagger := \begin{bmatrix} \boldsymbol{\Sigma}_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Regularization

Truncated singular value decomposition (TSVD)

- In terms of the SVD the LS solution is

$$\boldsymbol{x}^\delta = \boldsymbol{V}\tilde{\boldsymbol{x}}^\delta = \sum_{j=1}^{r} \frac{\boldsymbol{u}_j^\top \boldsymbol{y}^\delta}{\sigma_j} \boldsymbol{v}_j, \qquad \boldsymbol{V} = [\boldsymbol{v}_1 | \cdots | \boldsymbol{v}_n], \quad \boldsymbol{U} = [\boldsymbol{u}_1 | \cdots | \boldsymbol{u}_m].$$

- Since $\boldsymbol{u}_j^\top \boldsymbol{y}^\delta = \boldsymbol{u}_j^\top \boldsymbol{y} + \boldsymbol{u}_j^\top \boldsymbol{\delta}$, noise amplification occurs in those (singular vector expansion) components for which $|\boldsymbol{u}_j^\top \boldsymbol{\delta}| > \sigma_j$.
- Regularization: introduce (parameter-dependent) weighting factor

$$w_\alpha : (0, \sigma_1^2] \to [0, 1], \qquad w_\alpha(\sigma^2) := \begin{cases} 1 & \text{if } \sigma^2 > \alpha \\ 0 & \text{otherwise} \end{cases},$$

giving the (TSVD-)regularized solution

$$\boldsymbol{x}_\alpha^\delta := \sum_{j=1}^{r} \frac{w_\alpha(\sigma_j^2)}{\sigma_j} (\boldsymbol{u}_j^\top \boldsymbol{y}^\delta)\, \boldsymbol{v}_j = \sum_{\sigma_j > \alpha} \sigma_j^{-1} (\boldsymbol{u}_j^\top \boldsymbol{y}^\delta)\, \boldsymbol{v}_j.$$

- Denoting this regularization operator by R_α, we obtain

$$\boldsymbol{x}_\alpha^\delta = R_\alpha \boldsymbol{y}^\delta \quad \text{in place of} \quad \boldsymbol{x}^\delta = \boldsymbol{A}^\dagger \boldsymbol{y}^\delta.$$

- For $\|\tilde{\boldsymbol{\delta}}\|_2 = \|\boldsymbol{U}^\top \boldsymbol{\delta}\|_2 = \delta > 0$, assuming roughly uniform noise across components, this suggests $\tilde{\delta}_j := \boldsymbol{u}_j^\top \boldsymbol{\delta} \approx \delta/\sqrt{m}$ so that truncating the terms where $|\boldsymbol{u}_j^\top \boldsymbol{\delta}| > \sigma_j$ corresponds to $\alpha = |\boldsymbol{u}_j^\top \boldsymbol{\delta}| = |\tilde{\delta}_j| \approx \delta/\sqrt{m}$.
- Error

$$\boldsymbol{A}^\dagger \boldsymbol{y} - R_\alpha \boldsymbol{y}^\delta = \underbrace{\boldsymbol{A}^\dagger \boldsymbol{y} - R_\alpha \boldsymbol{y}}_{\text{approximation error}} + \underbrace{R_\alpha(\boldsymbol{y} - \boldsymbol{y}^\delta)}_{\text{data error}}$$

- By construction

$$\boldsymbol{A}^\dagger \boldsymbol{y} - R_\alpha \boldsymbol{y} = \sum_{j=1}^{r} \underbrace{\frac{1 - w_\alpha(\sigma_j^2)}{\sigma_j}}_{\to 0 \text{ as } \alpha \to 0} (\boldsymbol{u}_j^\top \boldsymbol{y})\, \boldsymbol{v}_j$$

- Data error: $\|\boldsymbol{\delta}\|_2 = \delta$ implies

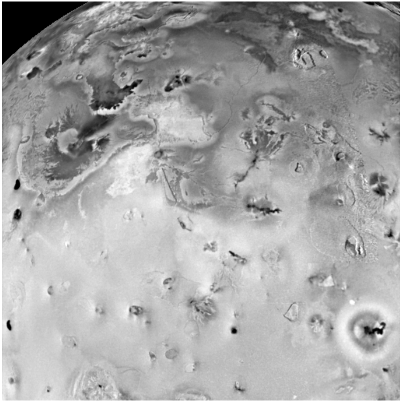
$$\|R_\alpha(\boldsymbol{y} - \boldsymbol{y}^\delta)\|_2^2 = \left\| \sum_{j=1}^{r} \underbrace{\frac{w_\alpha(\sigma_j^2)}{\sigma_j}}_{\le \alpha^{-1/2}} (\boldsymbol{u}_j^\top \boldsymbol{\delta})\, \boldsymbol{v}_j \right\|_2^2 \le \frac{\delta^2}{\alpha}$$

and therefore the regularization scheme of choosing

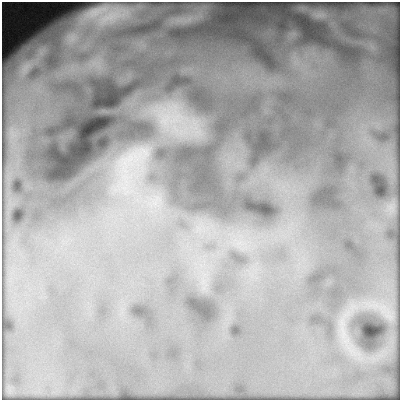
$$\alpha = \alpha(\delta) := \delta^p, \quad 0 < p < 2$$

leads to $\boldsymbol{x}_\alpha^\delta \to \boldsymbol{x}^\dagger$ as $\delta \to 0$.

[Hansen, Nagy & O'Leary, 2006]: Image modeled as pixel vector $\boldsymbol{x}$, blurring due to atmosphere modelled by discrete convolution $\boldsymbol{A}$ (point-spread-function).

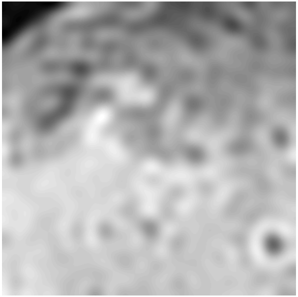


Original image $\boldsymbol{x}$

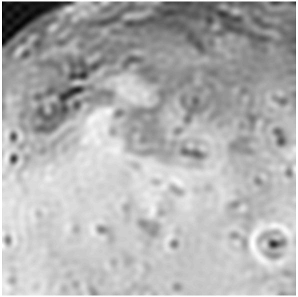


blurred and noisy $\boldsymbol{y} + \boldsymbol{\delta}$

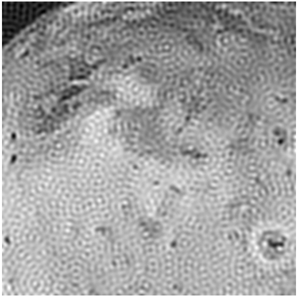
Three different reconstructions (deblurring) via TSVD.
(k refers to the truncation index of the singular values.)



$k = 658$



$k = 2813$



$k = 7243$

In place of $\boldsymbol{A}^\dagger \boldsymbol{y}^\delta$, choose $\boldsymbol{x}_\alpha^\delta$ to minimize the Tikhonov functional

$$\underbrace{\frac{1}{2}\|\boldsymbol{y}^\delta - \boldsymbol{A}\boldsymbol{x}\|_2^2}_{\text{data misfit}} \quad + \quad \underbrace{\frac{\alpha}{2}\|\boldsymbol{x}\|^2}_{\text{regularization functional}} \qquad \alpha > 0.$$

- Corresponds to weighting function $w_\alpha(\sigma^2) = \frac{\sigma^2}{\sigma^2+\alpha}$.
- Also leads to convergent regularization scheme as $\delta \to 0$.
- Implementation does not require SVD, can be computed, e.g., with iterative methods.
- Prototype of optimization-based schemes, also known as variational regularization.
- Different types of added information can be encoded into structure of regularizaton functional: smoothness, blockyness, sparsity etc.

Contents

1 Inverse Problems

1.1 Introduction

1.2 Ill-Conditioned Linear Problems, Regularization

1.3 Infinite-Dimensional Problems

1.4 Outlook

Now let $A : X \to Y$ be an linear operator between two Hilbert spaces X and Y.

- The problem of solving $Ax = y$ with $y \in Y$ is ill-posed in the sense of [Nashed, 1987] if $\mathrm{range}(A) = \{Ax : x \in X\}$ is not closed in Y.
- A class of operators always leading to ill-posed problems is that of compact operators with an infinite-dimensional range.
- For compact A there exist orthonormal systems $\{u_j\}_{j\in\mathbb{N}}$ and $\{v_j\}_{j\in\mathbb{N}}$ as well as a nonincreasing null-sequence of nonnegative singular values $(\sigma_j)_{j\in\mathbb{N}}$ such that

$$Ax = \sum_{j=1}^{\infty} \sigma_j (v_j, x)_X \, u_j.$$

- Picard condition: for a compact linear operator A with singular system (u_j, v_j, σ_j), an element $y \in \overline{\mathrm{range}(A)}$ also lies in $\mathrm{range}(A)$ if

$$\sum_{j=1}^{\infty} \frac{(u_j, y)_Y^2}{\sigma_j^2} < \infty.$$

For $A : X \to Y$ bounded and linear, $y \in Y$ we have

- $$\begin{aligned} Ax^\dagger = P_{\overline{\text{range}(A)}} y \quad &\Leftrightarrow \quad x^\dagger = \operatorname*{arg\,min}_{x \in X} \|y - Ax\|_Y \\ &\Leftrightarrow \quad A^* A x^\dagger = A^* y \quad (\text{normal equations}). \end{aligned}$$
- The solutions of the normal equations form a closed and convex set, which is nonempty iff $y \in \text{range}(A) \oplus \text{range}(A)^\perp \quad (\neq Y$ in general).
- The pseudoinverse

  $$A^\dagger : D(A^\dagger) \to X, \quad D(A^\dagger) := \text{range}(A) \oplus \text{range}(A)^\perp,$$

  is the linear mapping which assigns to $y \in D(A^\dagger)$ the unique minimum norm solution $x^\dagger$ of the normal equations.
- $A^\dagger$ is continuous iff $\text{range}(A)$ is closed.

For a compact operator $A : X \to Y$ with singular system $\{(u_j, v_j, \sigma_j)\}$ its pseudoinverse has the representation

$$A^\dagger y = \sum_{j=1}^{\infty} \sigma_j^{-1} (u_j, y)_Y \, v_j, \qquad y \in D(A^\dagger).$$

For a noisy vector $y^\delta \in Y$ we have

$$A^\dagger y - A^\dagger y^\delta = \sum_{j=1}^{\infty} \sigma_j^{-1} (u_j, y - y^\delta)_Y \, v_j.$$

Given the singular value sequence σ_j, one can find noise vectors $\|y - y^\delta\|_Y = \delta > 0$ such that, say, $\|A^\dagger y - A^\dagger y^\delta\|_X > 1$ for arbitrarily small δ.

Contents

1 Inverse Problems
1.1 Introduction
1.2 Ill-Conditioned Linear Problems, Regularization
1.3 Infinite-Dimensional Problems
1.4 Outlook

Outlook

Nonlinear problems

For nonlinear forward maps $F : X \to Y$ variational regularization of the problem

$$F(x) = y, \quad y \in F(X)$$

consists of minimizing

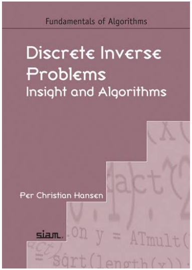
$$\|y - F(x)\|_Y^2 + R_\alpha(x)$$

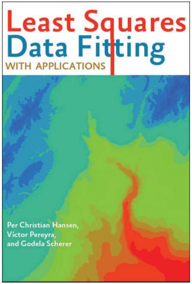
for a suitable regularization functional R_α.

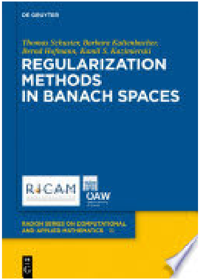
Ill-posedness in inherited by linearizations (cf. [Colton & Kress, 2013,Theorem 4.21].

Important issues not mentioned:

- Optimality of regularization schemes, construction of such schemes.
- Extension to Banach space setting.
- Determining the best value of regularization parameter (discrepancy principle, cross-validation, L-curve, . . .)
- Computational methods, large-scale implementations.







Accounting for Noise Distribution

Example 1

Consider forward map $\boldsymbol{F} : \mathbb{R}^+ \to \mathbb{R}^n$

$$x \mapsto \boldsymbol{F}(x) = \sqrt{x} \begin{bmatrix} 1 \\ 1/2 \\ \vdots \\ 1/n \end{bmatrix} =: \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{bmatrix}$$

and data $\boldsymbol{y} = \boldsymbol{F}(x) + \boldsymbol{\epsilon}$ where

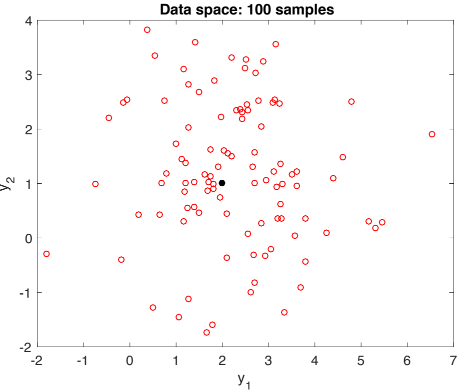
$$y_j = f_j(x) + \epsilon_j, \qquad j = 1, \dots, n, \quad \epsilon_j \sim N(0,1) \text{ i.i.d.}$$

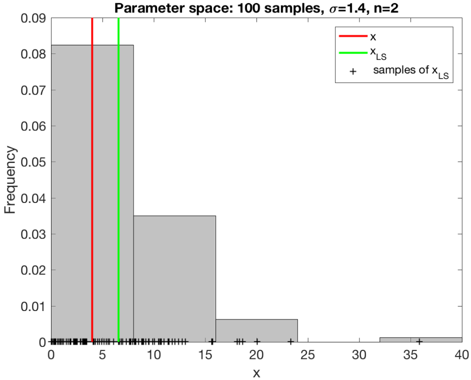
Least squares estimate

$$\hat{x}_{LS} = \operatorname*{arg\,min}_{x>0} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{F}(x)\|_2^2.$$

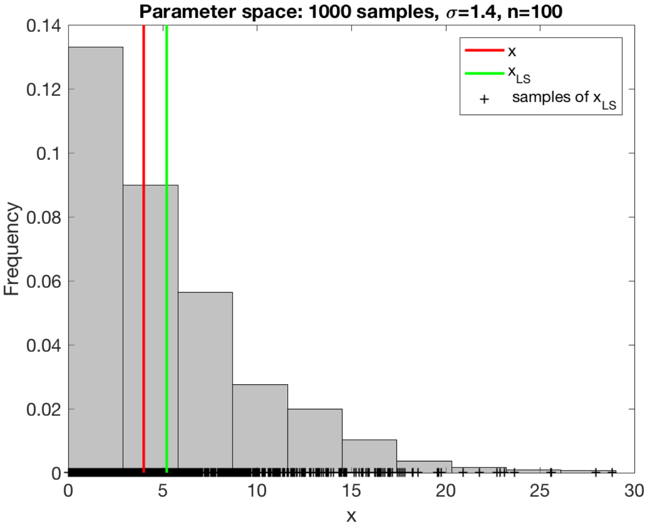
Accounting for Noise Distribution

Example 1



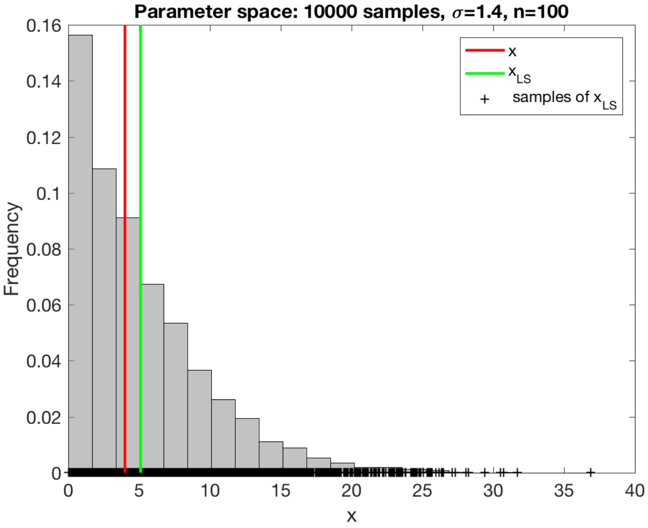


Accounting for Noise Distribution
Example 1



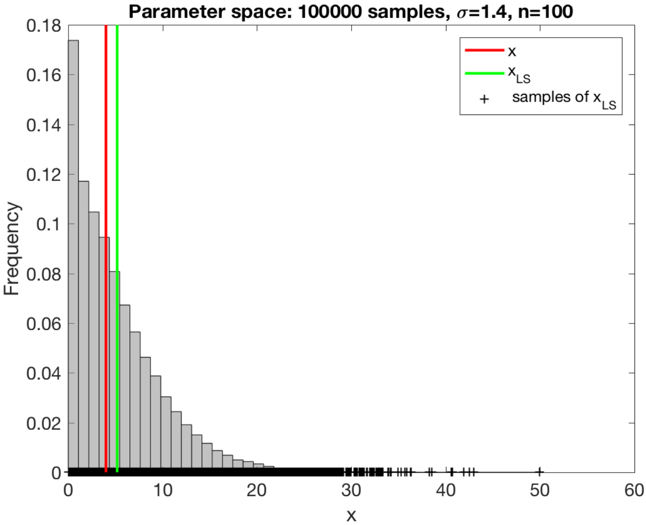
Accounting for Noise Distribution

Example 1



Accounting for Noise Distribution

Example 1



Accounting for Noise Distribution

Example 1

Observations:

- The LS estimate is biased.
- First-order optimality condition for (nonlinear) LS estimation

$$\sum_{j=1}^{n} f_j(\hat{x}) f_j'(\hat{x}) = \sum_{j=1}^{n} y_j f_j'(\hat{x})$$

  can be solved in this example to obtain

$$\hat{x}_{LS} = \left[\frac{\sum_{j=1}^{n} y_j / j}{\sum_{j=1}^{n} 1/j^2} \right]^2 .$$

- From the known Gaussian noise distribution we conclude

$$\mathbf{E}\left[\hat{x}_{LS}\right] = x + \sigma^2 / S_n, \qquad S_n := \sum_{j=1}^{n} 1/j^2 \overset{n\to\infty}{\longrightarrow} \frac{6}{\pi^2}.$$

Accounting for Noise Distribution

Example 1

- If we wish to estimate $z := \sqrt{x}$ instead of x, the optimality conditions yield

$$\hat{z}_{LS} = \frac{\sum_{j=1}^{n} y_j/j}{\sum_{j=1}^{n} 1/j^2}, \qquad \text{giving} \qquad \mathbf{E}\left[\hat{z}_{LS}\right] = \sqrt{x} = z,$$

  an unbiased estimate.
- One can remove the bias by subtracting σ^2/S_n, but this is no longer a LS-fit to the data.
- Note that an unbiased estimate of x is not obtained as the square of the unbiased estimate for $z = \sqrt{x}$. (Note: think of a sine wave signal varying aroind zero with zero mean noise. Best constant estimate of signal is zero, but power of signal is not zero.)
- Statistics: "Conditioning on estimates gives poor predictive distributions".

Accounting for Noise Distribution

Example 1

Does regularization help? Tikhonov would yield

$$\hat{x}_\alpha = \arg\min_{x>0} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{F}(x)\|_2^2 + \frac{\alpha}{2} x^2 \right)$$

- With regularization both estimates (for x and $\sqrt{x}$) are biased.
- The bias depends on the unknown value x.

Accounting for Noise Distribution

Example 1

Does regularization help? Tikhonov would yield

$$\hat{x}_\alpha = \arg\min_{x>0} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{F}(x)\|_2^2 + \frac{\alpha}{2} x^2 \right)$$

- With regularization both estimates (for x and $\sqrt{x}$) are biased.
- The bias depends on the unknown value x.

Take-away:

- LS estimates may not be quantitatively accurate.
- Regularization makes it harder to fix the bias.
- Best estimates do not map correctly through functions, but one can map distributions over possible values correctly.

Accounting for Noise Distribution

Example 2

Consider N measurements of a scalar μ with uniform noise

$$y_j = \mu + \epsilon_j, \qquad \epsilon_j \sim \mathrm{U}[-1,1], \qquad j = 1,\dots,N.$$

- Since $\mu - 1 \le y_j \le \mu + 1$ for all j, we have

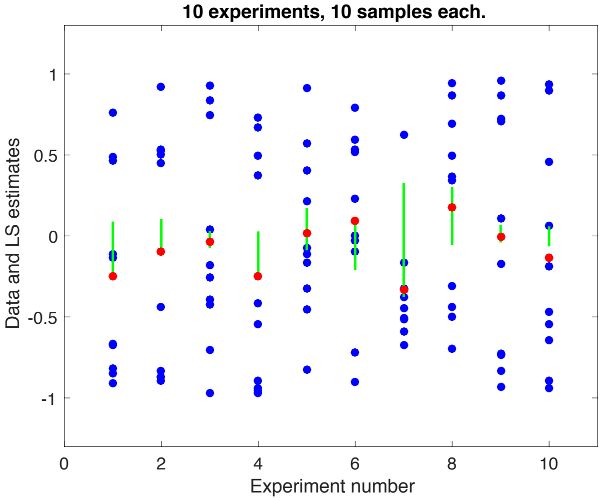
$$\max\{y_j\} - 1 \le \mu \le \min\{y_j\} + 1.$$

- The LS-estimate for μ from N measurements is

$$\hat{\mu}_{LS} = \frac{1}{N}\sum_{j=1}^{N} y_j, \qquad \text{with mean-square error } \frac{1}{\sqrt{3N}}.$$

- The MSE is the variance of the estimate, often quoted as the "error in the estimate".

10 experiments of 10 samples each, red dots: $\hat{\mu}_{LS}$, blue dots the data y_j, green line: feasible region for μ above:



Accounting for Noise Distribution

Observations:

- In 3 out of 10 experiments the LS estimate is not even feasible. (In Bayesian terms: outside the posterior distribution).
- The estimation variance $1/\sqrt{3N} \approx \pm 0.1826$ is sometimes larger, sometimes smalles thatn the actual error in $\hat{\mu}_{LS}$.
- The size of the feasible interval depends on the data, so is not a fixed value.
- An the number of measurements N increases, so does the chance that $\hat{\mu}_{LS}$ is infeasible. At the same time, the estimation error decreases.

We are thus more certain of an estimate that is more likely to be wrong.

- Solution of inverse problems sensitive to noise.
- Noise always present.
- Regularization formulated as optimization problem.
- Regularization replaces solution operator by nearby continuous operator controlled by regularization parameter.
- Goal: convergence to solution for vanishing noise.
- Issue: for finite noise, estimate provided by regularization methods does not convey the variation (distribution) of solution given noise structure.

Contents

1. Inverse Problems
2. Bayesian Inference
3. Sampling from the Posterior
4. An Inverse Problem for Groundwater Flow at WIPP

Contents

2 Bayesian Inference
2.1 Bayes' Rule
2.2 Estimating Probabilities with Bayes' Rule
2.3 A Measurement Model

Bayesian Inference

When the facts change I change my mind.
What do you do, sir.

(attributed to) J. M. Keynes, 1940

Statistical inference about a quantity of interest is described as the modification of the uncertainty about its value in the light of evidence, and Bayes' theorem precisely specifies how this modification should be made.

José M. Bernardo, Bayesian Statistics, 2003

Given probability space $(\Omega, \mathfrak{A}, \mathbf{P})$, $A, B \in \mathfrak{A}$, $\mathbf{P}(B) > 0$, then the conditional probability of A given B is defined by

$$\mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

Special cases:

(i) Mutually exclusive events

$$A \cap B = \emptyset \Rightarrow \mathbf{P}(A|B) = 0.$$

(ii) "B implies A"

$$B \subset A \Rightarrow \mathbf{P}(A|B) = 1.$$

(iii) A, B independent

$$\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B) \Rightarrow \mathbf{P}(A|B) = \mathbf{P}(A).$$

Solving for $\mathbf{P}(A \cap B)$, exchanging roles of A and B, assuming $\mathbf{P}(A) > 0$, gives

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\,\mathbf{P}(A)}{\mathbf{P}(B)}$$ Bayes' rule [Bayes, 1763]

Solving for $\mathbf{P}(A \cap B)$, exchanging roles of A and B, assuming $\mathbf{P}(A) > 0$, gives

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)\,\mathbf{P}(A)}{\mathbf{P}(B)}$$ Bayes' rule [Bayes, 1763]

Interpretations:

- A: unobservable state of nature, with prior probability $\mathbf{P}(A)$ of occurring;
- B: observable event, probability $\mathbf{P}(B)$ known as evidence;
- $\mathbf{P}(B|A)$: probability that A causes B to occur (likelihood);
- $\mathbf{P}(A|B)$: posterior probability of A knowing that B has occurred.
- Terms: inverse probability, Bayesian inference.

Given partition $\{A_j\}_{j\in\mathbb{N}}$ of Ω into exhaustive and exclusive disjoint events, de Morgan's rule and countable additivity give, assuming all $\mathbf{P}(A_j) > 0$,

$$\mathbf{P}(B) = \sum_{j\in\mathbb{N}} \mathbf{P}(B|A_j)\,\mathbf{P}(A_j) \qquad \text{(law of total probability)},$$

leading to another variant of Bayes' rule:

$$\mathbf{P}(A_k|B) = \frac{\mathbf{P}(B|A_k)\,\mathbf{P}(A_k)}{\sum_{j\in\mathbb{N}} \mathbf{P}(B|A_j)\,\mathbf{P}(A_j)},$$

giving posterior probability of each A_k after observing B.

- Incidence of disease among general population: 0.01 %
- Test has true positive rate (sensitivity) of 99.9 %.
- Same test has true negative rate (specificity) of 99.99 %.
- What is the chance that someone who tests positive actually has the disease?

- Incidence of disease among general population: 0.01 %
- Test has true positive rate (sensitivity) of 99.9 %.
- Same test has true negative rate (specificity) of 99.99 %.
- What is the chance that someone who tests positive actually has the disease?

Answer (Bayes' formula, total probability)

$$\mathbf{P}(\text{desease}|\text{pos}) = \frac{\mathbf{P}(\text{pos}|\text{disease}) \cdot \mathbf{P}(\text{disease})}{\mathbf{P}(\text{pos})}$$

where

$$\mathbf{P}(\text{pos}) = \mathbf{P}(\text{pos}|\text{disease}) \cdot \mathbf{P}(\text{disease}) + \mathbf{P}(\text{pos}|\text{no disease}) \cdot \mathbf{P}(\text{no disease})$$

giving

$$\mathbf{P}(\text{desease}|\text{pos}) = \frac{0.999 \cdot 0.0001}{0.999 \cdot 0.0001 + (1 - 0.9999) \cdot (1 - 0.0001)} \approx 0.4998$$

In [Gigerenzer, 1996]: Medical practitioners were given the following information regarding mammography screenings for breast cancer:

incidence: 1 %; sensitivity: 80 %; specificity: 90 %.

When asked to quantify the probability of the patient actually having breast cancer given a positive screening result ($\approx$ 7.5%), 95 out of 100 physicians estimated this probability to lie above 75%.
See also [Gigerenzer et al., 1998] for similar observations in AIDS counseling.

In [Gigerenzer, 1996]: Medical practitioners were given the following information regarding mammography screenings for breast cancer:

incidence: 1 %; sensitivity: 80 %; specificity: 90 %.

When asked to quantify the probability of the patient actually having breast cancer given a positive screening result ($\approx$ 7.5%), 95 out of 100 physicians estimated this probability to lie above 75%.
See also [Gigerenzer et al., 1998] for similar observations in AIDS counseling.

Alternative phrasing (of same answer using natural frequencies)

- Think of random sample 10,000 people.
- Of these, on average 1 will have the disease, 9,999 will not.
- The person who has the disease will almost certainly test positive.
- of the 9,999 healthy people, on average one will test (falsely) positive.
- Thus, roughly one out of every two positive patients actually has the disease.

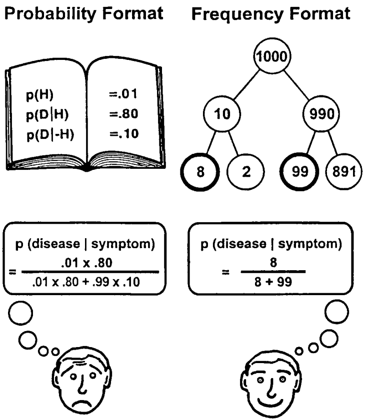


FIGURE 1. Bayesian computations are simpler when information is represented in a frequency format (right) than when it is represented in a probability format (left). p(H) = prior probability of hypothesis. H (breast cancer), p(D|H) = probability of data D (positive test) given H, and p(D|−H) = probability of D given −H (no breast cancer).

Sometimes the description of uncertainty is crucial for its transparent communication.

Source: **Gigerenzer, 1996**

Given two random variables (RV) X, Y, i.e., measurable functions

$$X, Y : \Omega \to \mathbb{R}$$

with probability density functions (pdf)

$$\mathbf{P}(X \le x) = \int_{-\infty}^{x} f_X(\xi)\,\mathrm{d}\xi, \quad \mathbf{P}(Y \le y) = \int_{-\infty}^{y} f_Y(\eta)\,\mathrm{d}\eta,$$

and joint pdf $f(x,y) = f_{X,Y}(x,y)$ (assumed to exist), then

$$\mathbf{P}(X \le x, Y \le y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(\xi,\eta)\,\mathrm{d}\eta\,\mathrm{d}\xi.$$

Conditional density of X given Y (given $Y = y$):

$$f_{X|Y}(x|y) = \frac{f(x,y)}{\int_{-\infty}^{\infty} f(\xi,y)\,\mathrm{d}\xi} = \frac{f(x,y)}{f_Y(y)}.$$

Interpretation:

- Joint density:

$$\mathbf{P}(X = x, Y = y) \mathrel{\hat{=}} f(x, y)\, \mathrm{d}(x, y).$$

- Marginal density:

$$f_Y(y) = \int_{-\infty}^{\infty} f(\xi, y)\, \mathrm{d}\xi, \qquad \mathbf{P}(Y = y) \mathrel{\hat{=}} f_Y(y)\, \mathrm{d}y.$$

- Conditional density:

$$f_{X|Y}(x|y) \mathrel{\hat{=}} \frac{\mathbf{P}(X = x, Y = y)}{\mathbf{P}(Y = y)\, \mathrm{d}y} \mathrel{\hat{=}} \mathbf{P}(X = x | Y = y)\, \mathrm{d}y.$$

Then Bayes' theorem states that

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)\, f_X(x)}{\int f_{Y|X}(y|x)\, f_X(x)\, \mathrm{d}x}.$$

- $f_{Y|X}(y|x)$ is now called the likelihood function.
- $\int f_{Y|X}(y|x)\, f_X(x)\, \mathrm{d}x$ is calles the normalizing factor or marginal.
- Short form:

$$f_{X|Y} \propto f_{Y|X}\, f_X.$$

Contents

2 Bayesian Inference
2.1 Bayes' Rule
2.2 Estimating Probabilities with Bayes' Rule
2.3 A Measurement Model

Problem:(cf. [Gorroochurn, 2012, Chapter 14])

- Given $A \in \mathfrak{A}$, suppose $p := \mathbf{P}(A) \in [0, 1]$ is unknown.
- Assume A has occurred in k out of n independent and identical trials.
- For $0 \leq p_1 < p_2 \leq 1$, what is the probability that $p \in (p_1, p_2)$?

Problem:(cf. [Gorroochurn, 2012, Chapter 14])

- Given $A \in \mathfrak{A}$, suppose $p := \mathbf{P}(A) \in [0,1]$ is unknown.
- Assume A has occurred in k out of n independent and identical trials.
- For $0 \le p_1 < p_2 \le 1$, what is the probability that $p \in (p_1, p_2)$?

Solution:

$$\mathbf{P}(p_1 < p < p_2) = \frac{(n+1)!}{k!(n-k)!} \int_{p_1}^{p_2} p^k (1-p)^{n-k} \, \mathrm{d}p.$$

- Classical probability (Bernoulli, Laplace): given probability $p = \mathbf{P}(A)$, how many independent trials n are necessary to be "morally certain" that A occurs $k = pn$ times?
- Inverse probability (Bayes): Given occurrence rates and notion of prior probability for A, what is $\mathbf{P}(A)$?
  In the literature of Bayes' day often called the "probability of causes".

> *The principle of indifference asserts that if there is no known reason for predicating of our subject one rather than another of several alternatives, then relatively to such knowledge the assertions of each of these alternatives have an equal probability.*
>
> *J. M. Keynes, 1921*

- Rule for assigning epistemic probabilities: in the absence of further information, use the uniform distribution.
- Taken as intuitively obvious by [J. Bernoulli, Ars Conjectandi, 1713].
- Later used by [Laplace, 1774] to define classical probability.
- Originally known as *"Principle of insufficient reason"*, (cf. Leibniz' "principle of sufficient reason"[1]), current term due to [Keynes, 1921].
- Can lead to contradictions (numerous paradoxes in literature).

[1] "For every fact F, there must be an explanation why F is the case."

Problem: A box contains a large number N of black and white balls. We draw n balls with replacement, of which k turn out to be black, $n-k$ white. What is the conditional probability that the next draw will yield a black ball?

Problem: A box contains a large number N of black and white balls. We draw n balls with replacement, of which k turn out to be black, $n-k$ white. What is the conditional probability that the next draw will yield a black ball?

Solution:

$$\mathbf{P}(\text{next draw black}|k \text{ of } n \text{ previous draws black}) = \frac{k+1}{n+2}.$$

Problem: A box contains a large number N of black and white balls. We draw n balls with replacement, of which k turn out to be black, $n-k$ white. What is the conditional probability that the next draw will yield a black ball?

Solution:

$$\mathbf{P}(\text{next draw black}|k \text{ of } n \text{ previous draws black}) = \frac{k+1}{n+2}.$$

Problem: [Laplace, 1814] What is the probability that the sun will rise tomorrow, given that it has risen on each day of the past 5000 years?

Bayesian Inference

Laplace's rule of succession

Problem: A box contains a large number N of black and white balls. We draw n balls with replacement, of which k turn out to be black, $n-k$ white. What is the conditional probability that the next draw will yield a black ball?

Solution:

$$\mathbf{P}(\text{next draw black}|k \text{ of } n \text{ previous draws black}) = \frac{k+1}{n+2}.$$

Problem: [Laplace, 1814] What is the probability that the sun will rise tomorrow, given that it has risen on each day of the past 5000 years?

Solution: $n = 5000 \cdot 365.2426 = 1,826,213$, $k = n$,

$$\mathbf{P}(\text{sunrise tomorrow}) = \frac{1,826,214}{1,826,215} \approx 0.9999995.$$

> *But this number is incomparably greater for him who, recognizing in the totality of phenomena the principal regulator of days and seasons, sees that nothing at the present moment can arrest the course of it.*
>
> *P.-S. Laplace, Essai Philosophique sur les Probabilités,1814*

[Taleb & Blythe, 2011], [B. Russell, 1912], [Gigerenzer, 2014]

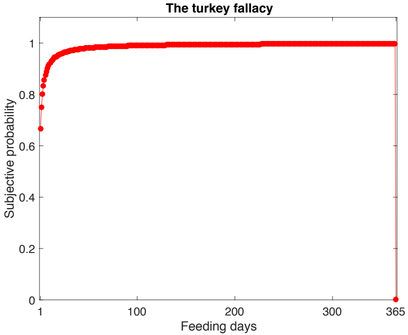
- A turkey is fed by the farmer every day for many months.
- The turkey applies Laplace's rule of succession and feels more confident with every passing day.

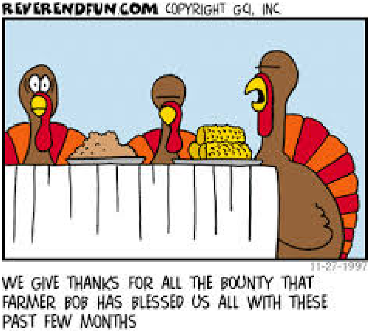
Bayesian Inference

Model uncertainty or the turkey fallacy

[Taleb & Blythe, 2011], [B. Russell, 1912], [Gigerenzer, 2014]

- A turkey is fed by the farmer every day for many months.
- The turkey applies Laplace's rule of succession and feels more confident with every passing day.
- . . . until Thanksgiving.





- The turkey had too much confidence in his model of uncertainty; he was missing important information (unknown unknowns).
- Fundamental question in epistemology (the theory of knowledge), known as the Problem of Induction [D. Hume, 1748]
- [K. Popper, 1959] postulated that induction is not possible, that scientific theories can only be falsified.
- The turkey illusion is the belief that a risk can be calculated when it cannot.
- [F. Knight, 1921]: distinction between known risk ("risk") and unknown risk ("uncertainty"). Uncertainty in this sense requires more tools than probability.

Contents

2 Bayesian Inference
2.1 Bayes' Rule
2.2 Estimating Probabilities with Bayes' Rule
2.3 A Measurement Model

Assume we have performed N measurements $\boldsymbol{y} = (y_1, \dots, y_N)$ of the length ℓ of a rod with error model

$$Y = \ell + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2), \qquad \text{i.e.} \qquad Y \sim \mathrm{N}(\ell, \sigma^2).$$

What is our state of knowledge about ℓ?

Assume we have performed N measurements $\boldsymbol{y} = (y_1, \dots, y_N)$ of the length ℓ of a rod with error model

$$Y = \ell + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2), \qquad \text{i.e.} \qquad Y \sim \mathrm{N}(\ell, \sigma^2).$$

What is our state of knowledge about ℓ?

For independent measurements the likelihood function is

$$f(\boldsymbol{y}|\ell) = \frac{1}{(\sigma\sqrt{2\pi})^N} \exp\left(-\sum_{j=1}^{N} \frac{(y_j - \ell)^2}{2\sigma^2}\right)$$

Assume we have performed N measurements $\boldsymbol{y} = (y_1, \dots, y_N)$ of the length ℓ of a rod with error model

$$Y = \ell + \epsilon, \quad \epsilon \sim \mathrm{N}(0, \sigma^2), \qquad \text{i.e.} \qquad Y \sim \mathrm{N}(\ell, \sigma^2).$$

What is our state of knowledge about ℓ?

For independent measurements the likelihood function is

$$f(\boldsymbol{y}|\ell) = \frac{1}{(\sigma\sqrt{2\pi})^N} \exp\left(-\sum_{j=1}^{N} \frac{(y_j - \ell)^2}{2\sigma^2} \right)$$

The posterior for ℓ starting with a prior density f is then

$$f(\ell|\boldsymbol{y}) = \frac{K}{(\sigma\sqrt{2\pi})^N} \exp\left(-\sum_{j=1}^{N} \frac{(y_j - \ell)^2}{2\sigma^2} \right) f(\ell)$$

with normalization constant K.

Fox et al., Lecture Notes, 2016

Rearranging terms, factoring out quantities not dependent on ℓ, gives

$$f(\ell|\boldsymbol{y}) \propto \exp\left[-\frac{N}{2\sigma^2}\exp(\ell-\overline{y})^2\right] f(\ell), \qquad \overline{y} := \frac{1}{N}\sum_{j=1}^{N} y_j.$$

- Effect of data collection: multiplication of prior $f(\ell)$ by Gaussian of mean $\overline{y}$ and variance σ^2/N.
- If prior $f(\ell)$ approximately uniform (flat) around $\overline{y}$, then posterior almost completely determined by data.
- For Gaussian posterior mean, median and mode all coincide, so that natural best estimate is

$$\hat{\ell} := \overline{y} \quad \text{with uncertainty measure} \quad \frac{\sigma}{N}.$$

Now assume measurement error variance σ unknown; include in Bayesian formulation by considering σ as new parameter (likelihood function same as before):

$$
\begin{aligned}
f(\ell,\sigma|\boldsymbol{y}) &= K\, f(\boldsymbol{y}|\ell,\sigma)\, f(\ell,\sigma) \\
&= \frac{K}{(\sigma\sqrt{2\pi})^N} \exp\left(-\frac{N}{2\sigma^2}\left[(x-\overline{y})^2 + s^2\right]\right) f(\ell,\sigma)
\end{aligned}
$$

with K normalization constant, $s := 1/N(\sum_{j=1}^N y_j^2) - \overline{y}^2$.

Now assume measurement error variance σ unknown; include in Bayesian formulation by considering σ as new parameter (likelihood function same as before):

$$\begin{aligned} f(\ell,\sigma|\boldsymbol{y}) &= K\, f(\boldsymbol{y}|\ell,\sigma)\, f(\ell,\sigma) \\ &= \frac{K}{(\sigma\sqrt{2\pi})^N} \exp\left(-\frac{N}{2\sigma^2}\left[(x-\overline{y})^2+s^2\right]\right) f(\ell,\sigma) \end{aligned}$$

with K normalization constant, $s := 1/N(\sum_{j=1}^N y_j^2) - \overline{y}^2$.

For flat prior normalized posterior density given by

$$f(\ell,\sigma|\boldsymbol{y}) = \sqrt{\frac{8}{N\pi}} \left(\frac{Ns^2}{2}\right)^{N/2} \frac{1}{s^2\sigma^N\Gamma(N/2-1)} \exp\left(-\frac{N}{2\sigma^2}\left[(\ell-\overline{y})^2+s^2)\right]\right)$$

Now assume measurement error variance σ unknown; include in Bayesian formulation by considering σ as new parameter (likelihood function same as before):

$$\begin{aligned} f(\ell,\sigma|\boldsymbol{y}) &= K\, f(\boldsymbol{y}|\ell,\sigma)\, f(\ell,\sigma) \\ &= \frac{K}{(\sigma\sqrt{2\pi})^N} \exp\left(-\frac{N}{2\sigma^2}\left[(x-\overline{y})^2+s^2\right]\right) f(\ell,\sigma) \end{aligned}$$

with K normalization constant, $s := 1/N(\sum_{j=1}^N y_j^2) - \overline{y}^2$.

For flat prior normalized posterior density given by

$$f(\ell,\sigma|\boldsymbol{y}) = \sqrt{\frac{8}{N\pi}}\left(\frac{Ns^2}{2}\right)^{N/2} \frac{1}{s^2\sigma^N\Gamma(N/2-1)} \exp\left(-\frac{N}{2\sigma^2}\left[(\ell-\overline{y})^2+s^2)\right]\right)$$

Peak of this function at

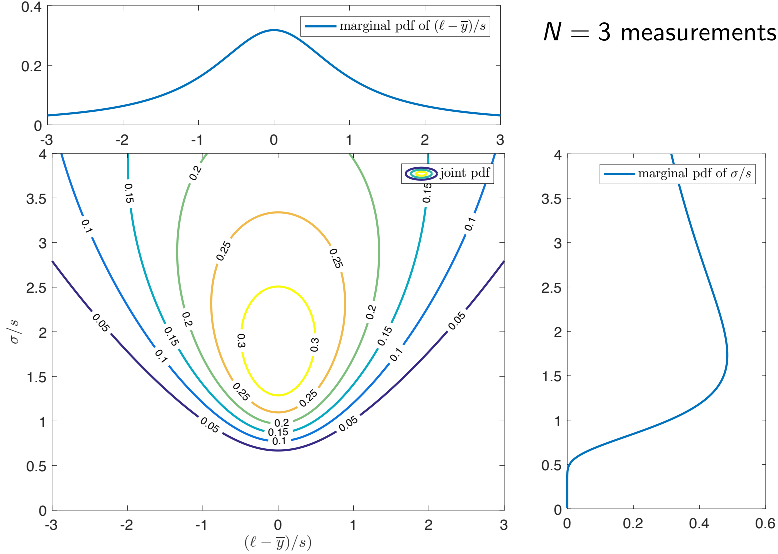
$$\ell_{\mathrm{MAP}} = \overline{y}, \quad \sigma_{\mathrm{MAP}} = s,$$

where MAP stands for maximum a posteriori estimate, which coincides with the maximum likelihood estimate.

We can find the marginal densities of the posterior by integrating out the unconsidered variable:

$$f(\ell|\boldsymbol{y}) = \frac{\Gamma(N/2-1/1)}{\Gamma(N/2-1)} \frac{s^{N-2}}{\sqrt{\pi}\left[(\ell-\overline{y})^2+s^2\right]^{(N-1/2)}}$$

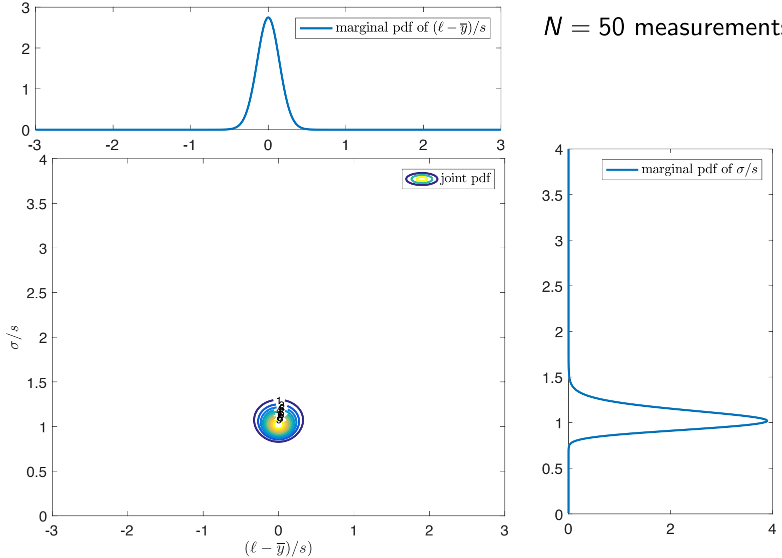
$$f(\sigma|\boldsymbol{y}) = \frac{2}{\Gamma(N/2-1)} \left(\frac{Ns^2}{s}\right)^{N/2-1} \sigma^{1-N} \exp\left(-\frac{Ns^2}{2\sigma^2}\right).$$



$N = 3$ measurements

Bayesian Inference

Example: A measurement model



$N = 50$ measurements

- Bayes theorem as mathematical model of incorporating new observations with current state of knowledge.
- For UQ: method of updating probabilities.
- Describes step from prior probability distribution to posterior probability distribution.
- Can incorporate uncertainty also in statistical parameters ("hyperparameters").

Contents

1. Inverse Problems
2. Bayesian Inference
3. Sampling from the Posterior
4. An Inverse Problem for Groundwater Flow at WIPP

Contents

3 Sampling from the Posterior

3.1 Monte Carlo Integration

3.2 Markov Chains

3.3 Markov Chain Monte Carlo

3.4 Proposal Distributions

Monte Carlo Integration

Given a RV (quantity of interest) X with a known distribution, information on its variability can be obtained from statistical quantities such as

- Expected value

$$\mathbf{E}[X] = \int x\, f_X(x)\, \mathrm{d}x$$

- Higher moments:

$$\mathbf{E}\left[X^k\right], \quad k \in \mathbb{N}.$$

- Cumulative distribution function

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, \mathrm{d}x.$$

- Probability of events

$$\mathbf{P}(x \in A) = \int \chi_A(x)\, f_X(x)\, \mathrm{d}x, \qquad \chi_A(x) = \begin{cases} 1, & x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

Monte Carlo Integration

- Given a device for generating a sequence $\{X_k\}$ of i.i.d. realizations of a given random variable X, basic MC simulation uses the approximation

$$\mathbf{E}[X] \approx \frac{S_N}{N}, \qquad S_N = X_1 + \cdots + X_N.$$

- By the SLLN, $\frac{S_N}{N} \to \mathbf{E}[X]$ a.s.
- Similarly, for a measurable function f, $\mathbf{E}[f(X)] \approx \frac{1}{N}\sum_{k=1}^{N} f(X_k)$.
- For a RV $X \in L^2(\Omega; R)$ the standardized RV

$$X^* := \frac{X - \mathbf{E}[X]}{\sqrt{\mathbf{Var}\, X}} \quad \text{has} \quad \mathbf{E}[X^*] = 0, \ \mathbf{Var}\, X^* = 1.$$

- If $\mu = \mathbf{E}[X]$, $\sigma^2 = \mathbf{Var}\, X$, then $\mathbf{E}[S_N] = N\mu$, $\mathbf{Var}\, S_N = N\sigma^2$ and, by the CLT,

$$S_N^* = \frac{S_N - N\mu}{\sqrt{N}\sigma} \to N(0,1).$$

- Since

$$\mathbf{E}\left[\left(\frac{S_N}{N}-\mu\right)^2\right] = \mathbf{Var}\,\frac{S_N}{N} = \frac{\sigma^2}{N} \to 0 \qquad (N\to\infty)$$

we have L^2-convergence of S_N/N to μ and, for any $\epsilon > 0$,

$$\mathbf{P}\left\{\left|\frac{S_N}{N}-\mu\right| > N^{-1/2+\epsilon}\right\} \le \frac{\sigma^2}{N^{2\epsilon}}, \tag{3.1}$$

i.e., as the number N of samples increases, the probability of the error being larger than $O(N^{-1/2+\epsilon})$ converges to zero for any $\epsilon > 0$.

- If $\rho := \mathbf{E}\left[|X-\mu|^3\right] < \infty$, then the Berry-Esseen bound further gives

$$|\mathbf{P}(S_N^* \le x) - \Phi(x)| \le C\,\frac{\rho}{\sigma^3\sqrt{N}}, \tag{3.2}$$

where Φ denotes the cdf of N$(0,1)$.

- For a RV $Z \sim \mathrm{N}(0,1)$ and $x \in \mathbb{R}$, this implies

$$\mathbf{P}(S_N^* \le x) = \mathbf{P}(Z \le x) + O(N^{-1/2})$$

and therefore

$$\begin{aligned}\mathbf{P}(|S_N^*| \le x) &= \mathbf{P}(S_N^* \le x) - \mathbf{P}(S_N^* < -x) \\ &= \mathbf{P}(Z \le x) - \mathbf{P}(Z < -x) + O(N^{-1/2}) \\ &= \mathbf{P}(|Z| \le x) + O(N^{-1/2}) \\ &= \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + O(N^{-1/2})\end{aligned}$$

where

$$\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) = 2\,\Phi(x) - 1.$$

- If the $O(N^{-1/2})$-term is assumed negligible, this can be used to construct (asymptotic) confidence intervals for S_N^*, i.e., the MC estimate S_N/N.

True confidence intervals are obtained if we carry along the bound in the Berry-Esseen estimate (3.2), denoted by B_N,

$$-B_N \leq \mathbf{P}(S_N^* \leq x) - \Phi(x) \leq B_N$$

i.e., for $R \geq 0$ we have

$$\begin{aligned}\mathbf{P}(|S_N^*| \leq R) &= \mathbf{P}(S_N^* \leq R) - \mathbf{P}(S_N^* < -R)\\ &\geq \Phi(R) - B_N - (\Phi(-R) + B_N)\\ &= \underbrace{\Phi(R) - \Phi(-R)}_{=:\gamma_R} - 2B_N\end{aligned}$$

and, in the same manner, $\mathbf{P}(|S_N^*| \leq R) \leq \gamma_R + 2B_N$, i.e.,

$$\gamma_R - 2\,B_N \leq \mathbf{P}\left(\mu \in \left[\frac{S_N}{N} - \frac{\sigma R}{\sqrt{N}}, \frac{S_N}{N} + \frac{\sigma R}{\sqrt{N}}\right]\right) \leq \gamma_R + 2\,B_N.$$

Contents

3 Sampling from the Posterior
3.1 Monte Carlo Integration
3.2 Markov Chains
3.3 Markov Chain Monte Carlo
3.4 Proposal Distributions

A stochastic process $X = \{X_n(\omega) : n \in \mathbb{N}_0\}$ with values in a set $\mathcal{S}$ is called a Markov chain (MC)[2] if
for all $A \subset \mathcal{S}$, for all $n \in \mathbb{N}_0$ and for all $x_0, x_1, \dots x_n \in \mathcal{S}$, there holds

$$\mathbf{P}(X_{n+1} \in A | X_n = x_n, X_{n-1} = x_{n-1}, \dots X_0 = x_0) = \mathbf{P}(X_{n+1} \in A | X_n = x_n),$$

i.e., the value of the chain is independent of its past history.

A MC X is time-homogeneous if its transition probabilities

$$\mathbf{P}(X_{n+1} \in A | X_n = x) = P(x, A) = \int_A p(x, y)\,\mathrm{d}y$$

do not depend on n. $P(x, A)$ is called the transition kernel which we assume absolutely continuous for every $x \in \mathcal{S}$.
The n-step transition densities are defined as

$$\mathbf{P}(X_{n+1} \in A | X_0 = x) = P^{(n)}(x, A) = \int_A p^{(n)}(x, y)\,\mathrm{d}y.$$

[2]The material on Markov chains and MCMC alorithms closely follows the excellent presentation in the lecture notes by Gareth Roberts at this link.

For a finite state space $|\mathcal{S}| = k < \infty$ the transition matrix $\boldsymbol{P} \in \mathbb{R}^{k \times k}$ is defined by

$$p_{i,j} = \mathbf{P}(X_{n+1} = i | X_n = j), \quad i, j, \in \mathcal{S}.$$

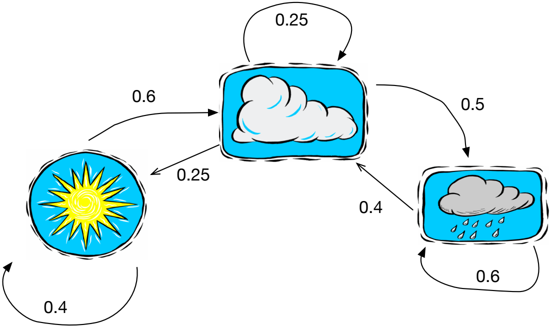
Markov Chains

Terminology

For a finite state space $|\mathcal{S}| = k < \infty$ the transition matrix $\boldsymbol{P} \in \mathbb{R}^{k \times k}$ is defined by

$$p_{i,j} = \mathbf{P}(X_{n+1} = i | X_n = j), \quad i, j, \in \mathcal{S}.$$

Example:



$$\boldsymbol{P} = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0.25 & 025 & 0.5 \\ 0 & 0.4 & 0.6 \end{bmatrix}$$

If the distribution of X_0 is given by the density $q^{(0)}$, then the density of X at time n is

$$q^{(n)}(x) = \int_{\mathcal{S}} q^{(0)}(y)\, p^{(n)}(x,y)\, \mathrm{d}y$$

and, for finite state spaces,

$$\boldsymbol{q}^{(n)} = \boldsymbol{q}^{(0)} \boldsymbol{P}^n.$$

Example: For the weather chain, if $\boldsymbol{q}^{(0)} = [1,0,0]$ (sunny on first day), then

$$\boldsymbol{q}^{(2)} = [1,0,0]\boldsymbol{P}^2 = [1,0,0] \begin{bmatrix} 0.31 & 0.39 & 0.3 \\ 0.1625 & 0.4125 & 0.425 \\ 0.1 & 0.34 & 0.56 \end{bmatrix} = [0.31, 0.1625, 0.1],$$

i.e., on day $n = 2$ there ist a 31 % chance on sunny weather.

Under certain regularity conditions the distribution of a MC converges to a limiting distribution, the stationary, invariant or equilibrium distribution, in which case the chain is said to be ergodic.

A MC is said to be irreducible if all states intercommunicate, i.e., if for all $i, j \in \mathcal{S}$ there is an $n \in \mathbb{N}_0$ such that $\mathbf{P}(X_n = i | X_0 = j) > 0$.

A MC is said to be recurrent if $\mathbf{P}(X_n = i,\ n > 0 | X_0 = i) > 0$ for all $i \in \mathcal{S}$.

A MC is said to be positive recurrent if $\mathbf{E}[T_{i,i}] < \infty$ for all $i \in \mathcal{S}$, where $T_{i,i}$ denotes the time of the first return to state i.
If X is ergodic with invariant distribution π, then $\pi(i) = 1/\mathbf{E}[T_{i,i}]$.

A MC is sait to be aperiodic if there do not exist $d \geq 2$ disjoint subsets $\mathcal{S}_1, \dots \mathcal{S}_d$ such that

$$\begin{aligned} \mathbf{P}(x, \mathcal{S}_{i+1}) &= 1 \qquad \text{for all } x \in \mathcal{S}_i, \quad 1 \leq i \leq d-1, \\ \mathbf{P}(x, \mathcal{S}_1) &= 1 \qquad \text{for all } x \in \mathcal{S}_d. \end{aligned}$$

Remarks:

1. Recurrence and aperiodicity are class properties: If a MC is irreducible, then all of $\mathcal{S}$ is one communicating class. Thus an irreducible chain is recurrent if one of its states is recurrent. The same is true for positive recurrence and aperiodicity.
2. Irreducibility essentially ensures there is no partition of $\mathcal{S}$ into subsets between which the chain cannot move.
3. (Positive) Recurrence ensures that the chain eventually visits every subset of the state space of positive measure (sufficiently often).
4. Perodicity states there exists a partition of $\mathcal{S}$ into subsets which are visited by the chain in cyclical sequential order.

If $\mathcal{X}$ denotes a continuous state space, a MC on $\mathcal{X}$ is called μ-reducible if there exists a measure μ on $\mathcal{X}$ such that for all $A \subset \mathcal{X}$ with $\mu(A) > 0$ and all $x \in \mathcal{X}$ there exists $n \in \mathbb{N}_0$ such that

$$P^n(x, A) > 0.$$

- Setting $\mu(A) = \delta_{x_0}(A)$ requires that state x_0 can be reached from every other state. (Therefore irreducibility ist stronger than μ-irreducibility.)
- Aperiodicity applies also to continuous MC.
- A MC that is μ-irreducible and aperiodic has a limit distribution.

The total variation distance $d_{TV}(P_1, P_2)$ between two probability measures P_1 and P_2 is defined as

$$d_{TV}(P_1, P_2) := \sup_{A \subset \mathcal{X}} |P_1(A) - P_2(A)|.$$

Theorem 3.1 (Limit distribution)

The distribution of an aperiodic μ-irreducible MC converges to a limit distribution π in the sense that

$$\lim_{n\to\infty} d_{TV}(P^n(x, \cdot), \pi(\cdot)) = 0 \quad \text{for } \pi\text{-almost all } x \in \mathcal{X}.$$

A MC is said to be Harris recurrent if for all $B \subset \mathcal{X}$ with $\pi(B) > 0$ and all $x \in \mathcal{X}$ there holds

$$\mathbf{P}(X_n \in B \text{ for some } n \in \mathbb{N} | X_0 = x) = 1.$$

Theorem 3.2

The distribution of an aperiodic Harris recurrent MC converges to a limit distribution π, i.e.,

$$\lim_{n\to\infty} d_{TV}(P^n(x,\cdot),\pi(\cdot)) = 0 \quad \text{for all } x \in \mathcal{X}.$$

- Because
$$q^n(A) := \mathbf{P}(X_n \in A) = \int q^{(0)}(x)P^n(x,A)\,\mathrm{d}x$$
it follows that $\lim_{n\to\infty}\mathbf{P}(X_n \in A) = \pi(A)$ for all $A \subset \mathcal{X}$ and all initial distributions $q^{(0)}$.
- Since Theorem 3.2 holds for any $q^{(0)}$, if we run an ergodic MC for a long time, it will reach a statistical equilibrium, regardless of its starting point.
- If we start a chain in equilibrium then it remains in equilibrium.
- We assume all MC in the following to be ergodic.

- If the chain begins in equilibrium, it stays there.
- This implies (dominated convergence theorem) that

$$\pi(x) = \int_{\mathcal{S}} \pi(y)\, p(y, x)\, \mathrm{d}y \qquad \text{(general balance relation).}$$

Lemma 3.3

A distribution π on $\mathcal{S}$ which satisfies the detailed balance relation

$$\pi(x)p(x, y) = \pi(y)p(y, x) \qquad \forall x, y \in \mathcal{S},$$

where $p(\cdot, \cdot)$ is the density of an ergodic MC X, is the stationary distribution of X.

- Detailed balance is sufficient, but not necessary for general balance.
- If detailed balance holds, then the MC is time-reversible.
  (Not all ergodic MCs are time-reversible.)

Theorem 3.4 (Ergodic theorem)

Let f be a real-valued function, X an ergodic MC with stationary distribution π and Y a RV with pdf π. If $\mathbf{E}_\pi\left[|f(Y)|\right] < \infty$, then

$$\overline{f}_N := \frac{1}{N}\sum_{n=1}^{N} f(X_n) \longrightarrow \mathbf{E}_\pi\left[f(Y)\right] \quad \text{as } N \to \infty \text{ with probability one.}$$

The ergodic theorem is a law of large numbers.
There is also a corresponding central limit theorem for MC, but this requires a stronger convergence (geometric ergodicity).

An ergodic MC with invariant distribution π is said to be geometrically ergodic if there exist $r \in (0,1)$ and a nonnegative function M on $\mathcal{S}$ with $\mathbf{E}_\pi[M(X)] < \infty$ such that

$$d_{TV}(P^n(x,\cdot),\pi(\cdot)) \le M(x)\, r^n \qquad \forall x \in \mathcal{S}, \forall n \in \mathbb{N}.$$

If M is bounded above then the chain is called uniformly ergodic.

Theorem 3.5 (Central limit theorem)

If X is geometrically ergodic and f such that $\mathbf{E}_\pi\left[f(Y)^{2+\epsilon}\right] < \infty$ for some $\epsilon > 0$ then

$$\frac{1}{N}\sum_{n=1}^{N} f(X_n) \xrightarrow{\text{dist}} \mathrm{N}\left(\mathbf{E}_\pi[f(Y)], \frac{\tau^2}{N}\right)$$

- Convergence in distribution.
- τ related to integrated autocorrelation time of X.

Contents

3 Sampling from the Posterior
3.1 Monte Carlo Integration
3.2 Markov Chains
3.3 Markov Chain Monte Carlo
3.4 Proposal Distributions

- Inspired by ergodic theorem: to approximate $\mathbf{E}[f(Y)]$ for function f and RV Y with pdf π.
- However, we cannot compute

$$\mathbf{E}_\pi[f(Y)] = \int f(y)\,\pi(y)\,\mathrm{d}y$$

directly, e.g. because cannot sample directly from π.
- Instead, Markov chain Monte Carlo (MCMC) methods construct an ergodic Markov chain with π as its limiting distribution and, having generated N samples by running the chain, compute approximation

$$\mathbf{E}_\pi[f(Y)] \approx \frac{1}{N}\sum_{n=1}^{N} f(X_n).$$

Ergodic theorem assures convergence as $N \to \infty$.

Family of methods for achieving this: "samplers".
(Many chains with invariant distribution π)

One of the most popular is the Metropolis-Hastings sampler [Rosenbluth & al., 1953], [Hastings, 1970]:

- $\forall\, x \in \mathcal{S}$: choose density $q(x,\cdot)$ specifying transition probability from x to another state in $\mathcal{S}$. (Should be easy to sample from.)
- $X_n = x$: Sample possible new state z according to $q(x,\cdot)$ (proposal).
- Accept proposal with acceptance probability

$$\alpha(x,z) = \min\left\{1, \frac{\pi(z)q(z,x)}{\pi(x)q(x,z)}\right\}.$$

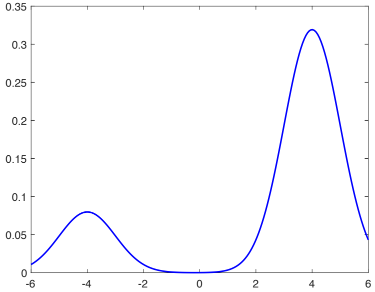
- If proposal accepted, set $X_{n+1} = z$, otherwise $X_{n+1} = x$.

Example: Bimodal normal mixture distribution

- Target density: (could also sample directly)

$$\pi(x) = \frac{p}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + \frac{1-p}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

where $0 < p < 1$.



$\mu_1 = -4, \mu_2 = 4,$
$\sigma_1 = \sigma_1 = 1,$
$p = 0.8.$

- Proposal density: sample w from $\mathrm{N}(0,1)$ and propose $z = x + w$, i.e., $z \sim \mathrm{N}(x,1)$, giving proposal density

$$q(x,z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-x)^2}{2}\right).$$

- Acceptance probability:

$$\begin{aligned}
\alpha(x,z) &= \min\left\{1, \frac{\pi(x)q(z,x)}{\pi(x)q(x,z)}\right\} \\
&= \min\left\{1, \frac{\pi(z)\frac{1}{\sqrt{2\pi}}\exp((x-z)^2/2)}{\pi(x)\frac{1}{\sqrt{2\pi}}\exp((x-z)^2/2)}\right\} \\
&= \min\left\{1, \frac{\pi(z)}{\pi(x)}\right\}.
\end{aligned}$$

- Much freedom in choosing proposal mechanism. Natural requirement:

$$\mathcal{S} = \operatorname{supp} \pi \subset \bigcup_{x \in \mathcal{S}} \operatorname{supp} q(x, \cdot).$$

- Note: acceptance probability depends on ratios of π, hence knowledge of normalizing constant not required.
- Acceptance probability terms chosen in order that detailed balance holds.
- Acceptance/rejection step needed to "steer" limit distribution of MC to target ("Metropolization")

Lemma 3.6

The transition kernel of the MH sampler is given by

$$p(x, y) = q(x, y)\alpha(x, y) + \mathbb{1}_{\{x=y\}} r(x)$$

where

$$r(x) = \begin{cases} \sum_{y \in \mathcal{S}} q(x, y) \left(1 - \alpha(x, y)\right), & \mathcal{S} \text{ discrete}, \\ \int_{\mathcal{S}} \left(1 - \alpha(x, y)\right) \, \mathrm{d}y, & \mathcal{S} \text{ continuous}. \end{cases}$$

Lemma 3.7

The MH chain satisfies the detailed balance relation with respect to π.

Proof: For $x \neq y$ we obtain

$$
\begin{aligned}
\pi(x)p(x,y) &= \pi(x)q(x,y)\alpha(x,y) \\
&= \min\left\{\pi(x)q(x,y), \pi(y)q(y,x)\right\} \\
&= \pi(y)q(y,x)\min\left\{\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)}, 1\right\} \\
&= \pi(y)p(y,x).
\end{aligned}
$$

Contents

3 Sampling from the Posterior
3.1 Monte Carlo Integration
3.2 Markov Chains
3.3 Markov Chain Monte Carlo
3.4 Proposal Distributions

Proposal Distributions

Gibbs sampler

To generate proposals from s d-variate distribution, the Gibbs sampler proceeds component by component:

- at state $\boldsymbol{x} = (x_1, \dots, x_d)$, denote

$$\boldsymbol{x}_{-i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots x_d), \qquad 1 \leq i \leq d.$$

- Choose component $i \in \{1, \dots, d\}$, propose new state

$$z = (x_1, \dots, x_{i-1}, y, x_{i+1}, \dots, x_d).$$

  with y sampled from full conditional density

$$\pi(y|x_{-i}) = \frac{\pi(z)}{\int \pi(x_1, \dots, x_{i-1}, w, x_{i+1}, \dots, x_d) \, \mathrm{d}w}.$$

- Acceptance probability equal to one. If full conditionals standard distributions they are easily sampled.

Proposal Distributions

Independence sampler

The independence sampler proposes states which are independent of the currenst state of the MC.

- For a fixed density f, proposal is

$$q(x, y) = f(y) \quad \forall x \in \mathcal{S}.$$

- Acceptance probability

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)f(x)}{\pi(x)f(y)}\right\}.$$

- Well understood, but often slow.
- Ergodic as long as $\operatorname{supp} \pi \subset \operatorname{supp} f$.

Metropolis et al. originally proosed to use symmetric proposal densities, i.e.,

$$q(x, y) = q(y, x).$$

The acceptance probability then simplifies to

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

Proposal Distributions

Random walk Metropolis-Hastings sampler

The random walk Metropolis-Hastings sampler makes proposals y according to a random walk

$$y = x + z$$

where z is drawn from a proposal density f.

- Proposal density: $q(x, y) = f(y - x)$.
- Acceptance probability:

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)f(x-y)}{\pi(x)f(y-x)}\right\}.$$

- Reduces to Metropolis sampler when density f symmetric about origin.

Sampling from the Posterior

Summary

- Posterior distribution of Bayesisn inference generally inaccessible.
- MCMC is a way of drawing samples from the posterior distribution.
- This suffices for computing statistical measures of QoI.
- MH sampler flexible approach for generating MC with given limit distribution.
- Samples correlated.
- Convergence needs to be assured.
- Note: for PDE forward models each MCMC step requires a PDE solve.

Contents

1. Inverse Problems
2. Bayesian Inference
3. Sampling from the Posterior
4. An Inverse Problem for Groundwater Flow at WIPP

Contents

4 An Inverse Problem for Groundwater Flow at WIPP
4.1 Gaussian Random Fields
4.2 Gaussian Random Field Models from Direct Observations
4.3 Solution of the Forward Problem
4.4 Bayesian Inversion
4.5 Improving MCMC Proposals in Hilbert Space
4.6 Numerical Results: Bayesian Inversion for WIPP

Gaussian Random Fields

Random field $\kappa : D \times \Omega \to \mathbb{R}$ is Gaussian iff for $n \in \mathbb{N}$ and $\boldsymbol{x}_i \in D$, $i = 1, \dots, n$,

$$(\kappa(\boldsymbol{x}_1), \dots, \kappa(\boldsymbol{x}_n)) \sim N(\boldsymbol{m}, \boldsymbol{C}), \qquad \boldsymbol{m} \in \mathbb{R}^n, \boldsymbol{C} \in \mathbb{R}^{n \times n}.$$

A GRF κ is determined by its mean and covariance function

$$m(\boldsymbol{x}) = \mathbf{E}\left[\kappa(\boldsymbol{x})\right], \qquad c(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{Cov}(\kappa(\boldsymbol{x}), \kappa(\boldsymbol{y})).$$

Gaussian Random Fields

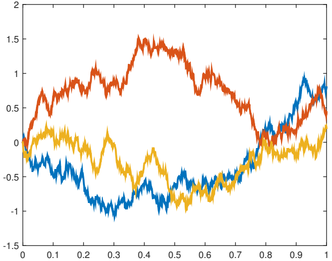
Random field $\kappa : D \times \Omega \to \mathbb{R}$ is Gaussian iff for $n \in \mathbb{N}$ and $\boldsymbol{x}_i \in D$, $i = 1, \dots, n$,

$$(\kappa(\boldsymbol{x}_1), \dots, \kappa(\boldsymbol{x}_n)) \sim N(\boldsymbol{m}, \boldsymbol{C}), \qquad \boldsymbol{m} \in \mathbb{R}^n, \boldsymbol{C} \in \mathbb{R}^{n \times n}.$$

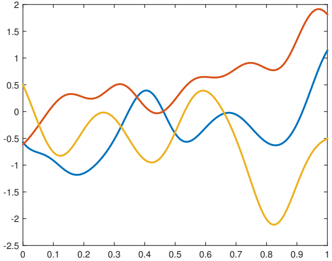
A GRF κ is determined by its mean and covariance function

$$m(\boldsymbol{x}) = \mathbf{E}\left[\kappa(\boldsymbol{x})\right], \qquad c(\boldsymbol{x}, \boldsymbol{y}) = \mathbf{Cov}(\kappa(\boldsymbol{x}), \kappa(\boldsymbol{y})).$$

Depending on m and c, realizations $\kappa(\cdot, \omega)$ are **P**-a.s. continuous (or smoother).



$c(x, y) = \min(x, y)$



$c(x, y) = \exp(-100\,|x - y|_2^2)$

Let κ be a Gaussian random field with a.s. continuous paths.

Can view κ also as random variable $\kappa : \Omega \to C(D)$ with values in $C(D)$.

Since $C(D) \hookrightarrow L^2(D)$ for bounded $D \subset \mathbb{R}^d$, we take $\kappa : \Omega \to L^2(D)$ as Hilbert space-valued random variable.

Let κ be a Gaussian random field with a.s. continuous paths.

Can view κ also as random variable $\kappa : \Omega \to C(D)$ with values in $C(D)$.

Since $C(D) \hookrightarrow L^2(D)$ for bounded $D \subset \mathbb{R}^d$, we take $\kappa : \Omega \to L^2(D)$ as Hilbert space-valued random variable.

For separable Hilbert spaces $\mathcal{H}$ we can (analogously to $\mathbb{R}^n$) define

- Lebesgue spaces $L^2(\Omega; \mathcal{H})$,
- expectations $\mathbf{E}[\kappa] \in \mathcal{H}$,
- covariances $\mathbf{Cov}(\kappa) \in \mathcal{L}(\mathcal{H})$,
- Gaussian random variables $\kappa \sim N(m, C)$.

Let κ be a Gaussian random field with a.s. continuous paths.

Can view κ also as random variable $\kappa : \Omega \to C(D)$ with values in $C(D)$.

Since $C(D) \hookrightarrow L^2(D)$ for bounded $D \subset \mathbb{R}^d$, we take $\kappa : \Omega \to L^2(D)$ as Hilbert space-valued random variable.

For separable Hilbert spaces $\mathcal{H}$ we can (analogously to $\mathbb{R}^n$) define

- Lebesgue spaces $L^2(\Omega; \mathcal{H})$,
- expectations $\mathbf{E}[\kappa] \in \mathcal{H}$,
- covariances $\mathbf{Cov}(\kappa) \in \mathcal{L}(\mathcal{H})$,
- Gaussian random variables $\kappa \sim N(m, C)$.

If κ GRF with m, c as mean and covariance function, then $\kappa \sim N(m, C)$ with

$$C\phi(\boldsymbol{x}) = \int_D c(\boldsymbol{x}, \boldsymbol{y})\phi(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y}$$

(and vice versa).

Consider a CONS $\{\phi_k\}_{k\in\mathbb{N}}$ of $\mathcal{H}$. Then for $\kappa \sim N(0, C)$

$$\kappa(\omega) = \sum_{k=1}^{\infty} \xi_k(\omega)\, \phi_k, \qquad \xi_k(\omega) = \langle \kappa(\omega), \phi_k \rangle,$$

and $\boldsymbol{\xi} = (\xi_k)_{k\in\mathbb{N}}$ is a Gaussian random variable in $\ell^2(\mathbb{R})$ with

$$\mathbf{E}\left[\xi_k\right] = 0, \qquad \mathbf{Cov}(\xi_k, \xi_l) = \langle C\phi_k, \phi_l \rangle.$$

Consider a CONS $\{\phi_k\}_{k\in\mathbb{N}}$ of $\mathcal{H}$. Then for $\kappa \sim N(0, C)$

$$\kappa(\omega) = \sum_{k=1}^{\infty} \xi_k(\omega)\,\phi_k, \qquad \xi_k(\omega) = \langle \kappa(\omega), \phi_k \rangle,$$

and $\boldsymbol{\xi} = (\xi_k)_{k\in\mathbb{N}}$ is a Gaussian random variable in $\ell^2(\mathbb{R})$ with

$$\mathbf{E}\left[\xi_k\right] = 0, \qquad \mathbf{Cov}(\xi_k, \xi_l) = \langle C\phi_k, \phi_l \rangle.$$

Common choice for CONS uses eigenpairs (λ_k, ϕ_k) of C; then

$$\boldsymbol{\xi} \sim N(0, \Lambda), \qquad \Lambda = \operatorname{diag}(\lambda_k)_{k\in\mathbb{N}}.$$

This yields Karhunen-Loève expansion (KLE) of random field $\kappa(\cdot, \omega)$.

Consider a CONS $\{\phi_k\}_{k\in\mathbb{N}}$ of $\mathcal{H}$. Then for $\kappa \sim N(0, C)$

$$\kappa(\omega) = \sum_{k=1}^{\infty} \xi_k(\omega)\,\phi_k, \qquad \xi_k(\omega) = \langle \kappa(\omega), \phi_k \rangle,$$

and $\boldsymbol{\xi} = (\xi_k)_{k\in\mathbb{N}}$ is a Gaussian random variable in $\ell^2(\mathbb{R})$ with

$$\mathbf{E}[\xi_k] = 0, \qquad \mathbf{Cov}(\xi_k, \xi_l) = \langle C\phi_k, \phi_l \rangle.$$

Common choice for CONS uses eigenpairs (λ_k, ϕ_k) of C; then

$$\boldsymbol{\xi} \sim N(0, \Lambda), \qquad \Lambda = \operatorname{diag}(\lambda_k)_{k\in\mathbb{N}}.$$

This yields Karhunen-Loève expansion (KLE) of random field $\kappa(\cdot, \omega)$.

For numerical purposes: truncate series after M terms

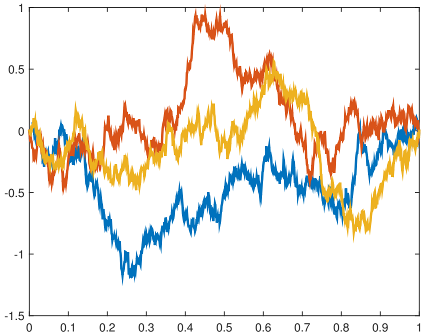
$$\kappa(x, \omega) \approx \kappa_M(x, \omega) = \sum_{m=1}^{M} \xi_m(\omega)\,\phi_m(x).$$

Gaussian Random Fields
Example:

Brownian bridge on $D = [0, 1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x, y) = \min(x, y) - xy$$

Realizations



Gaussian Random Fields

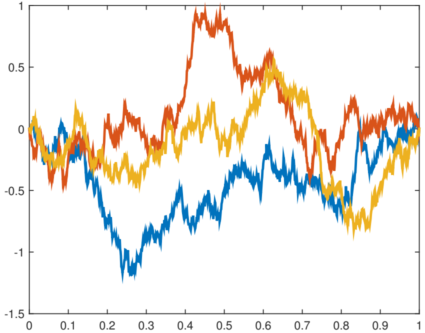
Example:

Brownian bridge on $D = [0, 1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x, y) = \min(x, y) - xy \quad \Leftrightarrow \quad \kappa(x, \omega) = \frac{\sqrt{2}}{\pi} \sum_{m=1}^{\infty} \xi_m(\omega) \sin(m\pi x),$$

with $\xi_m \sim N(0, m^{-2})$ i.i.d.

Realizations

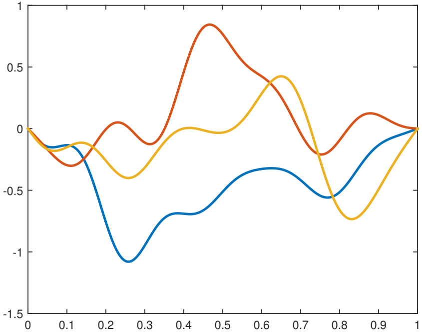


Brownian bridge on $D = [0, 1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x, y) = \min(x, y) - xy \quad \Leftrightarrow \quad \kappa(x, \omega) = \frac{\sqrt{2}}{\pi} \sum_{m=1}^{\infty} \xi_m(\omega) \sin(m\pi x),$$

with $\xi_m \sim N(0, m^{-2})$ i.i.d.

Realizations with $M = 10$ KL terms

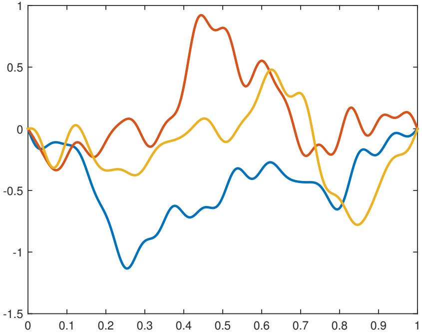


Brownian bridge on $D = [0,1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x,y) = \min(x,y) - xy \quad \Leftrightarrow \quad \kappa(x,\omega) = \frac{\sqrt{2}}{\pi} \sum_{m=1}^{\infty} \xi_m(\omega) \sin(m\pi x),$$

with $\xi_m \sim N(0, m^{-2})$ i.i.d.

Realizations with $M = 25$ KL terms



Gaussian Random Fields

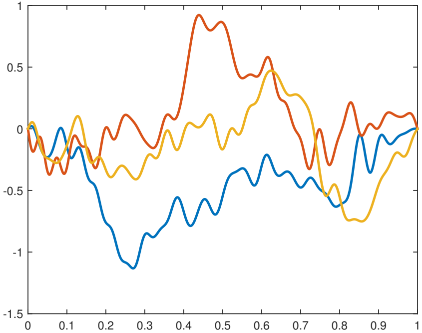
Example:

Brownian bridge on $D = [0, 1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x, y) = \min(x, y) - xy \quad \Leftrightarrow \quad \kappa(x, \omega) = \frac{\sqrt{2}}{\pi} \sum_{m=1}^{\infty} \xi_m(\omega) \sin(m\pi x),$$

with $\xi_m \sim N(0, m^{-2})$ i.i.d.

Realizations with $M = 50$ KL terms

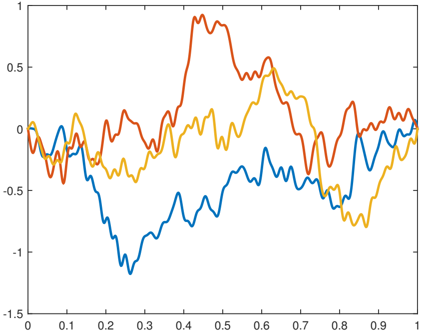


Brownian bridge on $D = [0,1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x,y) = \min(x,y) - xy \quad \Leftrightarrow \quad \kappa(x,\omega) = \frac{\sqrt{2}}{\pi} \sum_{m=1}^{\infty} \xi_m(\omega) \sin(m\pi x),$$

with $\xi_m \sim N(0, m^{-2})$ i.i.d.

Realizations with $M = 100$ KL terms



Gaussian Random Fields

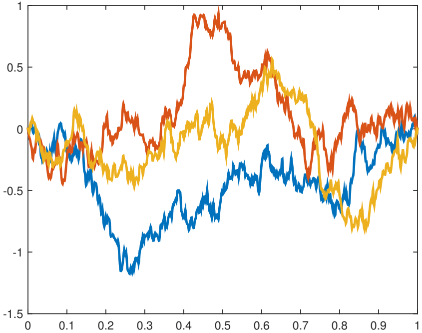
Example:

Brownian bridge on $D = [0,1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x,y) = \min(x,y) - xy \quad \Leftrightarrow \quad \kappa(x,\omega) = \frac{\sqrt{2}}{\pi}\sum_{m=1}^{\infty} \xi_m(\omega)\sin(m\pi x),$$

with $\xi_m \sim N(0, m^{-2})$ i.i.d.

Realizations with $M = 500$ KL terms



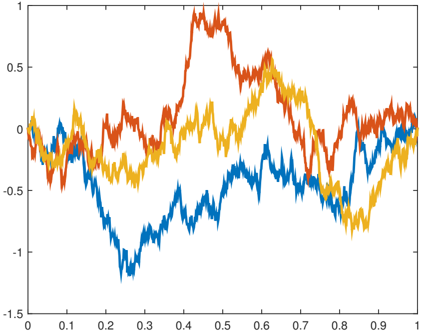
Gaussian Random Fields
Example:

Brownian bridge on $D = [0,1]$, i.e., Gaussian random field with

$$m(x) = 0, \quad c(x,y) = \min(x,y) - xy \quad \Leftrightarrow \quad \kappa(x,\omega) = \frac{\sqrt{2}}{\pi} \sum_{m=1}^{\infty} \xi_m(\omega) \sin(m\pi x),$$

with $\xi_m \sim N(0, m^{-2})$ i.i.d.

Realizations with $M = \infty$ KL terms



$$c(\boldsymbol{x}, \boldsymbol{y}) = c(r) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\,r}{\rho}\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}\,r}{\rho}\right), \quad r = \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

K_ν : modified Bessel function of order ν

Parameters

- σ^2 : variance
- ρ : correlation length
- ν : smoothness parameter

$$c(\boldsymbol{x}, \boldsymbol{y}) = c(r) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\, r}{\rho} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}\, r}{\rho} \right), \quad r = \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

K_ν : modified Bessel function of order ν

Parameters

σ^2 : variance

ρ : correlation length

ν : smoothness parameter

Special cases:

$\nu = \frac{1}{2}$:	$c(r) = \sigma^2 \exp(-\sqrt{2}r/\rho)$	exponential covariance
$\nu = 1$:	$c(r) = \sigma^2 \left(\frac{2r}{\rho}\right) K_1 \left(\frac{2r}{\rho}\right)$	Bessel covariance
$\nu \to \infty$:	$c(r) = \sigma^2 \exp(-r^2/\rho^2)$	Gaussian covariance

$$c(\boldsymbol{x},\boldsymbol{y}) = c(r) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\,r}{\rho}\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}\,r}{\rho}\right), \quad r = \|\boldsymbol{x}-\boldsymbol{y}\|_2$$

K_ν : modified Bessel function of order ν

Parameters

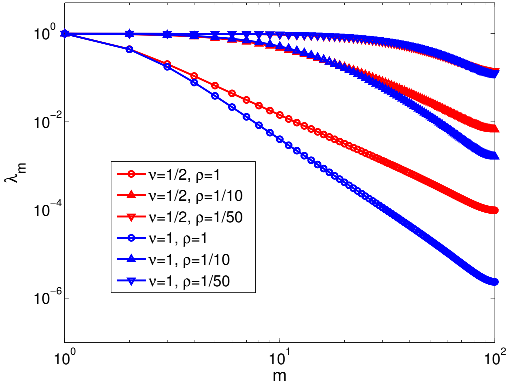
- σ^2 : variance
- ρ : correlation length
- ν : smoothness parameter

Special cases:

$\nu = \frac{1}{2}$:	$c(r) = \sigma^2 \exp(-\sqrt{2}r/\rho)$	exponential covariance
$\nu = 1$:	$c(r) = \sigma^2 \left(\frac{2r}{\rho}\right) K_1\left(\frac{2r}{\rho}\right)$	Bessel covariance
$\nu \to \infty$:	$c(r) = \sigma^2 \exp(-r^2/\rho^2)$	Gaussian covariance

Smoothness: Realizations $\kappa(\cdot,\omega)$ are k times differentiable $\Leftrightarrow \nu > k$.

Preasymptotic plateau (determined by correlation length ρ) before asymptotic decay sets in.

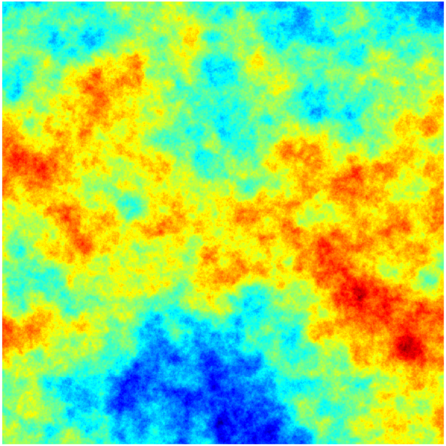


Eigenvalue decay, Matérn kernel, $D = [-1, 1]$.

Rate:

$$\lambda_m \sim m^{-(1+2\nu/d)} \quad (m \to \infty)$$

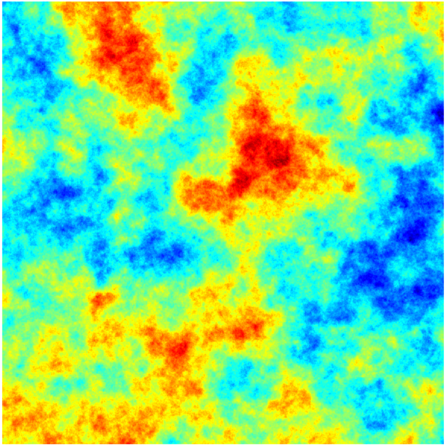
[Lord, Powell & Shardlow, 2014], [Widom, 1963]



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$

Gaussian Random Fields

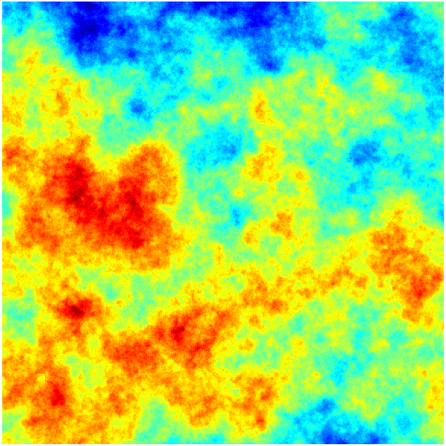
Realizations of GRF with Matérn covariance



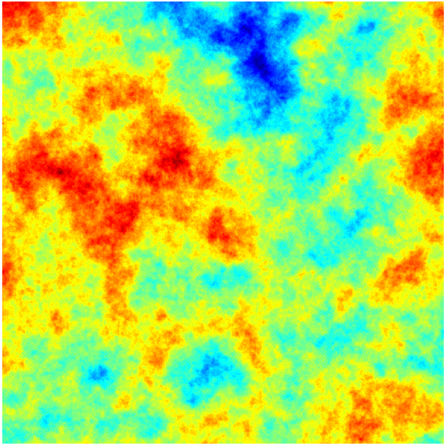
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$

Gaussian Random Fields

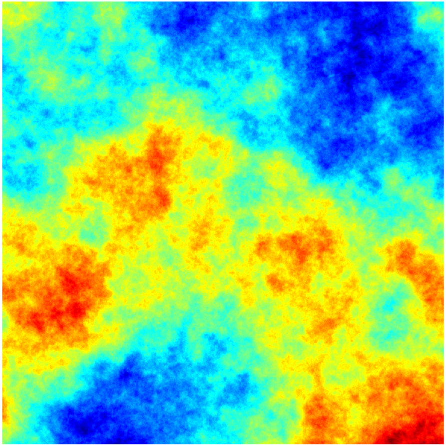
Realizations of GRF with Matérn covariance



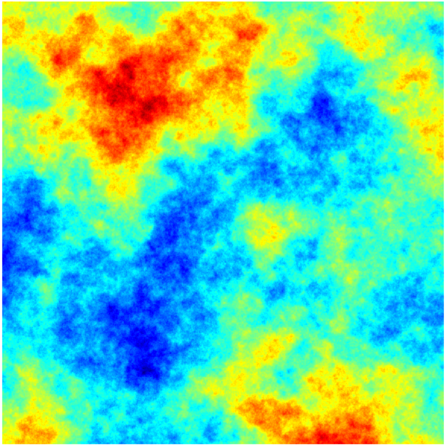
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$



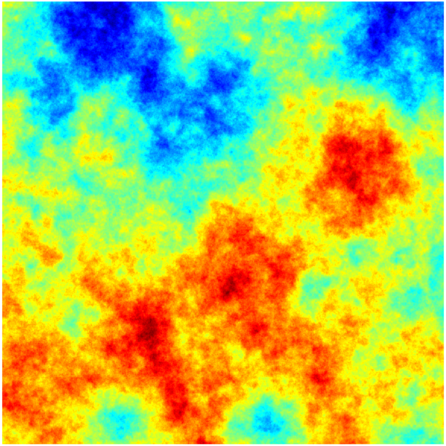
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$



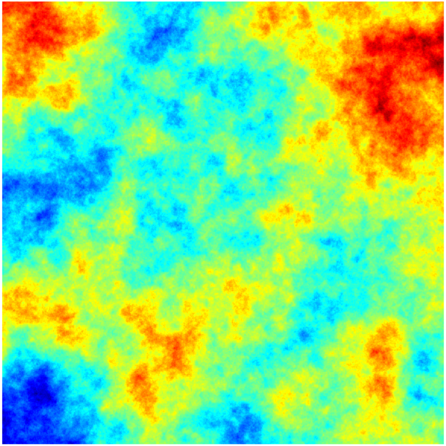
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$



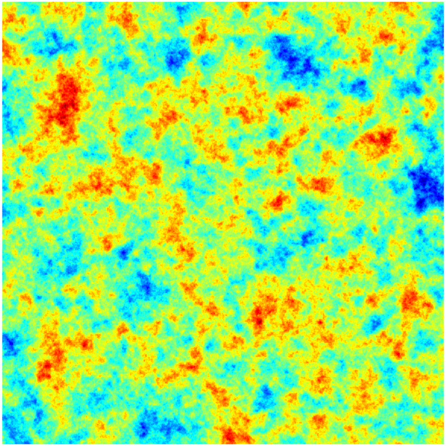
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.5$

Gaussian Random Fields

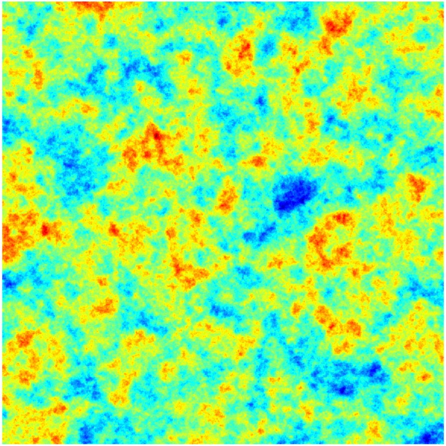
Realizations of GRF with Matérn covariance



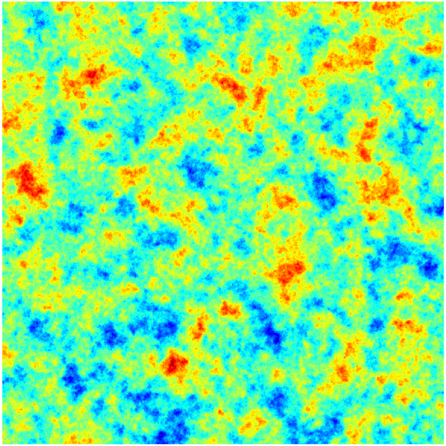
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$

Gaussian Random Fields

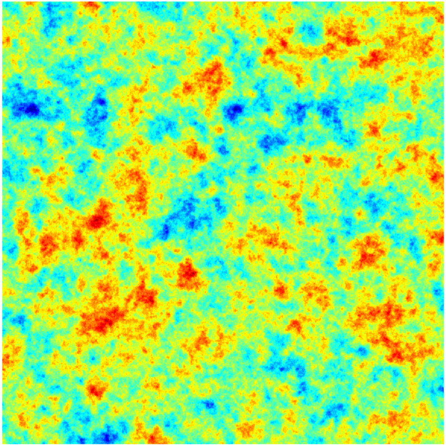
Realizations of GRF with Matérn covariance



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$



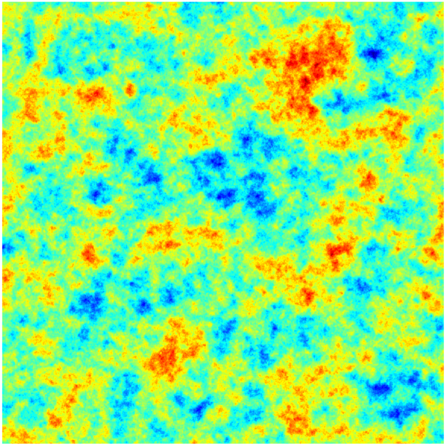
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$



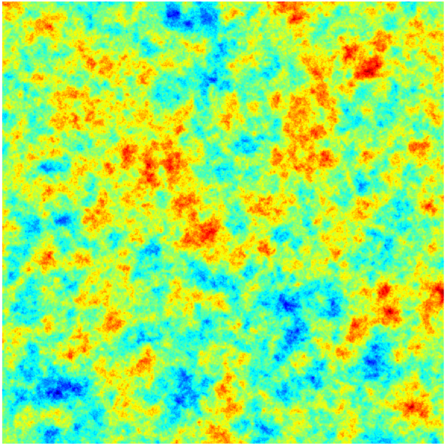
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$

Gaussian Random Fields

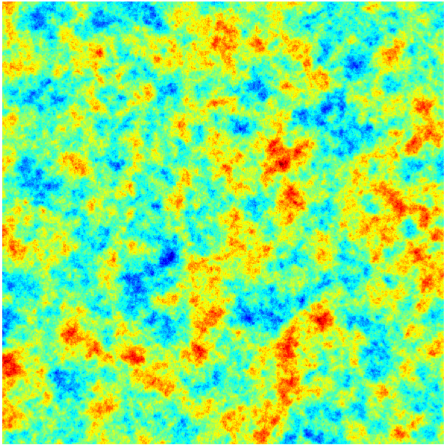
Realizations of GRF with Matérn covariance



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$



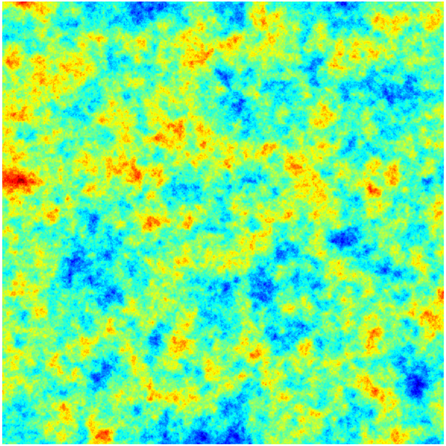
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$



Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$

Gaussian Random Fields

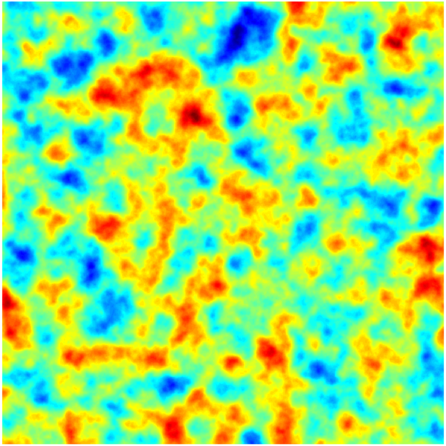
Realizations of GRF with Matérn covariance



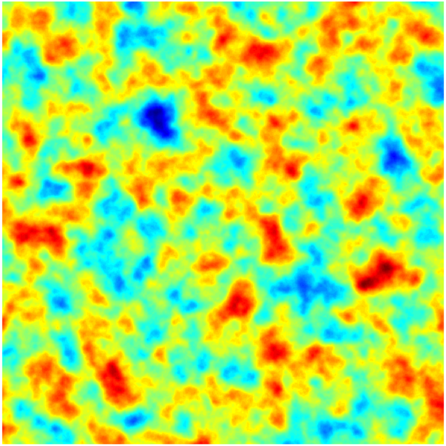
Matérn covariance, $\sigma = 1$, $\nu = \frac{1}{2}$, $\rho = 0.05$

Gaussian Random Fields

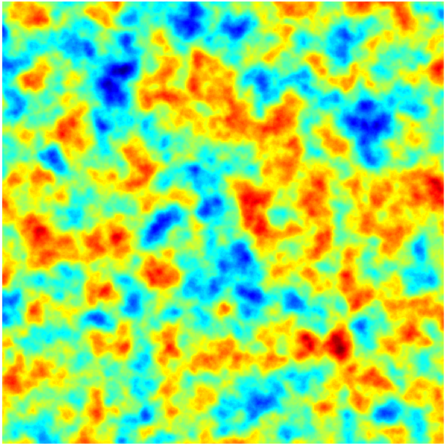
Realizations of GRF with Matérn covariance



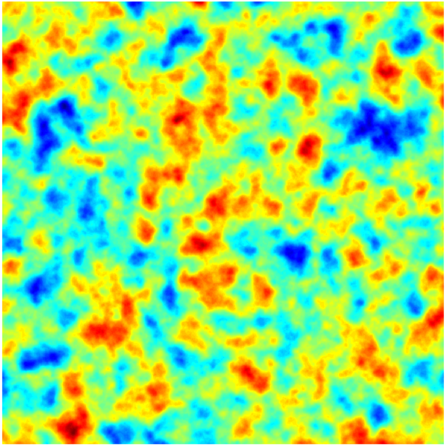
Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$



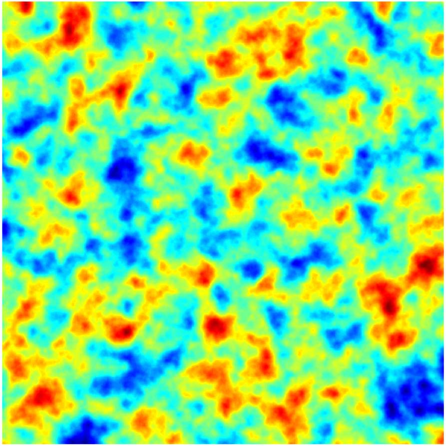
Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$



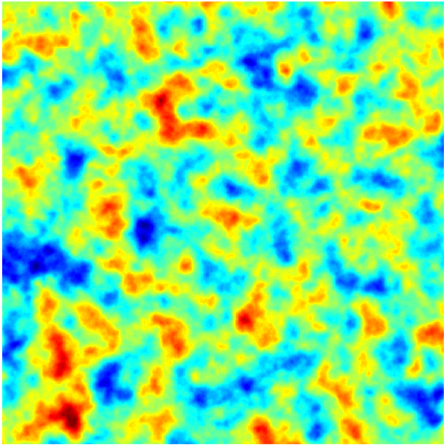
Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$



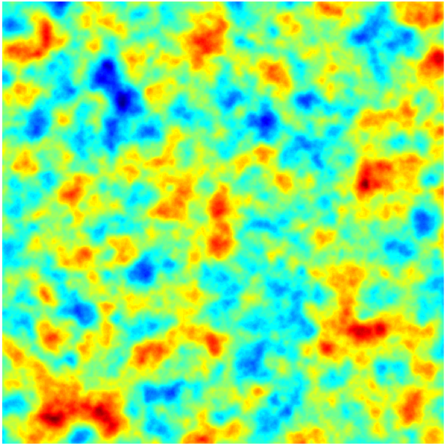
Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$

Gaussian Random Fields

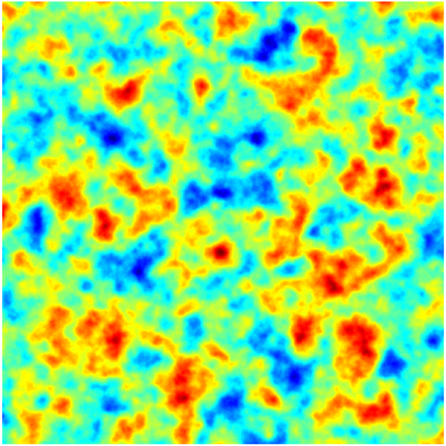
Realizations of GRF with Matérn covariance



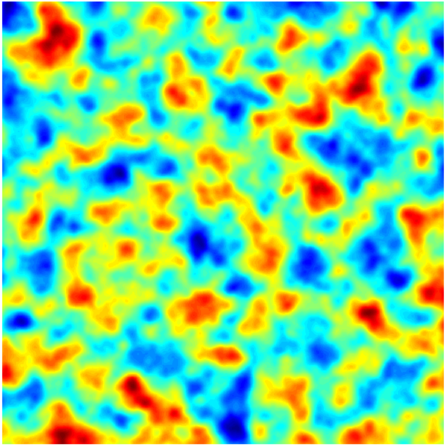
Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$



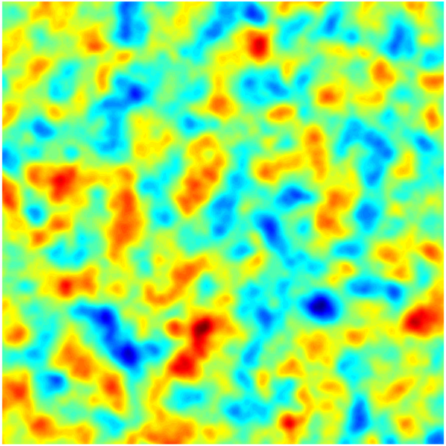
Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$



Matérn covariance, $\sigma = 1$, $\nu = \frac{3}{2}$, $\rho = 0.05$



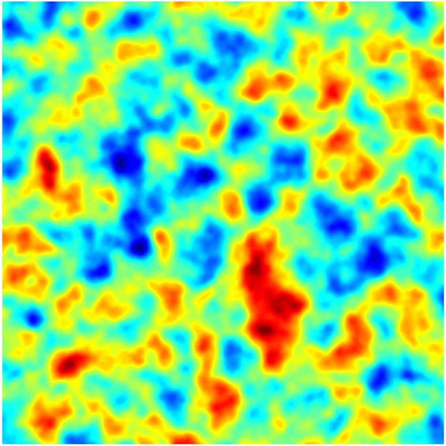
Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$



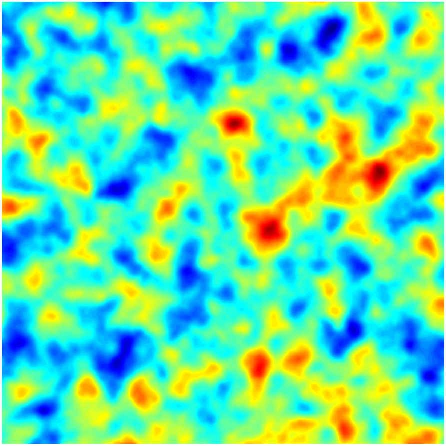
Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$

Gaussian Random Fields

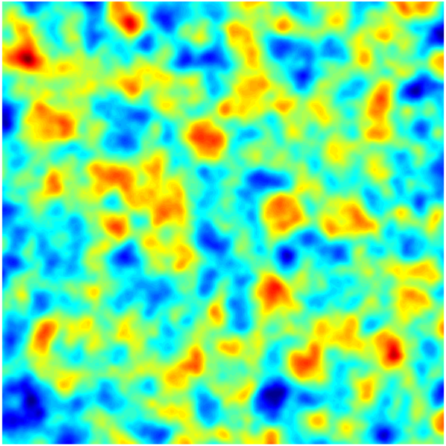
Realizations of GRF with Matérn covariance



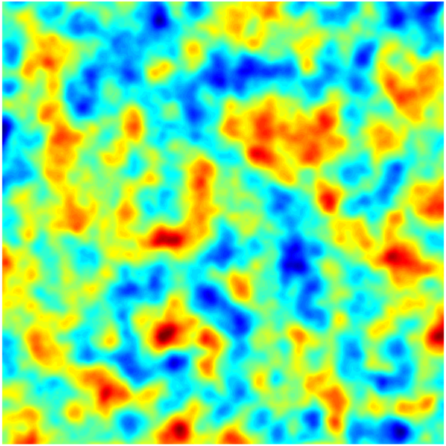
Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$



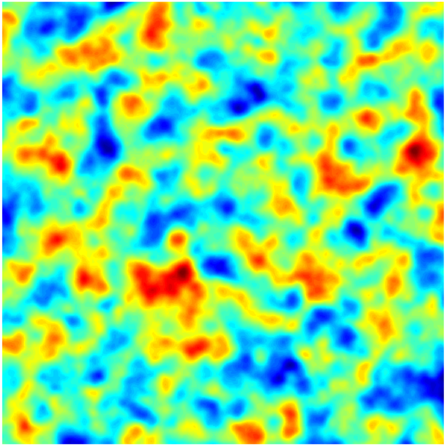
Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$



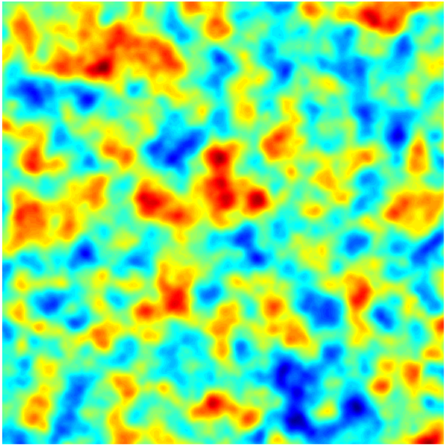
Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$

Gaussian Random Fields

Realizations of GRF with Matérn covariance



Matérn covariance, $\sigma = 1$, $\nu = \frac{5}{2}$, $\rho = 0.05$

Contents

4 An Inverse Problem for Groundwater Flow at WIPP
4.1 Gaussian Random Fields
4.2 Gaussian Random Field Models from Direct Observations
4.3 Solution of the Forward Problem
4.4 Bayesian Inversion
4.5 Improving MCMC Proposals in Hilbert Space
4.6 Numerical Results: Bayesian Inversion for WIPP

Assumptions on m and c:

- $m(\boldsymbol{x}) = \mathbf{E}[\kappa](\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i f_i(\boldsymbol{x})$, here: $n = 1, f_1 \equiv 1$
- c belongs to Matérn class of covariance functions:

$$c(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\,|\boldsymbol{x} - \boldsymbol{y}|_2}{\rho} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}\,|\boldsymbol{x} - \boldsymbol{y}|_2}{\rho} \right),$$

K_ν : modified Bessel function of order ν

with σ^2 variance, ρ correlation length, ν smoothness parameter

Assumptions on m and c:

- $m(\boldsymbol{x}) = \mathbf{E}[\kappa](\boldsymbol{x}) = \sum_{i=1}^{n} \beta_i f_i(\boldsymbol{x})$, here: $n = 1, f_1 \equiv 1$
- c belongs to Matérn class of covariance functions:

$$c(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sigma^2}{2^{\nu-1}\,\Gamma(\nu)} \left(\frac{2\sqrt{\nu}\,|\boldsymbol{x}-\boldsymbol{y}|_2}{\rho} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}\,|\boldsymbol{x}-\boldsymbol{y}|_2}{\rho} \right),$$

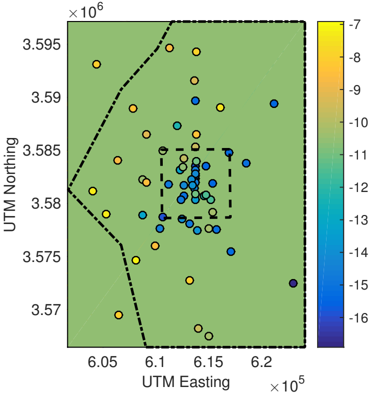
K_ν : modified Bessel function of order ν

with σ^2 variance, ρ correlation length, ν smoothness parameter

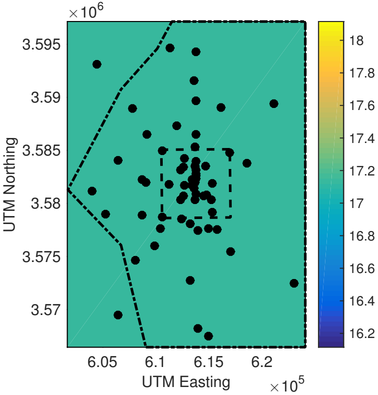
Results: Restricted maximum likelihood estimates given measurements of a are

$$\beta_1 = -10.55, \quad \sigma^2 = 17.15, \quad \rho = 6510, \quad \nu = 0.5 \text{ (fixed)}$$

With these REML estimates, pointwise mean and variance of log *a* obtained as:

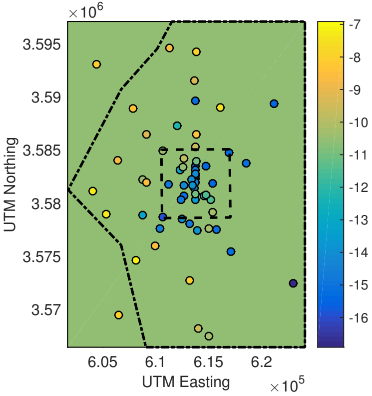


Mean, observations

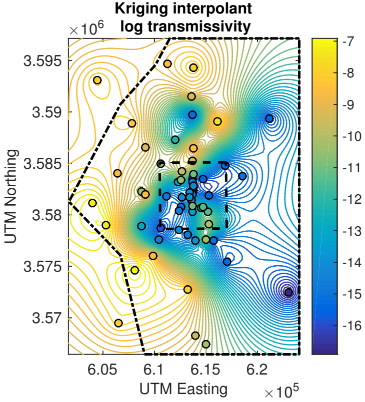


Variance, well locations

Better: geostatistical interpolant

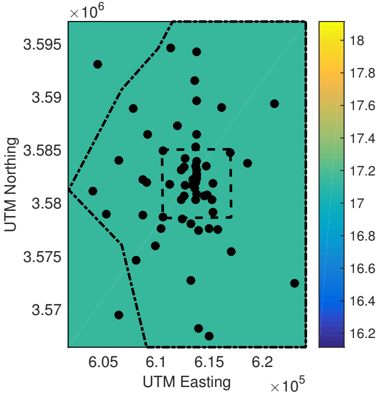


Mean observations

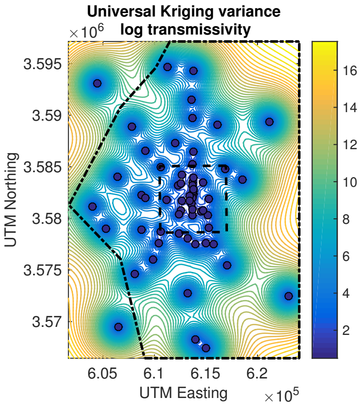


Kriging mean, observations

Better: geostatistical interpolant and its error



Variance, well locations



Kriging variance, well locations

Let κ be a zero-mean GRF with covariance function c.

Given observations $\{\kappa(\boldsymbol{x}_j) = \kappa_j\}_{j=1}^N$, compute best linear unbiased estimate

$$\hat{\kappa}(\boldsymbol{x}) = \sum_{j=1}^N m_j(\boldsymbol{x})\,\kappa(\boldsymbol{x}_j).$$

Let κ be a zero-mean GRF with covariance function c.

Given observations $\{\kappa(\boldsymbol{x}_j) = \kappa_j\}_{j=1}^N$, compute best linear unbiased estimate

$$\hat{\kappa}(\boldsymbol{x}) = \sum_{j=1}^{N} m_j(\boldsymbol{x})\, \kappa(\boldsymbol{x}_j).$$

Explicit solution

$$\hat{\kappa}(\boldsymbol{x}) = \boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{x}), \qquad \boldsymbol{C} = [c(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^N, \quad \boldsymbol{c}(\boldsymbol{x}) = [c(\boldsymbol{x}_i, \boldsymbol{x})]_{i=1}^N$$

with error covariance

$$\hat{c}(\boldsymbol{x}, \boldsymbol{x}) := \mathbf{E}\left[(\kappa(\boldsymbol{x}) - \hat{\kappa}(\boldsymbol{x}), \kappa(\boldsymbol{y}) - \hat{\kappa}(\boldsymbol{y}))\right] = c(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{c}(\boldsymbol{x})\boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{y})$$

Let κ be a zero-mean GRF with covariance function c.

Given observations $\{\kappa(\boldsymbol{x}_j) = \kappa_j\}_{j=1}^N$, compute best linear unbiased estimate

$$\hat{\kappa}(\boldsymbol{x}) = \sum_{j=1}^{N} m_j(\boldsymbol{x})\, \kappa(\boldsymbol{x}_j).$$

Explicit solution

$$\hat{\kappa}(\boldsymbol{x}) = \boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{x}), \qquad \boldsymbol{C} = [c(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^N, \quad \boldsymbol{c}(\boldsymbol{x}) = [c(\boldsymbol{x}_i, \boldsymbol{x})]_{i=1}^N$$

with error covariance

$$\hat{c}(\boldsymbol{x}, \boldsymbol{x}) := \mathbf{E}\left[(\kappa(\boldsymbol{x}) - \hat{\kappa}(\boldsymbol{x}), \kappa(\boldsymbol{y}) - \hat{\kappa}(\boldsymbol{y}))\right] = c(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{c}(\boldsymbol{x})\boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{y})$$

coincide with mean and covariance of GRF κ conditioned on $\{\kappa(\boldsymbol{x}_j) = \kappa_j\}_{j=1}^N$.

Let κ be a zero-mean GRF with covariance function c.

Given observations $\{\kappa(\boldsymbol{x}_j) = \kappa_j\}_{j=1}^N$, compute best linear unbiased estimate

$$\hat{\kappa}(\boldsymbol{x}) = \sum_{j=1}^N m_j(\boldsymbol{x})\, \kappa(\boldsymbol{x}_j).$$

Explicit solution

$$\hat{\kappa}(\boldsymbol{x}) = \boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{x}), \qquad \boldsymbol{C} = [c(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1}^N, \quad \boldsymbol{c}(\boldsymbol{x}) = [c(\boldsymbol{x}_i, \boldsymbol{x})]_{i=1}^N$$

with error covariance

$$\hat{c}(\boldsymbol{x}, \boldsymbol{x}) := \mathbf{E}\left[(\kappa(\boldsymbol{x}) - \hat{\kappa}(\boldsymbol{x}), \kappa(\boldsymbol{y}) - \hat{\kappa}(\boldsymbol{y}))\right] = c(\boldsymbol{x}, \boldsymbol{y}) - \boldsymbol{c}(\boldsymbol{x})\boldsymbol{C}^{-1}\boldsymbol{c}(\boldsymbol{y})$$

coincide with mean and covariance of GRF κ conditioned on $\{\kappa(\boldsymbol{x}_j) = \kappa_j\}_{j=1}^N$.

Note: Kriging coincides with radial basis interpolation with suitable radial basis.

Contents

4 An Inverse Problem for Groundwater Flow at WIPP

4.1 Gaussian Random Fields

4.2 Gaussian Random Field Models from Direct Observations

4.3 Solution of the Forward Problem

4.4 Bayesian Inversion

4.5 Improving MCMC Proposals in Hilbert Space

4.6 Numerical Results: Bayesian Inversion for WIPP

Solution of the Forward Problem

Parametric representation of random fields

Given Kriging mean $\hat{\kappa}$ and covariance $\hat{c}$, compute the KL expansion $\{\lambda_m, \phi_m\}_{m=1}^{\infty}$ for $\hat{c}$ and approximate, using $\xi_m \sim N(0, \lambda_m)$

$$\log a(\boldsymbol{x}, \omega) = \kappa(\boldsymbol{x}, \omega) \approx \hat{\kappa}(\boldsymbol{x}) + \sum_{m=1}^{M} \phi_m(\boldsymbol{x})\, \xi_m(\omega)$$

Given Kriging mean $\hat{\kappa}$ and covariance $\hat{c}$, compute the KL expansion $\{\lambda_m, \phi_m\}_{m=1}^{\infty}$ for $\hat{c}$ and approximate, using $\xi_m \sim N(0, \lambda_m)$

$$\log a(\boldsymbol{x}, \boldsymbol{\xi}(\omega)) = \kappa(\boldsymbol{x}, \omega) \approx \hat{\kappa}(\boldsymbol{x}) + \sum_{m=1}^{M} \phi_m(\boldsymbol{x})\, \xi_m(\omega)$$

Can think of $\log a$ as a function of $\boldsymbol{x}$ with random parameter $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M)$.

Solution of the Forward Problem

Parametric representation of random fields

Given Kriging mean $\hat{\kappa}$ and covariance $\hat{c}$, compute the KL expansion $\{\lambda_m, \phi_m\}_{m=1}^{\infty}$ for $\hat{c}$ and approximate, using $\xi_m \sim N(0, \lambda_m)$

$$\log a(\boldsymbol{x}, \boldsymbol{\xi}(\omega)) = \kappa(\boldsymbol{x}, \omega) \approx \hat{\kappa}(\boldsymbol{x}) + \sum_{m=1}^{M} \phi_m(\boldsymbol{x})\, \xi_m(\omega)$$

Can think of $\log a$ as a function of $\boldsymbol{x}$ with random parameter $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)$.

Solution $(\boldsymbol{u}, p)$ of (PDE-Mixed) pair of functions of $\boldsymbol{x}$ with random parameter $\boldsymbol{\xi}$,

$$(\boldsymbol{u}, p)(\cdot, \boldsymbol{\xi}) \in H_0(\mathrm{div}; D) \times L^2(D) \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^M,$$

analogously for the travel time of released particles:

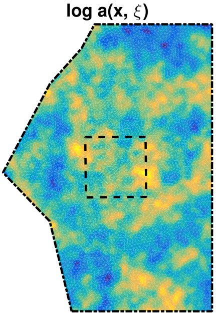
$$t_{\mathrm{exit}}(\boldsymbol{\xi}) \in \mathbb{R}, \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^M.$$

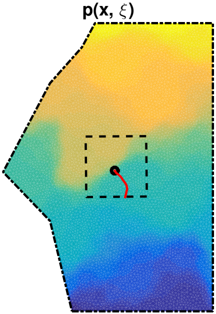
To approximate CDF of t_{exit},

$$F(t) = \mathbf{P}(t_{\text{exit}}(\boldsymbol{\xi}(\omega)) < t)$$

sample $\boldsymbol{\xi} \sim N(0, \Lambda)$ and solve (Mixed)/ODE many times. (Here many = 20,000.)

$$\boldsymbol{\xi} = \begin{pmatrix} 0.53 \\ 1.83 \\ -2.26 \\ 0.21 \\ 1.31 \\ \vdots \end{pmatrix}$$





$t_{\text{exit}}(\boldsymbol{\xi}) =$ 19,311 years

Solution of the Forward Problem

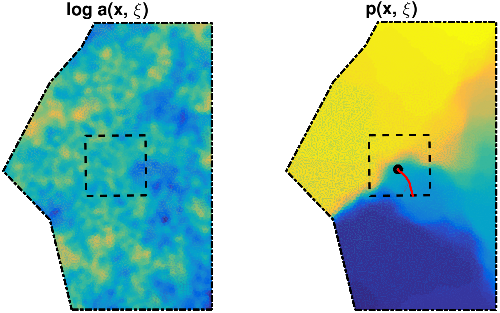
Parametric representation of random fields

To approximate CDF of t_{exit},

$$F(t) = \mathbf{P}(t_{\text{exit}}(\boldsymbol{\xi}(\omega)) < t)$$

sample $\boldsymbol{\xi} \sim N(0, \Lambda)$ and solve (Mixed)/ODE many times. (Here many = 20,000.)

$$\boldsymbol{\xi} = \begin{pmatrix} 0.57 \\ -0.97 \\ -0.71 \\ -0.09 \\ -0.17 \\ \vdots \end{pmatrix}$$



$t_{\text{exit}}(\boldsymbol{\xi}) =$ 8,272 years

Solution of the Forward Problem

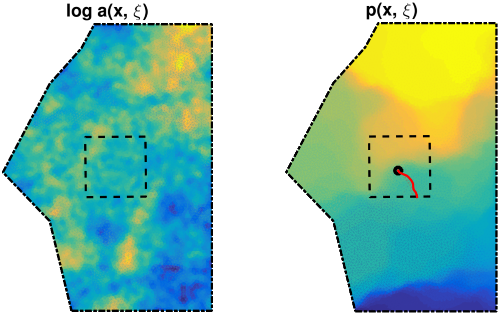
Parametric representation of random fields

To approximate CDF of t_{exit},

$$F(t) = \mathbf{P}(t_{\text{exit}}(\boldsymbol{\xi}(\omega)) < t)$$

sample $\boldsymbol{\xi} \sim N(0, \Lambda)$ and solve (Mixed)/ODE many times. (Here many = 20,000.)

$$\boldsymbol{\xi} = \begin{pmatrix} -1.03 \\ -1.76 \\ -2.01 \\ 1.32 \\ 0.63 \\ \vdots \end{pmatrix}$$



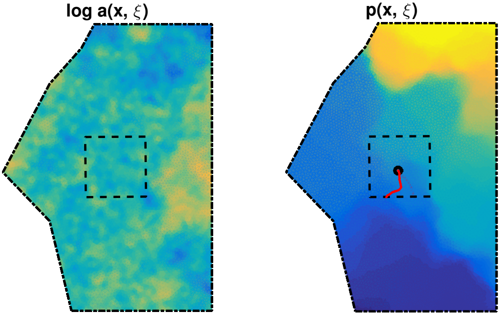
$t_{\text{exit}}(\boldsymbol{\xi}) =$ 31,237 years

To approximate CDF of t_{exit},

$$F(t) = \mathbf{P}(t_{\text{exit}}(\boldsymbol{\xi}(\omega)) < t)$$

sample $\boldsymbol{\xi} \sim N(0, \Lambda)$ and solve (Mixed)/ODE many times. (Here many = 20,000.)

$$\boldsymbol{\xi} = \begin{pmatrix} 0.03 \\ 2.30 \\ -0.19 \\ 0.58 \\ -0.39 \\ \vdots \end{pmatrix}$$



$t_{\text{exit}}(\boldsymbol{\xi}) =$ 102, 238 years

To approximate CDF of t_{exit},

$$F(t) = \mathbf{P}(t_{\text{exit}}(\boldsymbol{\xi}(\omega)) < t)$$

sample $\boldsymbol{\xi} \sim N(0, \Lambda)$ and solve (Mixed)/ODE many times. (Here many = 20,000.)

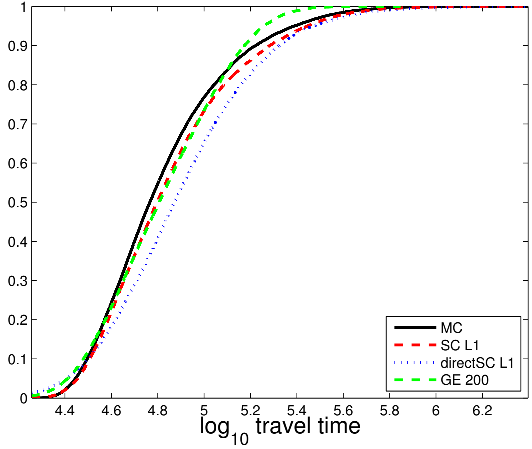
More efficient:

- Compute cheaper surrogate of mapping $\boldsymbol{\xi} \mapsto (\boldsymbol{u}, p)(\cdot, \boldsymbol{\xi})$ or $\boldsymbol{\xi} \mapsto t_{\text{exit}}(\boldsymbol{\xi})$.
- Evaluate surrogate 20,000 times.

We tried two kind of surrogates: [Cliffe et al., 2016]

- polynomial approximation based on sparse grid collocation operators (popular among numerical analysts in UQ community)
- Gaussian process emulators based on GRF approach/Kriging for mapping (popular among statisticians in UQ community)

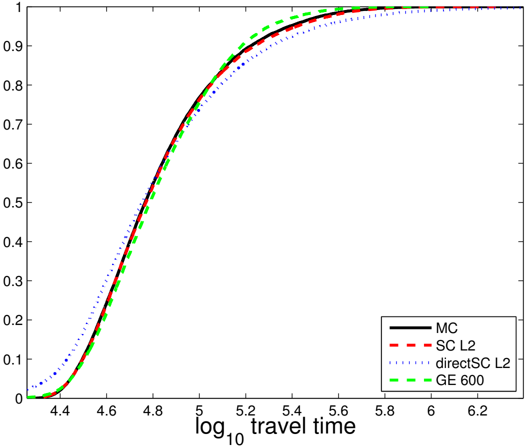
Plain Monte Carlo and Polynomials for $(\boldsymbol{u}, p)$, Polynomials for t_{exit}, Gaussian process emulators for t_{exit} for increasing degrees/number of training points



Solution of the Forward Problem

Results for surrogates (M = 20)

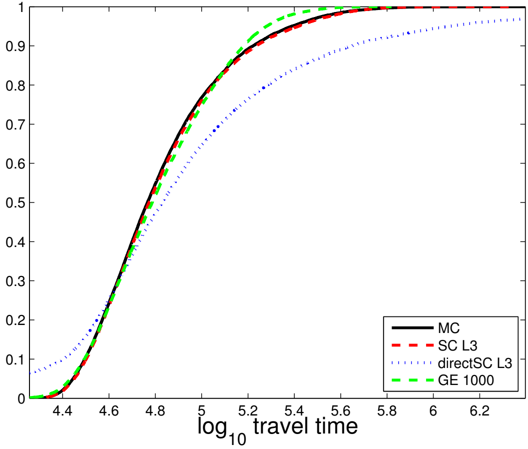
Plain Monte Carlo and Polynomials for $(\boldsymbol{u}, p)$, Polynomials for t_{exit}, Gaussian process emulators for t_{exit} for increasing degrees/number of training points



Solution of the Forward Problem

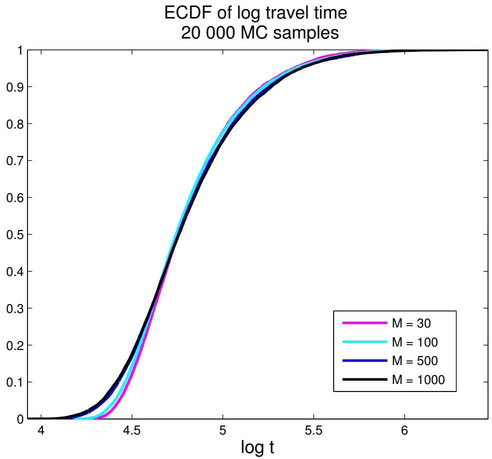
Results for surrogates (M = 20)

Plain Monte Carlo and Polynomials for $(\boldsymbol{u}, p)$, Polynomials for t_{exit}, Gaussian process emulators for t_{exit} for increasing degrees/number of training points



Solution of the Forward Problem

Effect of KL truncation



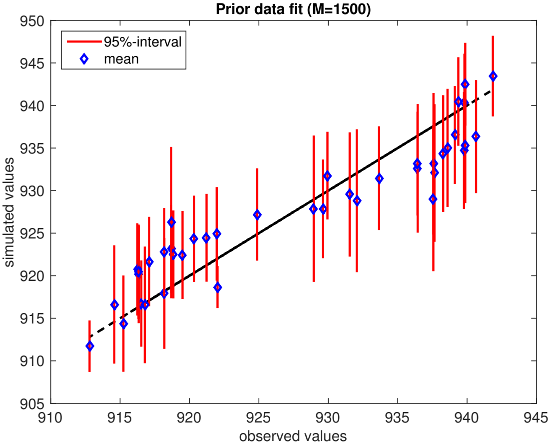
Problem seems to require substantially more than $M = 100$ KL terms.
This makes the use of surrogates infeasible.

Contents

4 An Inverse Problem for Groundwater Flow at WIPP
4.1 Gaussian Random Fields
4.2 Gaussian Random Field Models from Direct Observations
4.3 Solution of the Forward Problem
4.4 Bayesian Inversion
4.5 Improving MCMC Proposals in Hilbert Space
4.6 Numerical Results: Bayesian Inversion for WIPP

So far: random field model for log a based only on (direct) measurements of log a.

How well, does this random field model accomodate the observed values of p?



$$\boxed{\boldsymbol{y} \overset{!}{=} G(\boldsymbol{\xi})}$$

- Problem severely underdetermined.
- Observational noise: instead of $\boldsymbol{y}$, may observe perturbed data

$$\boldsymbol{y}^{\mathrm{obs}} = G(\boldsymbol{\xi}) + \boldsymbol{\epsilon},$$

  possibly not in range of G.
- G strongly smoothing, reconstruction unstable (ill-posed problem).
- Variational formulation (output least squares): determine $\boldsymbol{\xi}$ to minimize data misfit functional

$$\Phi(\boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{y}^{\mathrm{obs}} - G(\boldsymbol{\xi})\|^2.$$

- Regularization: include additional information to constrain solution.

- **Restriction to compact set.** If G is defined on a Banach space X, let E denote a reflexive Banach space compactly embedded in X, with norm $\|\cdot\|_E$. Instead of minimizing $\Phi(\boldsymbol{\xi})$ on X, restrict $\boldsymbol{\xi}$ to

$$E_\alpha := \{\boldsymbol{\xi} \in E : \|\boldsymbol{\xi}\|_E \leq \alpha\}, \qquad \alpha > 0.$$

  Then any minimizing sequence $\{\boldsymbol{\xi}_n\}$ in E_α contains a weakly convergent subsequence with limit $\boldsymbol{\xi}^\star \in E_\alpha$ such that $\phi(\boldsymbol{\xi}^\star) = \inf_{\boldsymbol{\xi}\in E_\alpha} \phi(\boldsymbol{\xi})$.
- **Tikhonov regularization**. Add a penalization term to data misfit functional and minimize the Tikhonov functional

$$I(\boldsymbol{\xi}) = \Phi(\boldsymbol{\xi}) + \frac{\alpha}{2}\|\boldsymbol{\xi}\|_E^2, \qquad \alpha > 0,$$

  over E. Again, minimizing sequences have weakly convergent subsequences with limit attaining $\inf_{\boldsymbol{\xi}\in E} I(\boldsymbol{\xi})$.

  Analysis of selection strategies for regularization parameter α in limit of vanishing noise $\|\boldsymbol{\epsilon}\|$ can be found in [Engl et al., 1996], [Hofmann, 2015].

- **Bounded noise** with noise level $\delta > 0$ (ususal deterministic formulation):

$$\|\boldsymbol{\epsilon}\| \le \delta.$$

- **Random noise**: $\boldsymbol{\epsilon}$ has known multivariate probability distribution, e.g. with (Lebesgue) density $\rho = \rho(\boldsymbol{\epsilon})$. For given $\boldsymbol{\xi}$, the observation data

$$\boldsymbol{y}^{\text{obs}} = G(\boldsymbol{\xi}) + \boldsymbol{\epsilon}$$

is a random vector with density $\rho(\boldsymbol{y}^{\text{obs}} - G(\boldsymbol{\xi}))$.

For centered Gaussian noise:

$$\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{C}), \qquad \boldsymbol{C} \in \mathbb{R}^{K \times K} \text{ positive definite,}$$

we have

$$\rho_{\boldsymbol{y}|\boldsymbol{\xi}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^K \det C}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - G(\boldsymbol{\xi}))^\top \boldsymbol{C}^{-1}(\boldsymbol{y} - G(\boldsymbol{\xi}))\right).$$

The conditional density $\rho_{\boldsymbol{y}|\boldsymbol{\xi}}$ is the likelihood of observing $\boldsymbol{y}$ given $\boldsymbol{\xi}$. Maximizing this is equivalent to minimizing the negative log-likelihood

$$-\log \rho_{\boldsymbol{y}|\boldsymbol{\xi}}(\boldsymbol{y}) = \frac{1}{2}\log\left((2\pi)^K \det \boldsymbol{C}\right) + \frac{1}{2}\|\boldsymbol{y} - G(\boldsymbol{\xi})\|^2_{\boldsymbol{C}^{-1}},$$

i.e., the data misfit functional adapted to the covariance of the Gaussian noise

$$\Phi(\boldsymbol{\xi}) := \frac{1}{2}\|\boldsymbol{y} - G(\boldsymbol{\xi})\|^2_{\boldsymbol{C}^{-1}}.$$

A solution for the inverse problem with random noise may be defined as the maximum likelihood estimator $\hat{\boldsymbol{\xi}}$ obtained by minimizing the negative log-likelihood.

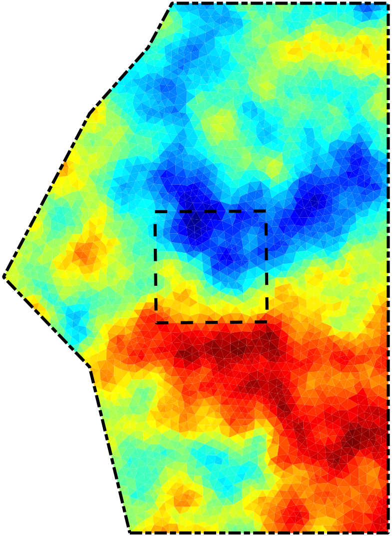
For centered Gaussian noise we recover the (weighted) output least squares solution.

- We now add uncertainty in the parameter to be estimated to our model by introducing a probability measure μ_0 on the space X containing the parameter $\boldsymbol{\xi}$.
- In the finite-dimensional case $\boldsymbol{\xi} \in \mathbb{R}^M$, we may assume μ_0 to have a (Lebesgue) density ρ_0.
- Viewing μ_0 as the prior probability distribution for $\boldsymbol{\xi}$, Bayes' theorem yields the posterior density for $\boldsymbol{\xi}$ after making the observations $\boldsymbol{y}$ as

$$\rho_{\boldsymbol{\xi}|\boldsymbol{y}}(\boldsymbol{\xi}) = \frac{\rho_{\boldsymbol{y}|\boldsymbol{\xi}}(\boldsymbol{y})\,\rho_0(\boldsymbol{\xi})}{Z}, \qquad Z := \int_{\mathbb{R}^M} \rho_{\boldsymbol{y}|\boldsymbol{\xi}}(\boldsymbol{y})\,\rho_0(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}$$

- We observe that the posterior measure $\mu^{\boldsymbol{y}}$ of $\boldsymbol{\xi}$ conditioned on the observation $\boldsymbol{y}$ is absolutely continuous with respect to the prior measure μ_0 ($\mu \ll \mu_0$) and that its Radon-Nikodym derivative satisfies

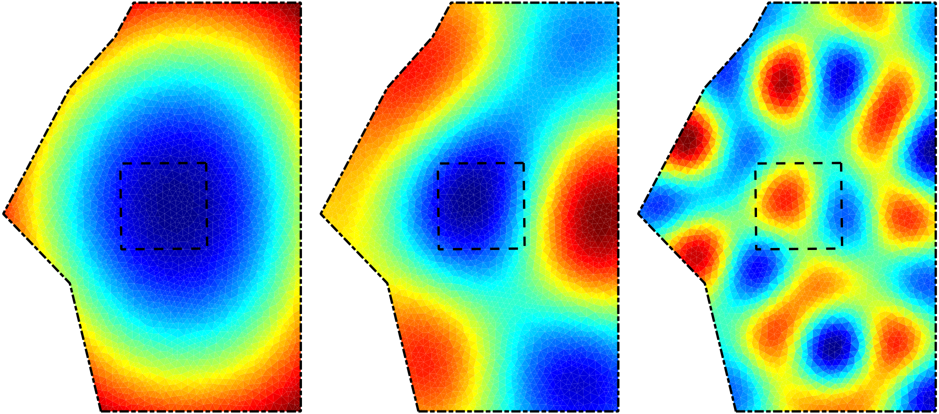
$$\frac{\mathrm{d}\mu^{\boldsymbol{y}}}{\mathrm{d}\mu_0}(\boldsymbol{\xi}) \propto \exp\left(-\Phi(\boldsymbol{\xi};\boldsymbol{y})\right).$$



Groundwater flow example:

Realization of $\log a(\boldsymbol{\xi})$, piecewise constant on mesh with 5135 triangles.

$$\log a(\boldsymbol{\xi}) = \sum_{m=1}^{\infty} \sqrt{\lambda_m}\phi_m(\boldsymbol{x})\xi_m.$$



$m = 1$ $m = 8$ $m = 35$

We want to incorporate into our model

$$\log a(\boldsymbol{x}, \boldsymbol{\xi}(\omega)) = \hat{\kappa}(\boldsymbol{x}) + \sum_{m=1}^{\infty} \phi_m(\boldsymbol{x})\, \xi_m(\omega), \qquad \xi_m \sim N(0, \lambda_m)$$

the available (noisy) observations of p at certain locations $\boldsymbol{x}_j \in D$.

We want to incorporate into our model

$$\log a(\boldsymbol{x}, \boldsymbol{\xi}(\omega)) = \hat{\kappa}(\boldsymbol{x}) + \sum_{m=1}^{\infty} \phi_m(\boldsymbol{x})\, \xi_m(\omega), \qquad \xi_m \sim N(0, \lambda_m)$$

the available (noisy) observations of p at certain locations $\boldsymbol{x}_j \in D$.

In stochastic terms: condition random field $\log a(\cdot, \omega)$ resp. random vector $\boldsymbol{\xi}(\omega)$ on the data $p(\boldsymbol{x}_j) = p_j, j = 1, \ldots, K$.

But due to nonlinearity of the mapping $\log a \mapsto p$, no nice explicit solution in this case.

We want to incorporate into our model

$$\log a(\boldsymbol{x}, \boldsymbol{\xi}(\omega)) = \hat{\kappa}(\boldsymbol{x}) + \sum_{m=1}^{\infty} \phi_m(\boldsymbol{x})\, \xi_m(\omega), \qquad \xi_m \sim N(0, \lambda_m)$$

the available (noisy) observations of p at certain locations $\boldsymbol{x}_j \in D$.

In stochastic terms: condition random field $\log a(\cdot, \omega)$ resp. random vector $\boldsymbol{\xi}(\omega)$ on the data $p(\boldsymbol{x}_j) = p_j, j = 1, \dots, K$.

But due to nonlinearity of the mapping $\log a \mapsto p$, no nice explicit solution in this case.

Bayes' rule provides expression for conditional probability measure.
For events, A, B, $\mathbf{P}(A), \mathbf{P}(B) > 0$:

$$\underbrace{\mathbf{P}(A|B)}_{\text{posterior probability}} = \frac{\overbrace{\mathbf{P}(B|A)}^{\text{likelihood}} \quad \overbrace{\mathbf{P}(A)}^{\text{prior probability}}}{\underbrace{\mathbf{P}(B)}_{\text{evidence}}}$$

- Random noise is multivariate Gaussian: $\boldsymbol{\epsilon} \sim N(0, \Sigma)$
- Prior measure is Gaussian measure on $\mathscr{H}$: $\mu_0 = N(0, C_0)$
- Forward map $G : \mathscr{H} \to \mathbb{R}^k$ is continuous and $\forall \alpha > 0 \ \exists K_\alpha < \infty$:

$$|G(\boldsymbol{\xi})| \leq K_\alpha \exp(\alpha \|\boldsymbol{\xi}\|^2_{\mathscr{H}}).$$

- $\boldsymbol{\xi} \sim \mu_0$ and $\boldsymbol{\epsilon} \sim N(0, \Sigma)$ are independent

- Random noise is multivariate Gaussian: $\boldsymbol{\epsilon} \sim N(0, \Sigma)$
- Prior measure is Gaussian measure on $\mathscr{H}$: $\mu_0 = N(0, C_0)$
- Forward map $G : \mathscr{H} \to \mathbb{R}^k$ is continuous and $\forall \alpha > 0\ \exists K_\alpha < \infty$:

$$|G(\boldsymbol{\xi})| \le K_\alpha \exp(\alpha \|\boldsymbol{\xi}\|_{\mathscr{H}}^2).$$

- $\boldsymbol{\xi} \sim \mu_0$ and $\boldsymbol{\epsilon} \sim N(0, \Sigma)$ are independent

Then the conditional probability measure $\mu^{\boldsymbol{y}}$ is given by Bayes' rule:

Theorem 4.1 (Bayes' rule [Stuart (2010)],[Dashti & Stuart (2016)])

The posterior measure $\mu^{\boldsymbol{y}}$ is given by

$$\mu^{\boldsymbol{y}}(\mathrm{d}\boldsymbol{\xi}) \propto \exp(-\Phi(\boldsymbol{\xi}; \boldsymbol{y}))\, \mu_0(\mathrm{d}\boldsymbol{\xi}), \qquad \Phi(\boldsymbol{\xi}; \boldsymbol{y}) = \frac{1}{2}|\boldsymbol{y} - G(\boldsymbol{\xi})|_{\Sigma^{-1}}^2.$$

- Parameter-to-observable map

$$G : \ell^2(\mathbb{N}) \to \mathbb{R}^J, \qquad \boldsymbol{\xi} \to \kappa(\boldsymbol{\xi}) \to p(\boldsymbol{\xi}) \to \{p(\boldsymbol{\xi})|_{x=x_j}\}_{j=1}^{k} =: \boldsymbol{p}^{\mathrm{obs}}.$$

- Gaussian random field in Hilbert space $\mathcal{H}$, e.g., $\mathcal{H} = L^2(D)$:

$$\kappa(x,\omega) = \hat{\kappa}(x) + \sum_{m\geq 1} \xi_m(\omega)\, \phi_m(x), \qquad \{\phi_m\}_{m=1}^{\infty} \text{ CONS of } \mathcal{H}.$$

- Parameter-to-observable map

$$G : \ell^2(\mathbb{N}) \to \mathbb{R}^J, \qquad \boldsymbol{\xi} \to \kappa(\boldsymbol{\xi}) \to p(\boldsymbol{\xi}) \to \{p(\boldsymbol{\xi})|_{x=x_j}\}_{j=1}^{k} =: \boldsymbol{p}^{\text{obs}}.$$

- Gaussian random field in Hilbert space $\mathcal{H}$, e.g., $\mathcal{H} = L^2(D)$:

$$\kappa(x,\omega) = \hat{\kappa}(x) + \sum_{m\geq 1} \xi_m(\omega)\, \phi_m(x), \qquad \{\phi_m\}_{m=1}^{\infty} \text{ CONS of } \mathcal{H}.$$

- Direct measurements of κ used to fit Gaussian prior μ_0 for κ resp. $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} \sim N(0, C_0) =: \mu_0 \qquad \text{on } \ell^2(\mathbb{N}).$$

- Parameter-to-observable map

$$G : \ell^2(\mathbb{N}) \to \mathbb{R}^J, \qquad \boldsymbol{\xi} \to \kappa(\boldsymbol{\xi}) \to p(\boldsymbol{\xi}) \to \{p(\boldsymbol{\xi})|_{x=x_j}\}_{j=1}^k =: \boldsymbol{p}^{\text{obs}}.$$

- Gaussian random field in Hilbert space $\mathcal{H}$, e.g., $\mathcal{H} = L^2(D)$:

$$\kappa(x,\omega) = \hat{\kappa}(x) + \sum_{m\geq 1} \xi_m(\omega)\, \phi_m(x), \qquad \{\phi_m\}_{m=1}^\infty \text{ CONS of } \mathcal{H}.$$

- Direct measurements of κ used to fit Gaussian prior μ_0 for κ resp. $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} \sim N(0, C_0) =: \mu_0 \qquad \text{on } \ell^2(\mathbb{N}).$$

- Merge indirect data $\boldsymbol{p}^{\text{obs}}$ by conditioning prior $\boldsymbol{\xi} \sim \mu_0$ on

$$\boldsymbol{p}^{\text{obs}} = G(\boldsymbol{\xi}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, \Sigma) \text{ Gaussian noise.}$$

Method of choice in Bayesian inference: Markov chain Monte Carlo sampling

- Construct a Markov chain $\{\boldsymbol{\xi}_n(\omega)\}_{n\in\mathbb{N}}$ with $\mathbf{P}\big(\boldsymbol{\xi} \in \cdot \,|\, p(\boldsymbol{x}_j) = p_j \;\forall j\big)$ as its limiting (invariant) distribution.
- Simple method: Metropolis-Hastings update
- Dimension-independent variants
- Let the chain run long enough (to converge) and take samples along the path $\boldsymbol{\xi}_n$ for Monte Carlo.

Need to evaluate $p(\boldsymbol{x}_j, \boldsymbol{\xi})$ resp. solve (PDE-mixed) many, many times (≈500,000) due to burn-in and autocorrelation.

Sampling-free alternatives: Filtering methods (EnKF, PC + EnKF), may be arbitarily wrong [Ernst, Sprungk & Starkloff, 2015].

Markov chain $(\boldsymbol{\xi}_n)_{n\in\mathbb{N}}$ in ℓ^2 with transition kernel

$$Q(\boldsymbol{\eta}, A) := \mathbf{P}(\boldsymbol{\xi}_{n+1} \in A | \boldsymbol{\xi}_n = \boldsymbol{\eta}), \qquad A \in \mathscr{B}(\ell^2)$$

which is reversible w.r.t. μ:

$$Q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta})\, \mu(\mathrm{d}\boldsymbol{\xi}) = Q(\boldsymbol{\eta}, \mathrm{d}\boldsymbol{\xi})\, \mu(\mathrm{d}\boldsymbol{\eta}) \qquad \Rightarrow \quad \mu = \mu Q.$$

Markov chain $(\boldsymbol{\xi}_n)_{n\in\mathbb{N}}$ in ℓ^2 with transition kernel

$$Q(\boldsymbol{\eta}, A) := \mathbf{P}(\boldsymbol{\xi}_{n+1} \in A | \boldsymbol{\xi}_n = \boldsymbol{\eta}), \qquad A \in \mathscr{B}(\ell^2)$$

which is reversible w.r.t. μ:

$$Q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta})\, \mu(\mathrm{d}\boldsymbol{\xi}) = Q(\boldsymbol{\eta}, \mathrm{d}\boldsymbol{\xi})\, \mu(\mathrm{d}\boldsymbol{\eta}) \qquad \Rightarrow \qquad \mu = \mu Q.$$

Then – under suitable conditions – we have for QoI t_{exit}

$$\frac{1}{N}\sum_{n=1}^{N} t_{\text{exit}}(\boldsymbol{\xi}_n) \quad \xrightarrow{N\to\infty} \quad \int t_{\text{exit}}(\boldsymbol{\xi})\, \mu(\mathrm{d}\boldsymbol{\xi}) = \mathbf{E}_\mu\left[t_{\text{exit}}\right].$$

Mean squared error $\propto N^{-\frac{1}{2}}$, constant is sum of autocovariances:

$$\sum_{k=-\infty}^{\infty} \gamma(k), \qquad \gamma(k) = \mathbf{Cov}\left(t_{\text{exit}}(\boldsymbol{\xi}_1), t_{\text{exit}}(\boldsymbol{\xi}_{1+k})\right), \qquad \boldsymbol{\xi}_1 \sim \mu.$$

Rapid decay of autocovariance function $\gamma \quad \Rightarrow \quad$ high statistical efficiency.

Metropolis-Hastings (MH) MCMC where $\boldsymbol{\xi}_n \to \boldsymbol{\xi}_{n+1}$ is as follows:

1. Propose new state $\boldsymbol{\eta}$ according to proposal kernel $q(\boldsymbol{\xi}_n, \mathrm{d}\boldsymbol{\eta})$, e.g.,

$$\boldsymbol{\eta} \sim q(\boldsymbol{\xi}_n, \cdot) = N(\boldsymbol{\xi}_n, s^2 C_0), \qquad s \in \mathbb{R}_+ \text{ stepsize.}$$

2. Accept proposal $\boldsymbol{\eta}$ with probability $\alpha(\boldsymbol{\xi}_n, \boldsymbol{\eta})$: draw $a \sim \mathrm{U}[0,1]$ and set

$$\boldsymbol{\xi}_{n+1} = \begin{cases} \boldsymbol{\eta}, & a \le \alpha(\boldsymbol{\xi}_n, \boldsymbol{\eta}), \\ \boldsymbol{\xi}_n, & \text{otherwise.} \end{cases}$$

Metropolis-Hastings (MH) MCMC where $\boldsymbol{\xi}_n \to \boldsymbol{\xi}_{n+1}$ is as follows:

1. Propose new state $\boldsymbol{\eta}$ according to proposal kernel $q(\boldsymbol{\xi}_n, \mathrm{d}\boldsymbol{\eta})$, e.g.,

$$\boldsymbol{\eta} \sim q(\boldsymbol{\xi}_n, \cdot) = N(\boldsymbol{\xi}_n, s^2 C_0), \qquad s \in \mathbb{R}_+ \text{ stepsize.}$$

2. Accept proposal $\boldsymbol{\eta}$ with probability $\alpha(\boldsymbol{\xi}_n, \boldsymbol{\eta})$: draw $a \sim \mathrm{U}[0,1]$ and set

$$\boldsymbol{\xi}_{n+1} = \begin{cases} \boldsymbol{\eta}, & a \leq \alpha(\boldsymbol{\xi}_n, \boldsymbol{\eta}), \\ \boldsymbol{\xi}_n, & \text{otherwise.} \end{cases}$$

Resulting transition kernel of MH chain:

$$Q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) = \alpha(\boldsymbol{\xi}, \boldsymbol{\eta}) q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) + \underbrace{\left[1 - \int \alpha(\boldsymbol{\xi}, \zeta)\, q(\boldsymbol{\xi}, \mathrm{d}\zeta)\right]}_{= \text{ rejection probability}} \delta_{\boldsymbol{\xi}}(\mathrm{d}\boldsymbol{\eta}),$$

Sufficient for reversibility w.r.t. μ is the choice

$$\alpha(\boldsymbol{\xi}_k, \boldsymbol{\eta}) = \min\left\{1, \frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(\boldsymbol{\xi}_k, \boldsymbol{\eta})\right\},$$

where $\nu(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) := q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta})\, \mu(\mathrm{d}\boldsymbol{\xi}), \qquad \nu^\top(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) := \nu(\mathrm{d}\boldsymbol{\eta}, \mathrm{d}\boldsymbol{\xi}).$

Sufficient for reversibility w.r.t. μ is the choice

$$\alpha(\boldsymbol{\xi}_k, \boldsymbol{\eta}) = \min\left\{1, \frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(\boldsymbol{\xi}_k, \boldsymbol{\eta})\right\},$$

where $\qquad \nu(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) := q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta})\,\mu(\mathrm{d}\boldsymbol{\xi}), \qquad \nu^\top(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) := \nu(\mathrm{d}\boldsymbol{\eta}, \mathrm{d}\boldsymbol{\xi}).$

In finite dimensions $\frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}$ is simply ratio of densities (w.r.t. Lebesgue measure).
E.g., if q has density $\rho(|\boldsymbol{\xi} - \boldsymbol{\eta}|)$, then

$$\frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{\pi(\boldsymbol{\eta})}{\pi(\boldsymbol{\xi})}, \qquad \mu(\mathrm{d}\boldsymbol{\xi}) \propto \pi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}.$$

Sufficient for reversibility w.r.t. μ is the choice

$$\alpha(\boldsymbol{\xi}_k, \boldsymbol{\eta}) = \min\left\{1, \frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(\boldsymbol{\xi}_k, \boldsymbol{\eta})\right\},$$

where $\qquad \nu(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) := q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta})\, \mu(\mathrm{d}\boldsymbol{\xi}), \qquad \nu^\top(\mathrm{d}\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}) := \nu(\mathrm{d}\boldsymbol{\eta}, \mathrm{d}\boldsymbol{\xi}).$

In finite dimensions $\frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}$ is simply ratio of densities (w.r.t. Lebesgue measure). E.g., if q has density $\rho(|\boldsymbol{\xi} - \boldsymbol{\eta}|)$, then

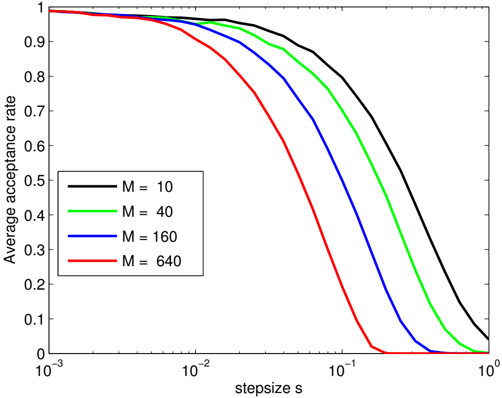
$$\frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{\pi(\boldsymbol{\eta})}{\pi(\boldsymbol{\xi})}, \qquad \mu(\mathrm{d}\boldsymbol{\xi}) \propto \pi(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi}.$$

In infinite dimensions, μ_0-reversibility of proposal q sufficient in order that

$$\frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(\boldsymbol{\xi}, \boldsymbol{\eta}) \text{ exists} \qquad \text{and} \qquad \boxed{\frac{\mathrm{d}\nu^\top}{\mathrm{d}\nu}(\boldsymbol{\xi}, \boldsymbol{\eta}) = \frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(\boldsymbol{\eta}) \left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(\boldsymbol{\xi})\right)^{-1} = \mathrm{e}^{\Phi(\boldsymbol{\xi}) - \Phi(\boldsymbol{\eta})}.}$$

Example: 2D groundwater flow model, synthetic data.
Acceptance rate vs. stepsize for increasing dimension M of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)$.

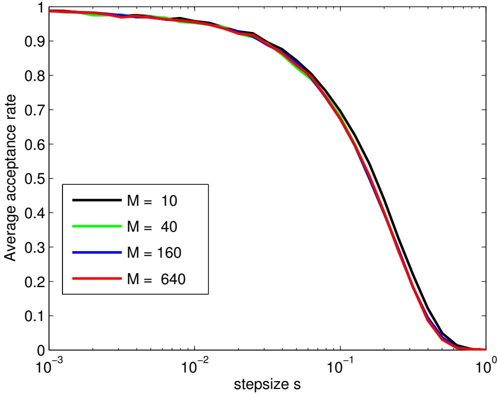
Random walk-proposal $q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2 C_0)$



average acceptance rate: $\bar{\alpha} = \mathbf{E}_\nu \left[\alpha(\boldsymbol{\xi}, \boldsymbol{\eta})\right]$

Example: 2D groundwater flow model, synthetic data.
Acceptance rate vs. stepsize for increasing dimension M of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_M)$.

pCN-proposal $q(\boldsymbol{\xi}) = N(\sqrt{1-s^2}\boldsymbol{\xi}, s^2 C_0)$



introduced in [Cotter, Roberts, Stuart & White, 2013]

Contents

4 An Inverse Problem for Groundwater Flow at WIPP

4.1 Gaussian Random Fields

4.2 Gaussian Random Field Models from Direct Observations

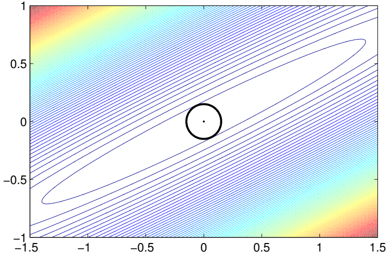
4.3 Solution of the Forward Problem

4.4 Bayesian Inversion

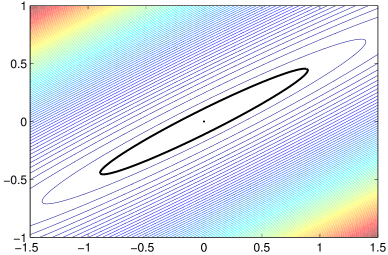
4.5 Improving MCMC Proposals in Hilbert Space

4.6 Numerical Results: Bayesian Inversion for WIPP

Example: $\mu = N(\mathbf{0}, \boldsymbol{C})$ in 2D, different Random Walk proposals

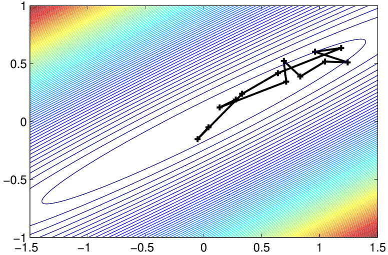


$$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2 \boldsymbol{I})$$

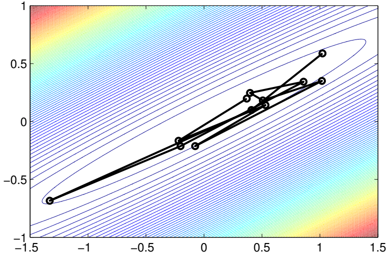


$$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2 \boldsymbol{C})$$

Example: $\mu = N(\mathbf{0}, \boldsymbol{C})$ in 2D, different Random Walk proposals

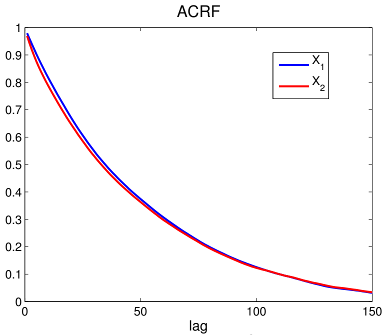


$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2\boldsymbol{I})$

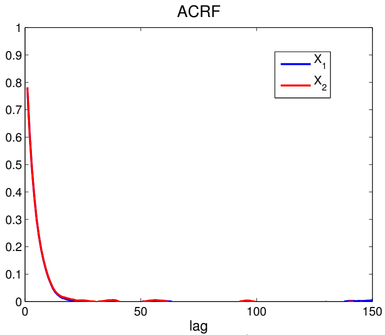


$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2\boldsymbol{C})$

Example: $\mu = N(\mathbf{0}, \boldsymbol{C})$ in 2D, different Random Walk proposals

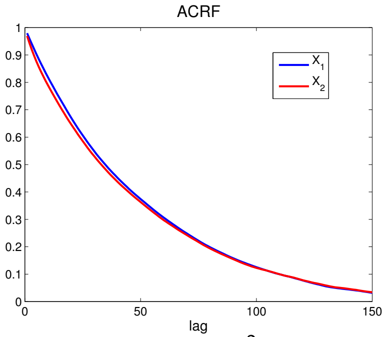


$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2\boldsymbol{I})$

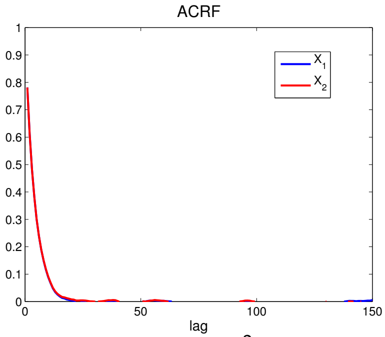


$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2\boldsymbol{C})$

Example: $\mu = N(\mathbf{0}, \boldsymbol{C})$ in 2D, different Random Walk proposals



$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2 \boldsymbol{I})$



$q(\boldsymbol{\xi}) = N(\boldsymbol{\xi}, s^2 \boldsymbol{C})$

Higher statistical efficiency of proposal with same covariance as μ shown in [Roberts & Rosenthal, 2001].

Improving MCMC Proposals in Hilbert Space

Gauss-Newton type approximation of posterior covariance

If forward map $G : \mathscr{H} \to \mathbb{R}^d$ were linear, $\mu_0 = N(0, C_0)$ and $\varepsilon \sim N(0, \Sigma)$, then

$$\mu = N(m, C), \qquad C = (C_0^{-1} + G^* \Sigma^{-1} G)^{-1}.$$

If forward map $G : \mathscr{H} \to \mathbb{R}^d$ were linear, $\mu_0 = N(0, C_0)$ and $\varepsilon \sim N(0, \Sigma)$, then

$$\mu = N(m, C), \qquad C = (C_0^{-1} + G^* \Sigma^{-1} G)^{-1}.$$

Idea: Gauss-Newton-type linear approximation of nonlinear G

$$G(\boldsymbol{\xi}) \approx \widetilde{G}(\boldsymbol{\xi}) := G(\boldsymbol{\xi}_0) + L\boldsymbol{\xi}, \qquad L = \nabla G(\boldsymbol{\xi}_0)$$

and use

$$C \approx \widetilde{C} = (C_0^{-1} + L^* \Sigma^{-1} L)^{-1}$$

as proposal covariance.

Improving MCMC Proposals in Hilbert Space

Gauss-Newton type approximation of posterior covariance

If forward map $G : \mathscr{H} \to \mathbb{R}^d$ were linear, $\mu_0 = N(0, C_0)$ and $\varepsilon \sim N(0, \Sigma)$, then

$$\mu = N(m, C), \qquad C = (C_0^{-1} + G^* \Sigma^{-1} G)^{-1}.$$

Idea: Gauss-Newton-type linear approximation of nonlinear G

$$G(\boldsymbol{\xi}) \approx \widetilde{G}(\boldsymbol{\xi}) := G(\boldsymbol{\xi}_0) + L\boldsymbol{\xi}, \qquad L = \nabla G(\boldsymbol{\xi}_0)$$

and use

$$C \approx \widetilde{C} = (C_0^{-1} + L^* \Sigma^{-1} L)^{-1}$$

as proposal covariance.

Good choice for $\boldsymbol{\xi}_0$ might be the maximum a posteriori estimator:

$$\boldsymbol{\xi}_{\text{MAP}} = \arg\min_{\boldsymbol{\xi}} \left(\Phi(\boldsymbol{\xi}) + \| C_0^{-1/2} \boldsymbol{\xi} \|^2 \right).$$

In place of prior covariance C_0, use approximated posterior covariance

$$\widetilde{C} = (C_0^{-1} + \Gamma)^{-1}, \qquad \Gamma \text{ positive, self-adjoint, bounded (otherwise arbitrary)},$$

for a Random Walk-like proposal kernel

$$\widetilde{q}_s(u) = N(P_s u, s^2 \widetilde{C}).$$

In place of prior covariance C_0, use approximated posterior covariance

$$\widetilde{C} = (C_0^{-1} + \Gamma)^{-1}, \qquad \Gamma \text{ positive, self-adjoint, bounded (otherwise arbitrary)},$$

for a Random Walk-like proposal kernel

$$\widetilde{q}_s(u) = N(P_s u, s^2 \widetilde{C}).$$

Enforcing reversibility of kernel $\widetilde{q}_s$ w.r.t. μ_0 – as for pCN-proposal – yields

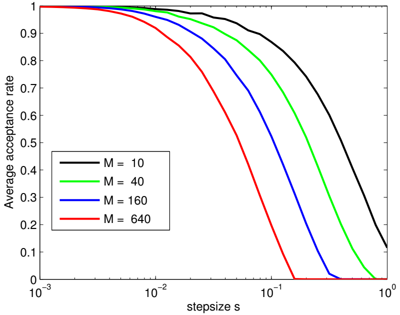
$$P_s = C_0^{1/2} \sqrt{I - s^2(I + H)^{-1}}\, C_0^{-1/2}, \qquad H := C_0^{1/2} \Gamma C_0^{1/2}.$$

We call $\widetilde{q}_s$ Gauss-Newton pCN-proposal (GNpCN), [Ernst & Sprungk, 2015].

Related approaches: [Law, 2013], [Cui, Law & Marzouk, 2014]

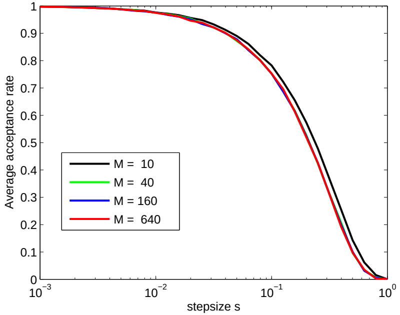
Improving MCMC Proposals in Hilbert Space
Random Walk vs. GNpCN

Same example as for pCN, $\widetilde{C} = (C_0^{-1} + L^*\Sigma^{-1}L)^{-1}$ where $L = \nabla G(\xi_{\mathrm{MAP}})$.



Anisotropic Random Walk

$\widetilde{q}_s(u) = N(u, s^2 \widetilde{C})$



GNpCN

$\widetilde{q}_s(u) = N(P_s u, s^2 \widetilde{C})$

- Markov operator $Q : L^2_\mu(\mathscr{H}) \to L^2_\mu(\mathscr{H})$ associated with kernel $Q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta})$:

$$Qf(\boldsymbol{\xi}) := \int_{\mathscr{H}} f(\boldsymbol{\eta})\, Q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}), \qquad f \in L^2_\mu(\mathscr{H}).$$

- Existence of an L^2-spectral gap of operator Q

$$0 < \operatorname{gap}(Q) = 1 - \|Q - \mathbf{E}_\mu\|_{L^2_\mu \to L^2_\mu}$$

implies geometric ergodicity/convergence to μ in total variation norm

$$\|\mu - \mu_0 Q^n\|_{\mathrm{TV}} \le C \exp(-r\, n).$$

Improving MCMC Proposals in Hilbert Space

Convergence

- Markov operator $Q : L^2_\mu(\mathscr{H}) \to L^2_\mu(\mathscr{H})$ associated with kernel $Q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta})$:

$$Qf(\boldsymbol{\xi}) := \int_{\mathscr{H}} f(\boldsymbol{\eta})\, Q(\boldsymbol{\xi}, \mathrm{d}\boldsymbol{\eta}), \qquad f \in L^2_\mu(\mathscr{H}).$$

- Existence of an L^2-spectral gap of operator Q

$$0 < \mathrm{gap}(Q) = 1 - \|Q - \mathbf{E}_\mu\|_{L^2_\mu \to L^2_\mu}$$

implies geometric ergodicity/convergence to μ in total variation norm

$$\|\mu - \mu_0 Q^n\|_{\mathrm{TV}} \le C \exp(-r\, n).$$

- For the pCN-proposal a (dimension-independent) L^2-spectral gap was proven under certain conditions on Φ in [Hairer, Stuart & Vollmer, 2014]
- [Rudolf & Sprungk, 2015]: derive spectral gap for GNpCN from that of pCN.

Contents

4 An Inverse Problem for Groundwater Flow at WIPP

4.1 Gaussian Random Fields

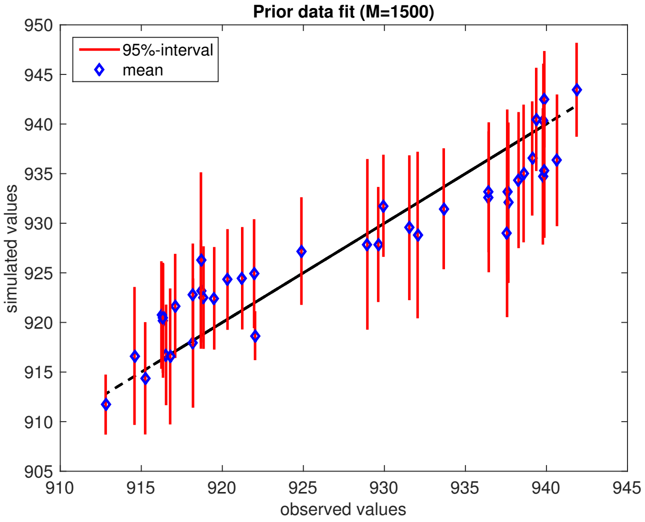
4.2 Gaussian Random Field Models from Direct Observations

4.3 Solution of the Forward Problem

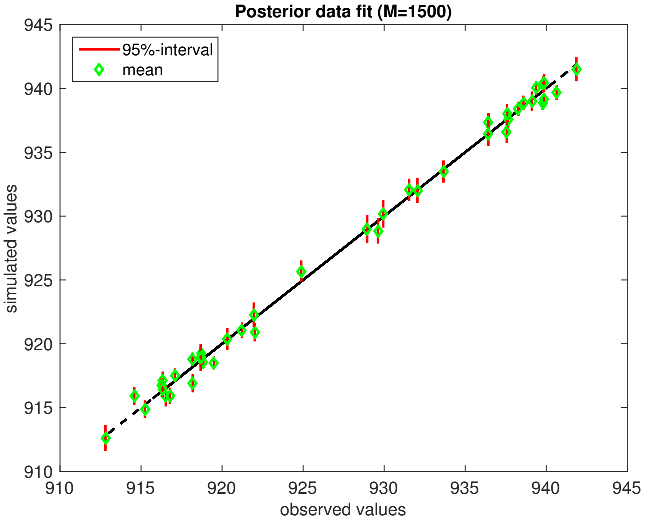
4.4 Bayesian Inversion

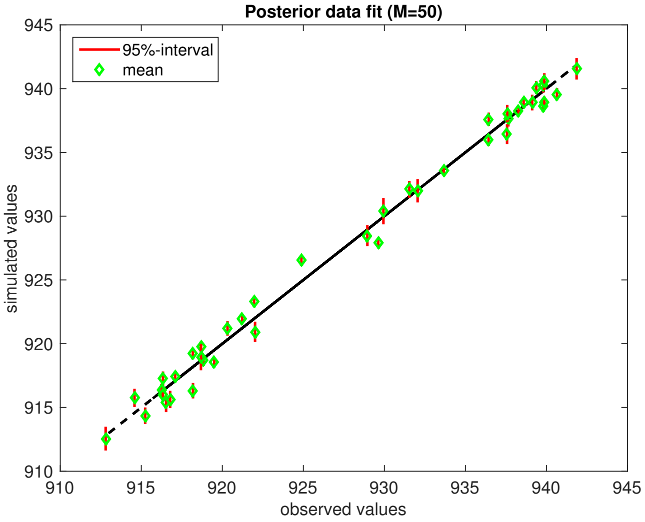
4.5 Improving MCMC Proposals in Hilbert Space

4.6 Numerical Results: Bayesian Inversion for WIPP

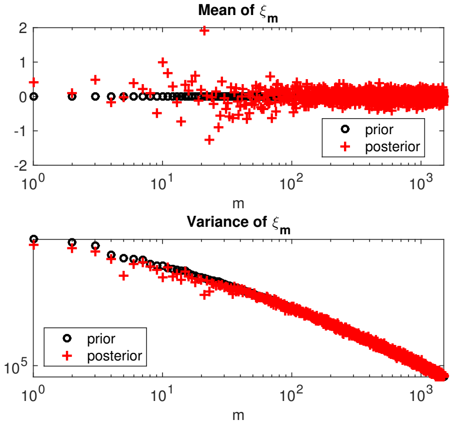


Prior data fit (M=1500)
95%-interval
mean
simulated values
observed values
950
945
940
935
930
925
920
915
910
905
910
915
920
925
930
935
940
945


Posterior data fit (M=1500)
95%-interval
mean
simulated values
observed values
945
940
935
930
925
920
915
910
910
915
920
925
930
935
940
945

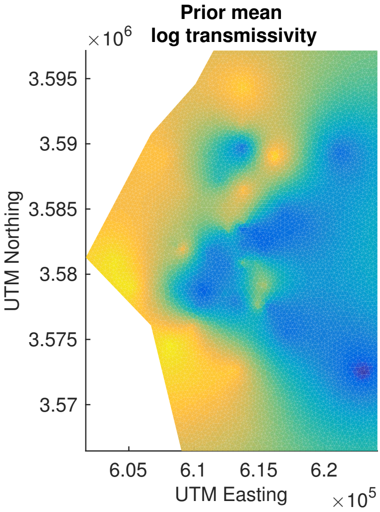


Posterior data fit (M=50)
95%-interval
mean
simulated values
observed values
910
915
920
925
930
935
940
945

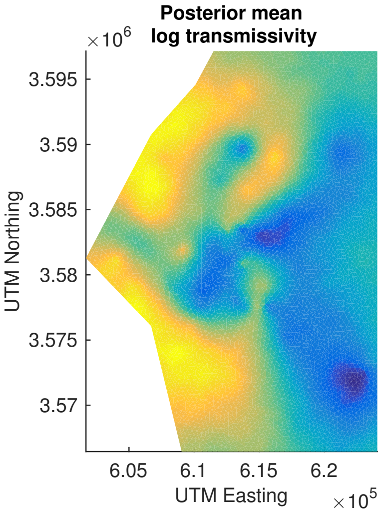


Significant changes only in a few ξ_m

$\Rightarrow$ Run chain only in smaller subspace and employ surrogates for solving PDE

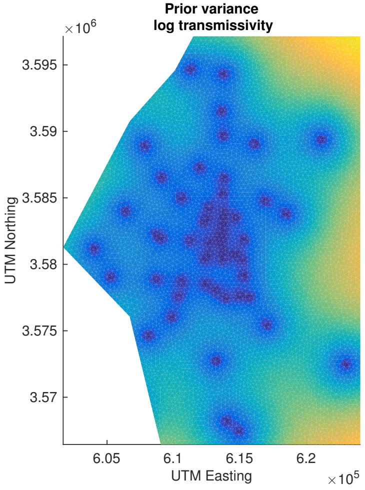


Prior mean
log transmissivity
×10^6
3.595
3.59
3.585
3.58
3.575
3.57
UTM Northing
6.05
6.1
6.15
6.2
UTM Easting
×10^5

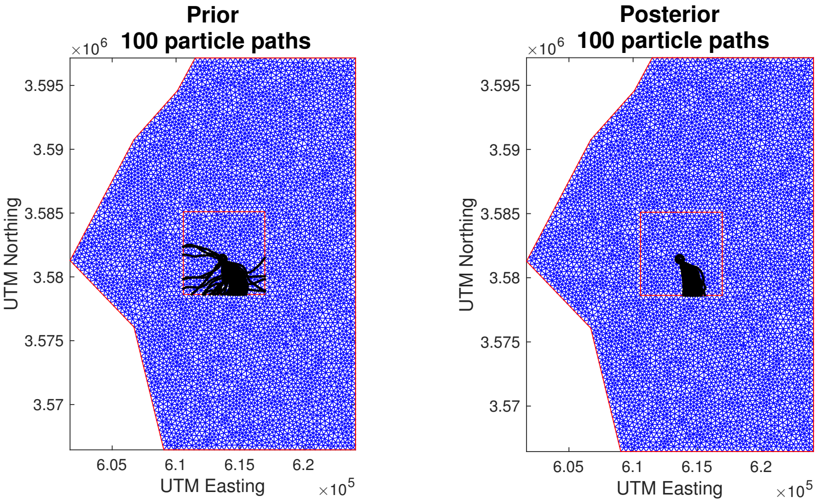


Posterior mean
log transmissivity
×10^6
3.595
3.59
3.585
3.58
3.575
3.57
UTM Northing
6.05
6.1
6.15
6.2
UTM Easting
×10^5

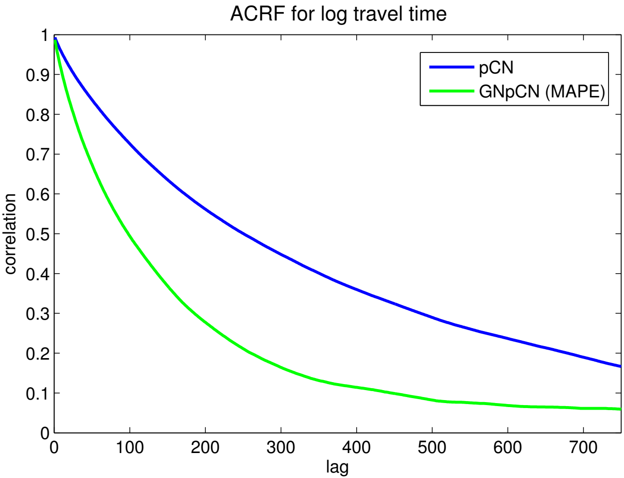
Bayesian Inversion
Results for WIPP



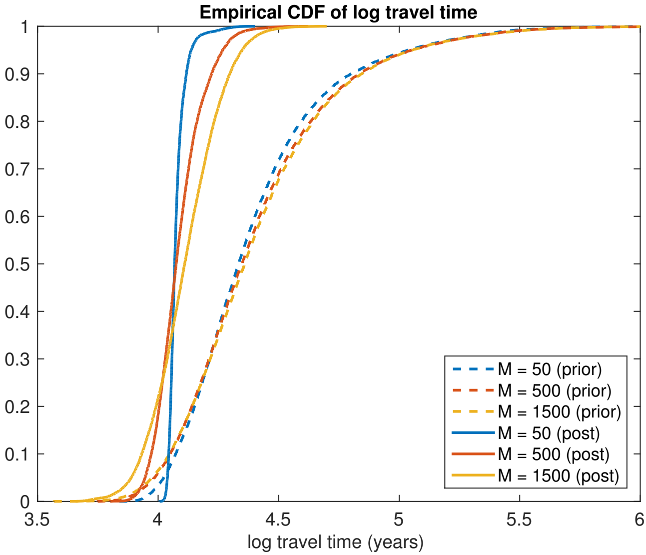
Bayesian Inversion
Results for WIPP



ACRF in QoI for pCN (blue) and GNpCN (green)



Bayesian Inversion
Results for WIPP



Summary

- WIPP case study: inverse problem for hydraulic conductivity and particle travel time.
- Estimation-based methods much improved by Bayesian inference on indirect observations.
- Method: MCMC in high (infinite-)dimensional space.

References I

Bertrand Russell.
The Problems of Philosophy.
Williams & Norgate, London, 1912.

Prakash Gorroochurn.
Classic Problems of Probability.
John Wiley & Sons, Hoboken, NJ, 2012.

Nassim Nicholas Taleb and Mark Blythe.
The Black Swan of Cairo: How suppressing volatility makes the world less predictable and more dangerous.
Foreign Affairs, 90(3):33–39, 2011.

Gerd Gigerenzer.
Risk Savvy: How to Make Good Decisions.
Viking, New York, 2014.

Colin Fox, Geoff K. Nicholls, and Sze M. Tan.
Lecture Notes, ELEC445 (Inverse Problems), University of Otago, 2016.

Gareth Roberts.
Lecture Notes: ST911 Fundamentals of Statistical Inference Part III, University of Warwick, 2012.

Andrew M. Stuart.
Inverse problems: a Bayesian perspective.
Acta Numerica, 19:451–559, 2010.

References II

Oliver G. Ernst, Björn Sprungk, and Hans-Jörg Starkloff.
Analysis of the ensemble and polynomial chaos Kalman filters in Bayesian inverse problems.
SIAM/ASA J. Uncertainty Quantification, 3(1):823–851, 2015.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer.
Regularization of Inverse Problems.
Springer Science & Business Media, 1996.

Bernd Hofmann.
On smoothness concepts in regularization for nonlinear inverse problems in Banach spaces.
In Roderick Melnik, editor, *Mathematical and Computational Modeling*, chapter 8, pages 192–221.
John Wiley & Sons, 2015.

Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer.
Spectral gaps for a Metropolis-Hastings algorithm in infinite dimensions.
Annals of Probability, 24(6):2455–2490, 2014.

Daniel Rudolf and Björn Sprungk.
On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm.
FOCM, 2016.

T. L. Cauffman, A. M. LaVenue, and J. P. McCord.
Ground-Water Flow Modeling of the Culebra Dolomite. Volume II: Data Base.
INTERA Inc. Contractor ReportContractor Report SAND89-7068/2, Sandia National Laboratories, Albuquerque, NM, 1990.