Illustrations of MCMC in Nonlinear Regression, Survival Analysis and Spatial Statistics

Ian McKeague
Department of Statistics
Florida State University

Eighth Summer Workshop
New Zealand Mathematics Research Institute
Napier, January 2002

Outline

- Brief introduction to MCMC
- Single-index models
- Application to climate prediction
- Bayesian inversion of ocean circulation data
- Bayesian hazard function estimation
- Point processes and disease clustering

BRIEF INTRODUCTION TO MCMC

If a distribution π is defined indirectly, it may be difficult to calculate the expectation

$$E_\pi f = \int f(x)\, \pi(dx)$$

analytically or even via classical Monte Carlo.

The idea behind the Markov chain Monte Carlo (MCMC) method is to generate n iterations of a Markov chain $\{X_i\}$ that has π as invariant law, and then approximate $E_\pi f$ by the ergodic average:

$$E_\pi f \approx \frac{1}{n} \sum_{i=1}^{n} f(X_i)$$

Important to discard early iterations (burn-in).

Typical MCMC samplers: Metropolis–Hastings, Gibbs

Metropolis–Hastings

Current state x, simulating a density π

Propose a candidate y from the *proposal* density $q(x, \cdot)$.

Acceptance probability

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

Easily checked that the transition distribution P satisfies detailed balance:

$$P(x, dy)\pi(dx) = \pi(dy)P(y, dx)$$

so π is an invariant distribution for P:

$$\begin{aligned}
\int P(x, A)\pi(dx) &= \int\int_A P(x, dy)\pi(dx) \\
&= \int\int_A \pi(dy)P(y, dx) \\
&= \int_A\int P(y, dx)\pi(dy) \\
&= \pi(A)
\end{aligned}$$

Metropolis

Symmetric proposal: $q(x, y) = q(y, x)$

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

Random-walk Metropolis

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

Gibbs

State space a finite product space.
Conditional density of ith component
(local characteristic):

$$q(x, y) = \pi(y_i | x_{-i})$$

Acceptance probability = 1 in this case.

SINGLE-INDEX MODELS

Flexible alternative to standard linear regression. Response Y_i allowed to be an arbitrary function f of a finite linear combination of predictors:

$$E(Y_i|X_i) = f(X_i'\boldsymbol{\theta}),\ i = 1, \ldots, n.$$

d-dimensional vector $\boldsymbol{\theta}$ (index vector) identifiable up to a constant factor

Link function f is an infinite-dimensional nuisance parameter

Background

Friedman and Stuetzle's (1981) projection pursuit regression

Extensive applications in econometrics

Weighted average derivative estimation

Stoker (1986), Powell, Stock and Stoker (1989)

Exploits the fact that $\boldsymbol{\theta}$ is proportional to the expected value of a weighted gradient of the regression function $r(x) = E(Y|X = x)$:

$$\delta \equiv E\left(p(X)\frac{\partial r}{\partial x}(X)\right) \propto \boldsymbol{\theta}.$$

Taking $p(x)$ as the density of X, integration by parts gives:

$$\begin{aligned}
\delta &= \int p(x)^2 \frac{\partial r}{\partial x}(x)\, dx \\
&= -2\int r(x)p(x)\frac{\partial p}{\partial x}(x)\, dx \\
&= -2E\left(r(X)\frac{\partial p}{\partial x}(X)\right) \\
&= -2E\left(Y\frac{\partial p}{\partial x}(X)\right).
\end{aligned}$$

Leads to $\sqrt{n}$-consistent estimators of the index vector.

Härdle and Stoker (1989), Härdle and Tsybakov (1993), Härdle, Marron and Tsybakov (1992), Samarov (1993), Hristache, Juditsky and Spokoiny (2001, Ann. Statist.)

Fails with high-dimensional predictors.

Semiparametric M-estimation

Least squares version:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\theta} \sum_{i=1}^{n} \left(Y_i - \widehat{f}_{\theta}(X_i'\boldsymbol{\theta})\right)^2,$$

$\widehat{f}_{\theta}$ a nonparametric estimator of

$$f_{\theta}(\cdot) = E(Y|X'\boldsymbol{\theta} = \cdot).$$

High-dimensional maximization problem.

In some cases the estimators are shown to be $\sqrt{n}$-consistent and asymptotically efficient.

Klein and Spady (1993), Ichimura (1993), Horowitz and Härdle (1996), Delecroix and Hristache (1999), Härdle, Hall and Ichimura (1993).

Bayesian approach

Antoniadis, Gregoire and McKeague (2001)

Offers the hope of more stable estimates, especially for small or moderate data sets with low signal-to-noise ratio.

Other recent Bayesian approaches in nonlinear regression: additive semiparametric regression (Biller, 2000), semiparametric hazard function regression (McKeague and Tighiouart, 2000), nonparametric regression with measurement error (Berry et al., 2000), and generalized linear models with parametric link function determination (Ntzoufras, Dellaportas and Forster, 2000).

Single-index model:

$$Y_i = f(X_i'\boldsymbol{\theta}) + \epsilon_i, \quad i = 1, \ldots, n$$

$\epsilon_i \sim$ iid $N(0, \sigma^2)$, $\boldsymbol{\theta} \in \mathbb{R}^d$, $||\boldsymbol{\theta}|| = 1$.

Hierarchical prior for $(\boldsymbol{\theta}, \sigma^2, f)$

First level: $\boldsymbol{\theta}$, σ^2, $\boldsymbol{\theta} \perp \sigma^2$
Second level: $f|\boldsymbol{\theta}, \sigma^2$

$\boldsymbol{\theta} \sim$ Fisher–von Mises:

$$p(\boldsymbol{\theta}) \propto \exp(\lambda_{\text{prior}} \boldsymbol{\theta}' \boldsymbol{\theta}_{\text{prior}}), \quad ||\boldsymbol{\theta}|| = 1$$

conc. par. $\lambda_{\text{prior}} \geq 0$, modal direction $\boldsymbol{\theta}_{\text{prior}}$.

$\sigma^2 \sim$ inverse-gamma:

$$p(\sigma^2) \sim \sigma^{-2(A+1)} \exp(-B^{-1}\sigma^{-2}), \quad \sigma^2 > 0,$$

$A > 0$, $B > 0$ tuning constants.

f represented by a random linear combination of B-spline basis functions:

$$f(t) = \sum_{j=1}^{K} \beta_j B_j(t).$$

Recast as a linear model:

$$\mathbf{Y} = \tilde{X}_\theta \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$\mathbf{Y} = (Y_1, \ldots, Y_n)'$
$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$
$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$
$\tilde{X}_\theta = B(X\boldsymbol{\theta})$
$X = (X_1, \ldots, X_n)'$
$B(\mathbf{t})_{ij} = B_j(t_i)$

B_j the jth B-spline basis function of degree q based on $m+1$ equispaced knots over $[a_\theta, b_\theta]$, $K = m + q$, $a_\theta = \min\{X_i'\boldsymbol{\theta}, i = 1, \ldots, n\} - \delta$, $b_\theta = \max\{X_i'\boldsymbol{\theta}, i = 1, \ldots, n\} + \delta$ for some $\delta > 0$.

B-splines provide an attractive system of basis functions and allow $\tilde{X}_\theta$ to be computed quickly (needed at each step of the MCMC).

Zellner's (1986) conjugate g-prior:

$$\boldsymbol{\beta} \sim N\left(\tilde{\boldsymbol{\beta}}_\theta, \sigma^2(\tilde{X}_\theta'\tilde{X}_\theta)^{-1}\right)$$

$\tilde{\boldsymbol{\beta}}_\theta = (\tilde{X}_\theta'\tilde{X}_\theta)^{-1}\tilde{X}_\theta'\mathbf{Y}$ least squares estimate.

Posterior

$$p(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta}|X, \mathbf{Y}) \;\; \propto \;\; p(\mathbf{Y}|X, \boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta}) p(\boldsymbol{\theta}, \sigma^2, \boldsymbol{\beta})$$

Integrate out $\sigma^2, \boldsymbol{\beta}$:

$$p(\boldsymbol{\theta}|X, \mathbf{Y}) \;\; \propto \;\; \left(S(\boldsymbol{\theta}) + \frac{2}{B} \right)^{-(A+\frac{n}{2})} \exp(\lambda_{\text{prior}} \boldsymbol{\theta}' \boldsymbol{\theta}_{\text{prior}})$$

$S(\boldsymbol{\theta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\tilde{X}_\theta (\tilde{X}_\theta' \tilde{X}_\theta)^{-1} \tilde{X}_\theta' \mathbf{Y}$ error sum of squares

Estimates

$\widehat{\boldsymbol{\theta}} =$ (normalized) posterior mean
$\widehat{f}(t) = \sum_{j=1}^{K} \tilde{\boldsymbol{\beta}}_{\widehat{\theta},j} B_j(t)$

Problem: severe over-fitting and high bias; no variable selection or regularization feature built into the prior for f.

Could modify the prior for $\boldsymbol{\beta}$: Smith and Kohn's (1996) *Bayesian variable selection*: β_j replaced by $\beta_j \epsilon_j$ with $\epsilon_j \sim$ iid Bernoulli(1/2).

No longer have a conjugate prior. Slows down the MCMC too much.

Regularized posterior

Replace $S(\boldsymbol{\theta})$ by

$$S_\rho(\boldsymbol{\theta}) = \mathbf{Y}'\mathbf{Y} - Y'\tilde{X}_\theta(\tilde{X}_\theta'\tilde{X}_\theta + \rho I)^{-1}\tilde{X}_\theta'\mathbf{Y}$$

$\rho \geq 0$ controls the penalty, $I = K \times K$ identity matrix.

Penalized least squares estimate:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_\theta &= (\tilde{X}_\theta'\tilde{X}_\theta + \rho I)^{-1}\tilde{X}_\theta'\mathbf{Y} \\ &= \arg\min_{\boldsymbol{\beta}}(||\mathbf{Y} - \tilde{X}_\theta\boldsymbol{\beta}||^2 + \rho||\boldsymbol{\beta}||^2) \end{aligned}$$

MCMC Algorithm

Random walk Metropolis on rotations in $\mathbb{R}^d$

Proposal distribution: Fisher–von Mises with modal vector given by the current value of $\boldsymbol{\theta}$.

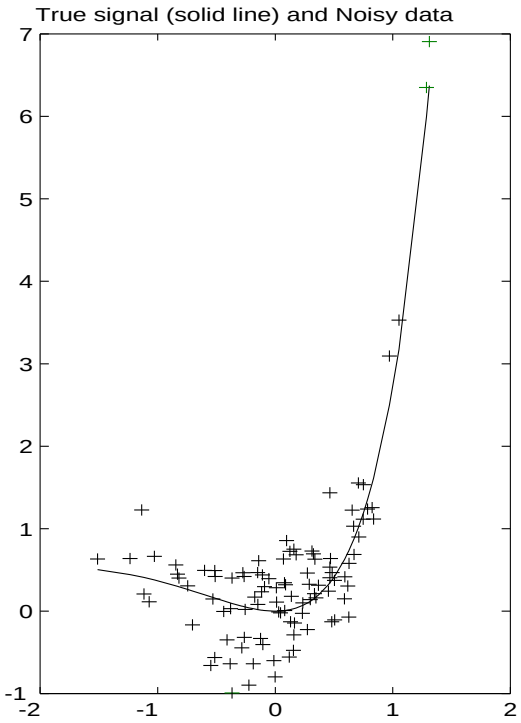
Start algorithm at the Hristache–Juditsky–Spokoiny (HJS) estimate.

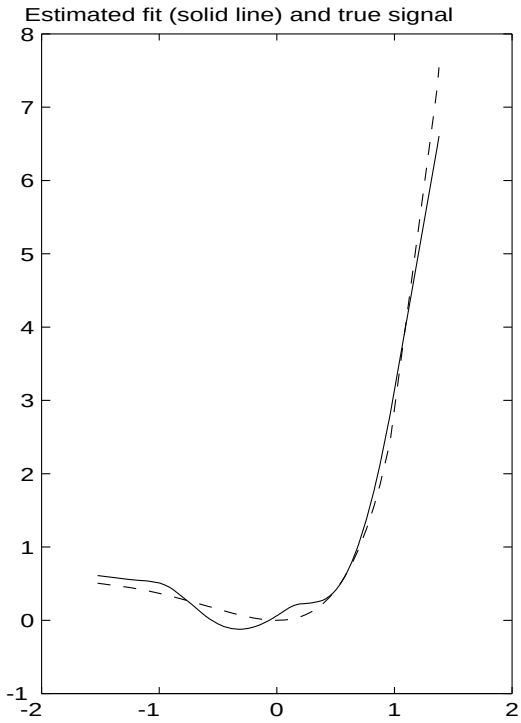
Ulrich (1984) gave a simple generator for Fisher–von Mises distributed random vectors

Sampler is geometrically ergodic, so good performance is expected in practice: ergodic averages converge rapidly, CLT holds, Monte Carlo error can be assessed by the method of batch means.

State space is compact, target and proposal pdf's bounded away from 0 and ∞, so Foster–Lyapunov drift conditions (Meyn and Tweedie 1993) readily checked.

Example with simulated data





Left: true link function (solid line) and response data at the true transformed design points. Right: estimated fit (solid line) and true fit (dashed line).

$n = 100$, $f(u) = u^2 e^u$, $\sigma = 0.5$

$X_i \sim$ iid uniform on $[-1, 1]^4$

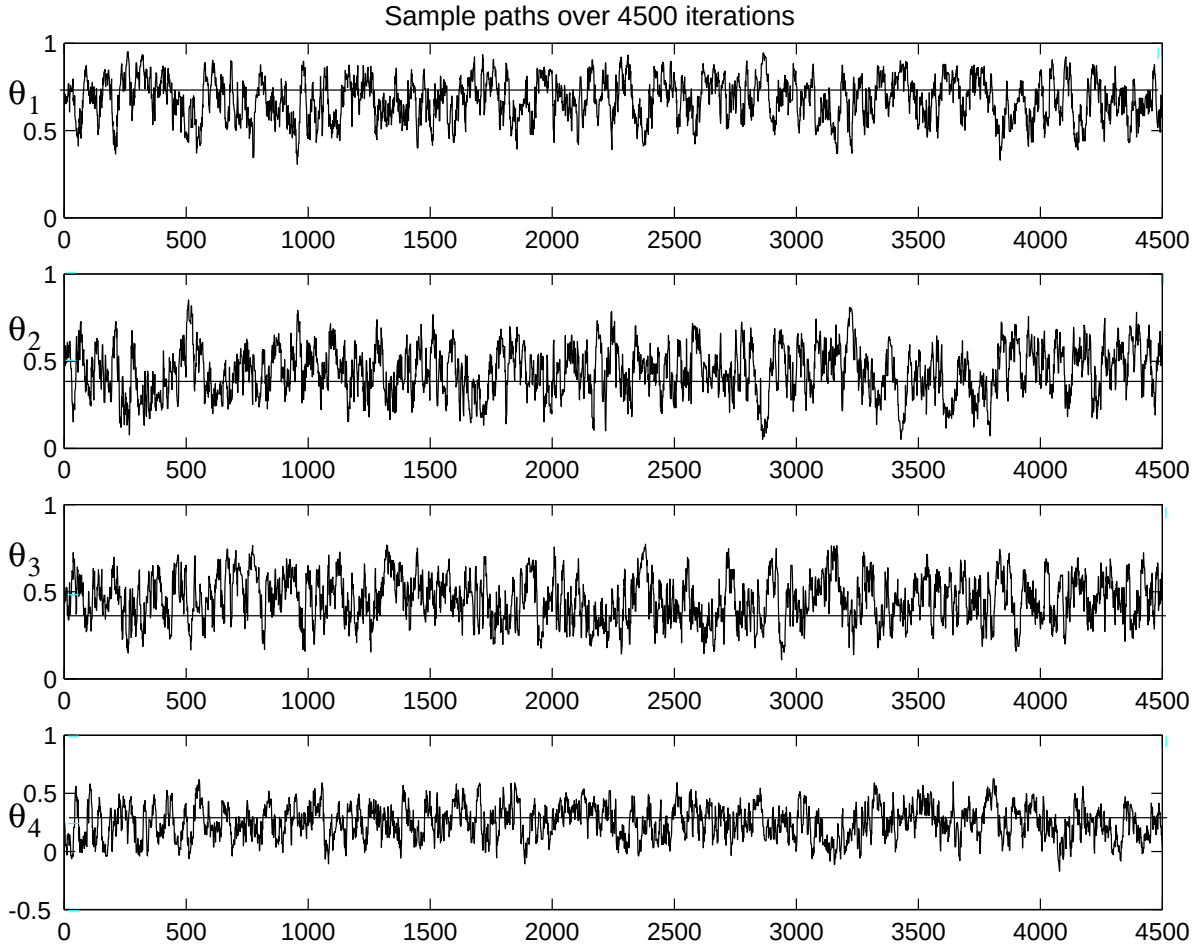
$\boldsymbol{\theta} = (2, 1, 1, 1)'/\sqrt{7}$, $\boldsymbol{\theta}_{\text{prior}} =$ HJS estimate

$\lambda_{\text{prior}} = 150$, $A = 0.001$, $B = 100$

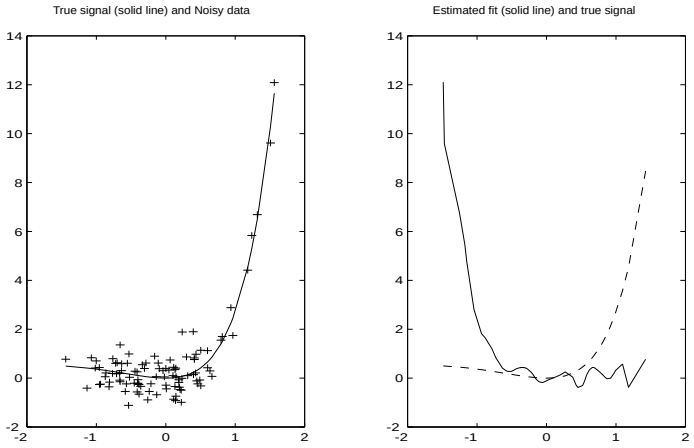
B-splines: 20 knots, degree $q = 2$, $\delta = .01$.
Regularization parameter: $\rho = 4$.
Proposal: $\lambda_{\text{prop}} = 1000$; acceptance rate 52%.

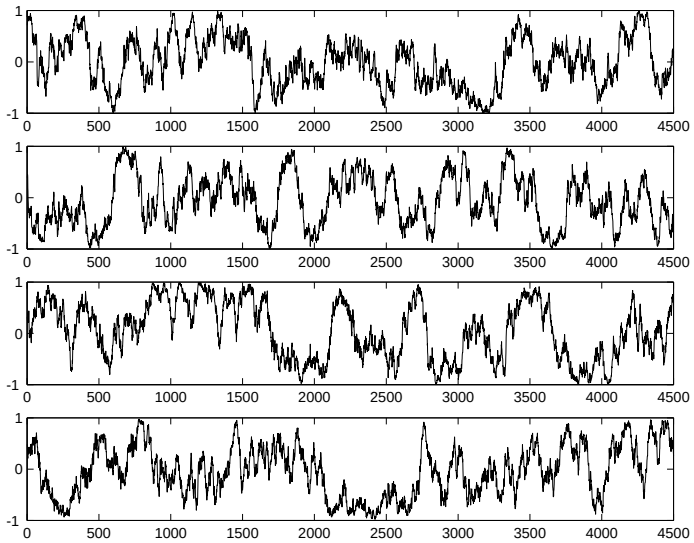
Burn-in of 500 iterations; 4000 iterations to obtain $\widehat{\boldsymbol{\theta}}$; Monte Carlo error negligible.

MCMC trace:

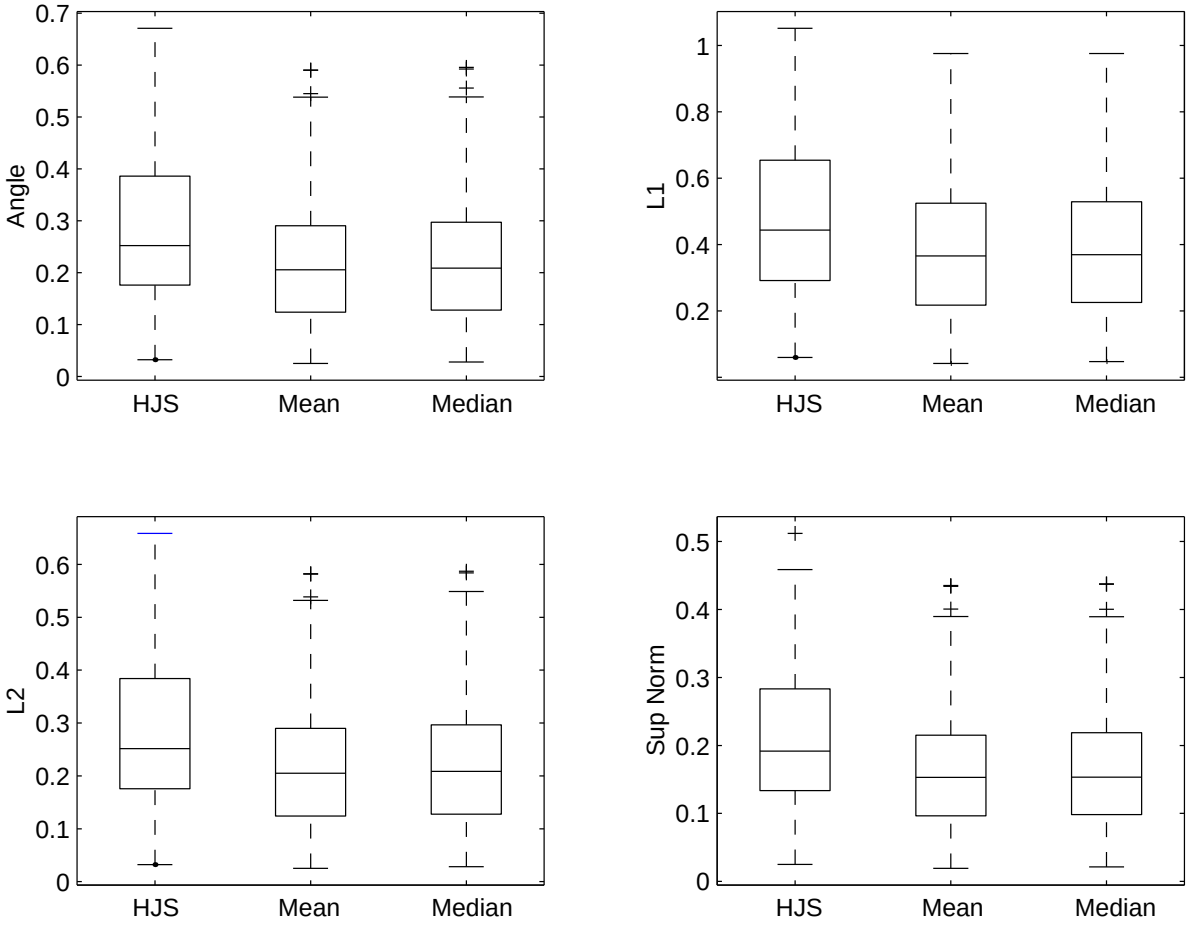


Absence of regularization has a disastrous effect on the estimates:





100 simulation runs



Posterior mean provides the best estimator, and clearly outperforms HJS.

Application to Climate Prediction

$\sim$10% of US GNP is modulated by weather, so there is great interest in developing a traded financial market (weather derivatives) based on atmospheric predictors.

Aquila Competition: predict heating-degree day pdf's for 13 US cities with 2 week lead outlooks; $50,000 prize every 3 months.

Clarke, van Gorda, McKeague:

(1) parametric time series model to capture the structure of temperature process

(2) single-index model to capture dependence of the time series parameters on various atmospheric predictors

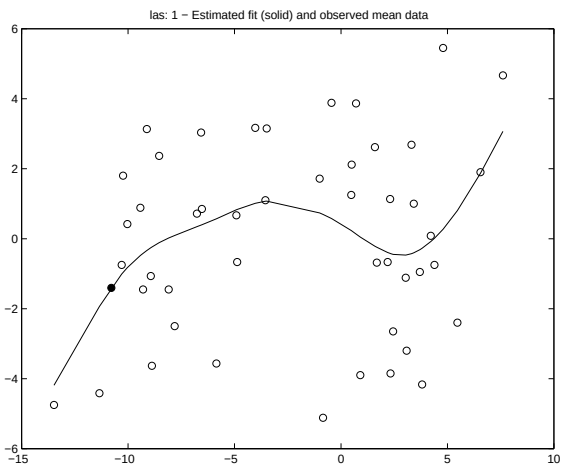
(3) simulation of (fitted) time series to obtain a predicted heating-degree day pdf

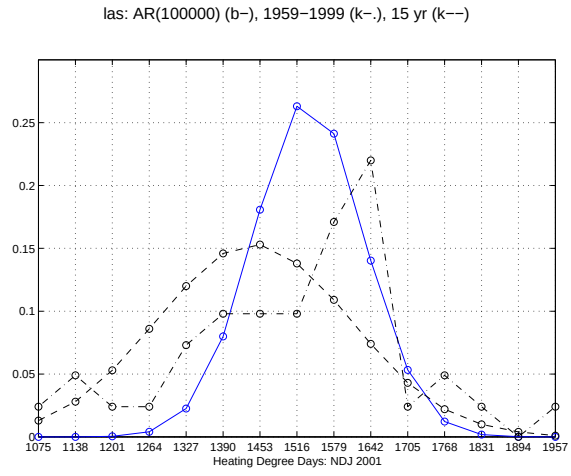
$n = 41$ years of data

AR(1) parameters: mean, lag-1 correlation, noise variance

Predictors: El Niño–Southern Oscillation index, North Atlantic Oscillation index

Las Vegas, November 2001:





Mean temperature (left) and predicted heating-degree day pdf (right) compared with the climatology histogram and a normal pdf with the climatology mean and variance (dashed lines).

BAYESIAN INVERSION OF OCEAN CIRCULATION DATA

Joint work with Kevin Speer and Seo-Eun Choi.

Background

Ocean circulation inverse problem: reconstruct the unknown climatological ocean flow field.

Since 1950, methods from geophysical fluid dynamics have been used to develop various models to attack this problem. Models arise from approximations to Navier–Stokes equations for a thin shell of temperature and salinity stratified fluid on a bumpy near-spheroidal body undergoing rapid rotation. Appropriate boundary conditions include no fluid flow into the bottom and sides, and assumptions concerning surface exchange of momentum, heat and moisture with the atmosphere.

Task of fitting such models to data (sparse observations from ships and satellites) leads to enormously complex ill-posed inverse problems.

Data: National Oceanic Data Center has collected data from hydrographic stations world-wide (including approximately 144,000 the North Atlantic alone). Used to construct climatological maps of ocean properties such as temperature, salinity and pressure (Levitus 1994).

Due to scarcity of data in any particular year and spatial irregularity of sampling stations, statistical smoothing methods are useful for estimating at space-time locations for which there are no measurements.

Lavine and Lozier (1999): used Markov random field models to construct temperature climatology maps for the North Atlantic.

Data assimilation: Combining data with general circulation models; constrained least squares (Wunsch 1996).

Bayesian inversion has some advantages: more attractive from a statistical modeling point of view, uncertainties in the reconstruction assessed by via error variances rather than Lagrange multipliers.

Recent work on Bayesian approaches to ill-posed inverse problems: Mosegaard and Tarantola (1995), Fox and Nicholls (1997), Nicholls and Fox (1998), Mosegaard (1998), Kaipio et al. (2000), Andersen et al (2001a, 2001b), and Higdon et al. (2001).

Stommel Gulf Stream model

$$\nabla^2\psi + \delta_1\frac{\partial\psi}{\partial x} = \delta_2 W,$$

$\psi = \psi(x, y)$, $0 < x, y < 1$, transport stream function

$W = W(x, y)$ wind stress curl

δ_1, δ_2 represent friction with the ocean floor, ocean depth, and earth's rotation (Coriolis force).

Problem: reconstruct ψ and W from sparse noisy observations.

Prior for (ψ, W)

Lattice approximation, sites (k, l)

Markov random field prior:

$$p(\psi, W) \propto \exp(-H(\psi, W))$$

(Gibbs distribution), energy function:

$$\begin{aligned} H(\psi, W) &= \sum_{(k,l)} \left[(\nabla^2 \psi)_{kl} + \delta_1 (\psi_x)_{kl} - \delta_2 W_{kl} \right]^2 \\ &\quad + \sum_{(k,l) \sim (k',l')} \left[\beta_\psi (\psi_{kl} - \psi_{k'l'})^2 + \beta_W (W_{kl} - W_{k'l'})^2 \right]. \end{aligned}$$

$(\psi_x)_{kl}$, $(\nabla^2 \psi)_{kl}$: lattice approximations to to the partial derivative and the Laplacian.

$\sim$ means 'nearest neighbors'

β_ψ, β_W: tuning parameters.

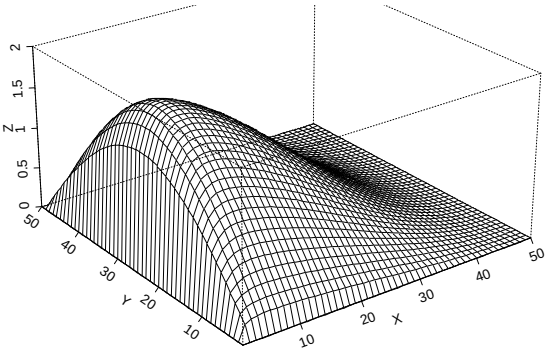
Example with simulated data

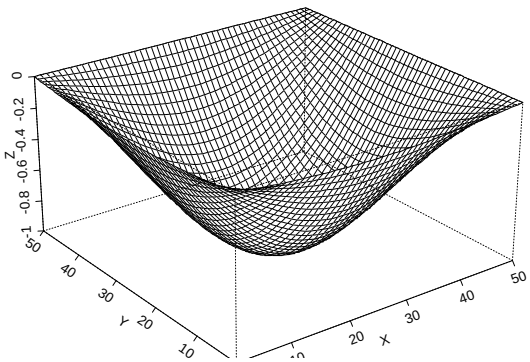
Imposed windstress curl:

$$W(x, y) = -\sin x \sin y$$

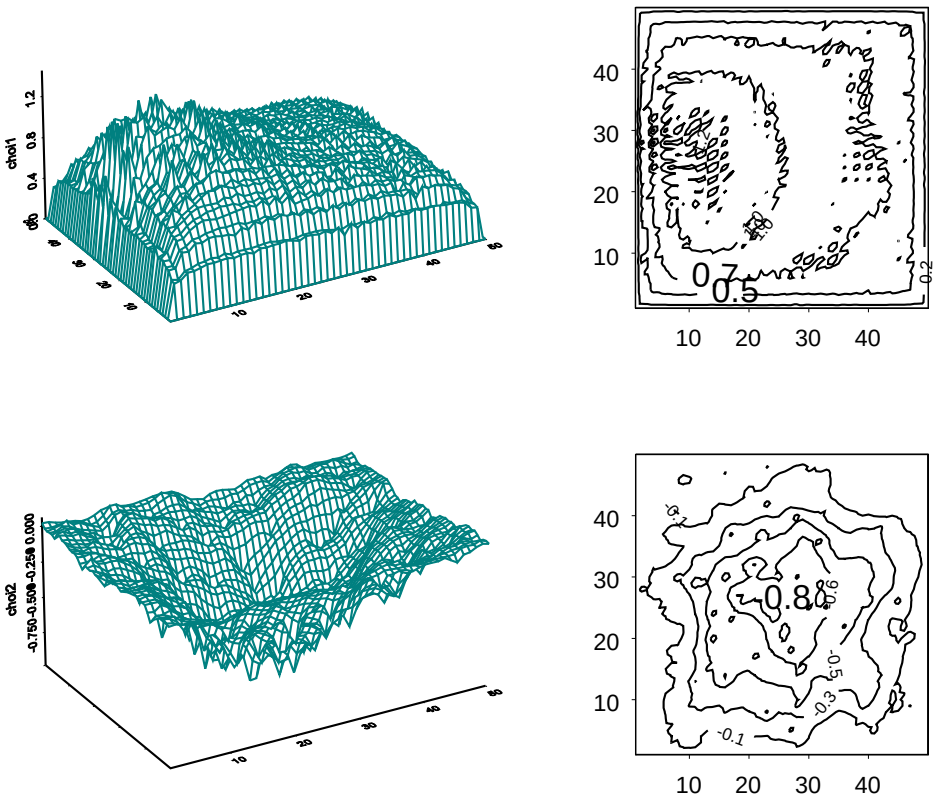
standard wind pattern for the North Atlantic.

Exact solution for ψ (left); imposed windstress curl W (right):





Gibbs sampler reconstructions



50×50 lattice

Data at 5% of sites

Gaussian measurement error

Reconstruction of ψ indicates a western boundary current (Gulf stream).