Riemann Manifold Methods in Bayesian Statistics

Ricardo Ehlers
ehlers@icmc.usp.br

Applied Maths and Stats
University of São Paulo, Brazil

Working Group in Statistical Learning
University College Dublin
September 2015

Bayesian inference is based on Bayes theorem:

$$\pi(\boldsymbol{\theta}) \propto L(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

We want to make inferences about a function $g(\boldsymbol{\theta})$ computing its posterior mean

$$\mathbb{E}_{\pi}[g(\boldsymbol{\theta})] = \int g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

typically analytically intractable.

Markov chain Monte Carlo

- Define transition densities $P(\boldsymbol{\theta}^t, \boldsymbol{\theta}^{t+1})$ of a Markov chain.
- Generate $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^m \sim \pi(\boldsymbol{\theta})$ (Target distribution).
- Under certain conditions,

$$\boldsymbol{\theta}^t \xrightarrow{t\to\infty} \pi(\boldsymbol{\theta}) \qquad \text{and} \qquad \frac{1}{m}\sum_{t=1}^{m} g(\theta_i^t) \xrightarrow{m\to\infty} \mathbb{E}_\pi(g(\theta_i)) \quad a.s.$$

- The chain is dependent by definition but the arithmetic mean of the chain values is a consistent estimator of the theoretical mean.

The Metropolis-Hastings Algorithm

At each iteration,

- sample a candidate value $\boldsymbol{\theta}' \sim q(\cdot, \boldsymbol{\theta})$.
- accept w.p.

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \min\left\{1, \frac{\pi(\boldsymbol{\theta}')\ q(\boldsymbol{\theta}, \boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})\ q(\boldsymbol{\theta}', \boldsymbol{\theta})}\right\}, \tag{1}$$

  where $q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is an arbitrary distribution which drives the general performance of the algorithm.

- $\pi(\boldsymbol{\theta})\ q(\boldsymbol{\theta}, \boldsymbol{\theta}')\ \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}') = \pi(\boldsymbol{\theta}')\ q(\boldsymbol{\theta}', \boldsymbol{\theta})\ \alpha(\boldsymbol{\theta}', \boldsymbol{\theta})$

Langevin Diffusions

Let $\boldsymbol{\xi} \in \mathbb{R}^D$ be a random vector with density $f(\boldsymbol{\xi})$. A Langevin diffusion with stationary distribution $f(\boldsymbol{\xi})$ is defined by the SDE,

$$d\boldsymbol{\xi}(t) = \frac{1}{2}\nabla_\xi \log f(\boldsymbol{\xi}(t))\, dt + d\mathbf{b}(t)$$

where $\mathbf{b}$ denotes a D-dimensional Brownian motion. This is the only non-explosive diffusion which is reversible with respect to f.

The Metropolis adjusted Langevin algorithm MALA is based on a first order Euler discretization giving the following proposal mechanism,

$$\boldsymbol{\xi}^{(t+1)} = \boldsymbol{\xi}^{(t)} + \frac{\epsilon^2}{2}\nabla_\xi \log f(\boldsymbol{\xi}^{(t)}) + \epsilon\mathbf{z},\ \mathbf{z} \sim N(0, \mathbf{I}_D) \qquad (2)$$

where ϵ is the integration step size.

Acceptance probability to ensure convergence to the invariant distribution.

1. A new value $\boldsymbol{\xi}'$ is sampled from a multivariate normal distribution with mean

$$\mu(\boldsymbol{\xi}^{(t)}, \epsilon) = \boldsymbol{\xi}^{(t)} + \frac{\epsilon^2}{2} \nabla_\xi \log f(\boldsymbol{\xi}^{(t)})$$

and variance-covariance matrix $\epsilon^2 \mathbf{I}_D$.

2. This value is accepted with probability given by,

$$\min \left\{ 1, \frac{f(\boldsymbol{\xi}')}{f(\boldsymbol{\xi}^{(t)})} \frac{\exp\{-||\boldsymbol{\xi}' - \mu(\boldsymbol{\xi}^{(t)}, \epsilon)||^2 / 2\epsilon^2\}}{\exp\{-||\boldsymbol{\xi}^{(t)} - \mu(\boldsymbol{\xi}', \epsilon)||^2 / 2\epsilon^2\}} \right\}$$

since the proposal distribution is $N(\mu(\boldsymbol{\xi}^{(t)}, \epsilon), \epsilon^2 \mathbf{I}_D)$.

Riemann Manifold MALA

Moves are according to a Riemann metric and is referred to as Riemann manifold MALA or MMALA. The proposal mechanism is now given by,

$$\xi_i' = \mu(\boldsymbol{\xi}^{(t)}, \epsilon)_i + \left\{ \epsilon \sqrt{\mathbf{G}^{-1}}(\boldsymbol{\xi}^{(t)}) \mathbf{z} \right\}_i, \tag{3}$$

$$\begin{aligned} \mu(\boldsymbol{\xi}^{(t)}, \epsilon)_i &= \xi_i^{(t)} + \frac{\epsilon^2}{2} \left\{ \mathbf{G}^{-1}(\boldsymbol{\xi}^{(t)}) \nabla_\xi \log f(\boldsymbol{\xi}^{(t)}) \right\}_i \\ &- \epsilon^2 \sum_{j=1}^{D} \left\{ \mathbf{G}^{-1}(\boldsymbol{\xi}^{(t)}) \frac{d\mathbf{G}(\boldsymbol{\xi}^{(t)})}{d\xi_j} \mathbf{G}^{-1}(\boldsymbol{\xi}^{(t)}) \right\}_{ij} \\ &+ \frac{\epsilon^2}{2} \sum_{j=1}^{D} \left\{ \mathbf{G}^{-1}(\boldsymbol{\xi}^{(t)}) \right\}_{ij} tr \left\{ \mathbf{G}^{-1}(\boldsymbol{\xi}^{(t)}) \frac{d\mathbf{G}(\boldsymbol{\xi}^{(t)})}{d\xi_j} \right\} \end{aligned} \tag{4}$$

In Bayesian applications, given a sample $y_1, \ldots, y_n$ then

$$\log f(\boldsymbol{\xi}) = \log f(\boldsymbol{\xi}|\mathbf{y}) = \log f(\boldsymbol{\xi}) + \log f(\mathbf{y}|\boldsymbol{\xi}) - \log f(\mathbf{y}),$$

$$\mathbf{G}(\boldsymbol{\xi}) = -E\left(\frac{d^2 \log f(\mathbf{y}|\boldsymbol{\xi})}{d\boldsymbol{\xi}^\top \boldsymbol{\xi}}\right) - \frac{d^2 \log f(\boldsymbol{\xi})}{d\boldsymbol{\xi}^\top \boldsymbol{\xi}},$$

1. Sample $\boldsymbol{\xi}'$ from a multivariate normal distribution with mean $\mu(\boldsymbol{\xi}^{(t)}, \epsilon)$ and variance matrix $\epsilon^2 \mathbf{G}(\boldsymbol{\xi}^{(t)})$.
2. Accept with probability $\min\{1, A\}$,

$$A = \frac{f(\boldsymbol{\xi}')}{f(\boldsymbol{\xi}^{(t)})} \frac{|\mathbf{G}(\boldsymbol{\xi}^{(t)})|^{-1/2} \exp\left\{-\frac{1}{2\epsilon^2}(\boldsymbol{\xi}' - \mu(\boldsymbol{\xi}^{(t)}, \epsilon))^T \mathbf{G}^{-1}(\boldsymbol{\xi}^{(t)})(\boldsymbol{\xi}' - \mu(\boldsymbol{\xi}^{(t)}, \epsilon))\right\}}{|\mathbf{G}(\boldsymbol{\xi}')|^{-1/2} \exp\left\{-\frac{1}{2\epsilon^2}(\boldsymbol{\xi}^{(t)} - \mu(\boldsymbol{\xi}', \epsilon))^T \mathbf{G}^{-1}(\boldsymbol{\xi}')(\boldsymbol{\xi}^{(t)} - \mu(\boldsymbol{\xi}', \epsilon))\right\}}$$

Stochastic Volatility Models (with M. Zevallos and L. Gasco)

A Stochastic Volatility model for a time series of financial returns,

$$\begin{aligned} y_t &= \beta \exp(h_t/2)\, \varepsilon_t, \\ h_t &= \phi h_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma^2) \end{aligned}$$

$E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = 1$, η_t and ε_t are independent $\forall t$, $\beta > 0$, $|\phi| < 1$ and ε_t follows a Gaussian, a generalized error distribution (GED) with shape ν or a t distribution with ν degrees of freedom. Defining $\boldsymbol{\theta} = (\beta, \sigma, \phi, \nu)$,

$$f(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) = \prod_{t=1}^{n} f(y_t|h_t, \beta)\; f(h_1|\phi, \sigma)\; \prod_{t=2}^{n} f(h_t|h_{t-1}, \phi, \sigma)$$

We implemented a hybrid method in which a MMALA scheme is applied for the parameters and a MALA scheme is applied for the volatilities.

- *Sample the latent variables* $\mathbf{h}$. Assuming the parameters as constants, apply (2) with $f = f(\mathbf{y}, \mathbf{h})$ and gradient ∇ calculated with respect to $\mathbf{h}$.
- *Sample parameters* $\boldsymbol{\theta}$. Given $(\mathbf{y}, \mathbf{h})$, apply (3) and (4) with $f = f(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta})f(\boldsymbol{\theta})$ and gradient ∇ calculated with respect to $\boldsymbol{\theta}$.

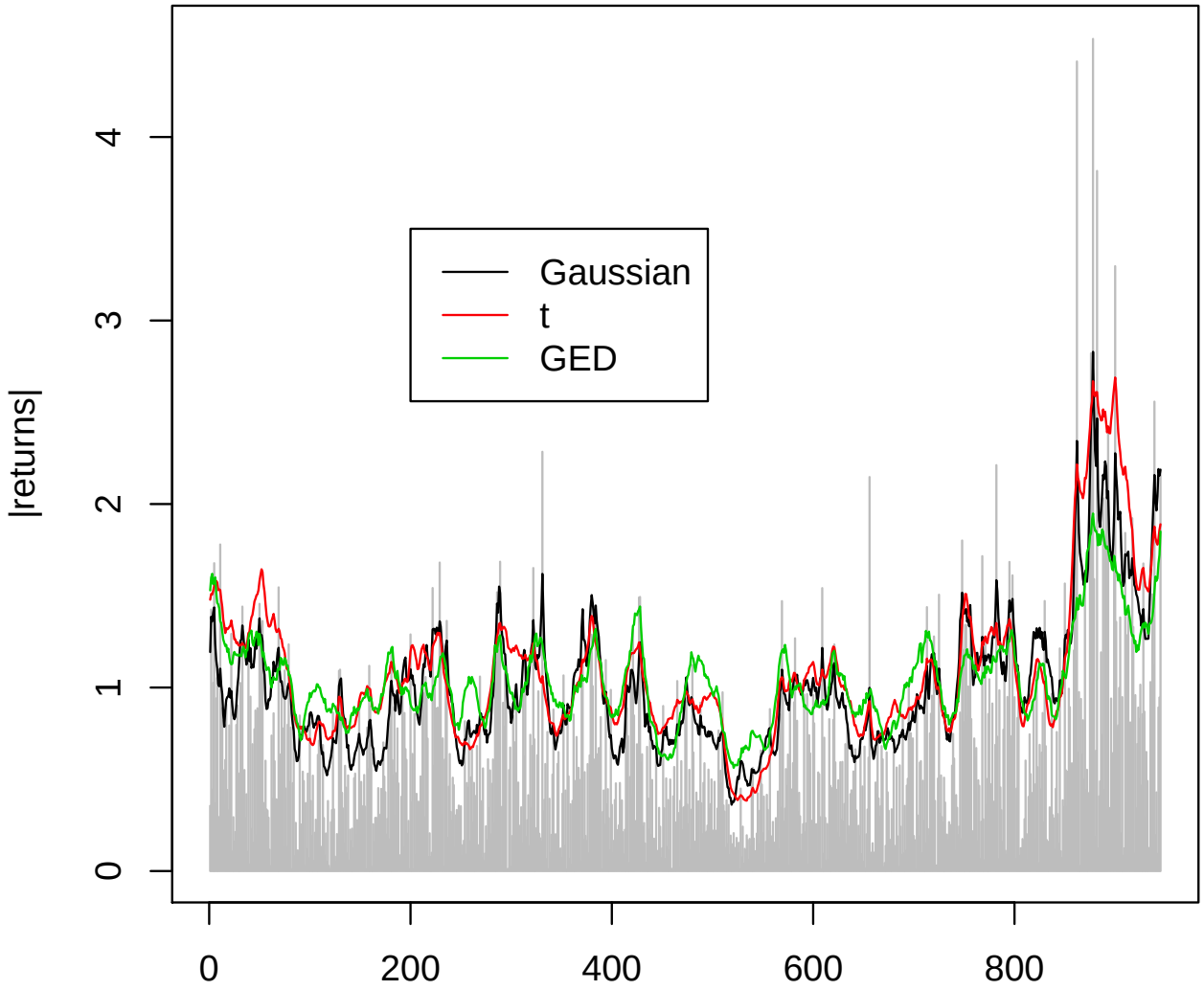
A simplified proposal mechanism is obtained when a constant curvature is assumed. In this case, the last two terms in (4) vanish and the proposal mean becomes,

$$\boldsymbol{\mu}(\boldsymbol{\theta}^{(t)}, \epsilon) = \boldsymbol{\theta}^{(t)} + \frac{\epsilon^2}{2}\mathbf{G}^{-1}(\boldsymbol{\theta}^{(t)})\ \nabla_\theta \log f(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}^{(t)}) f(\boldsymbol{\theta}^{(t)}).$$

Monte Carlo experiments. Bias and square root of the mean squared error of posterior means. Parameters: $\beta = 0.65$, $\phi = 0.98$, $\sigma = 0.15$ and $\nu = 1.6$ (for GED) and $\nu = 7$ (for Student's t).

Errors	Method	β		ϕ		σ		ν	
		bias	smse	bias	smse	bias	smse	bias	smse
Gaussian	MALA	-0.001	0.038	-0.022	0.028	0.051	0.056		
	MMALA	0.024	0.038	-0.011	0.015	0.000	0.014		
GED	MALA	-0.002	0.032	-0.042	0.051	0.090	0.099	-0.011	0.128
	MMALA	0.002	0.029	-0.027	0.032	0.050	0.054	0.048	0.115
Student's t	MALA	-0.003	0.031	-0.063	0.072	0.122	0.131	0.912	2.311
	MMALA	-0.010	0.030	-0.101	0.107	0.180	0.185	0.287	1.428

Absolute returns for the Pound/Dollar series and estimated volatilities using MMALA under the three different errors.



Stochastic simulation via Hamiltonian dynamics

For a closed and conservative system of particles the total energy is constant and given by the Hamiltonian function. For purposes of generate random values from an arbitrary distribution suppose that function is given by,

$$H(\boldsymbol{\theta}, \mathbf{p}) = \underbrace{-\log \pi(\boldsymbol{\theta}|D)}_{\substack{\text{potential} \\ \text{energy}}} + \underbrace{\mathbf{p}' M^{-1} \mathbf{p}/2}_{\substack{\text{kinetic} \\ \text{energy}}} \tag{5}$$

$\mathbf{p}$ is the momentum, M is a mass matrix (positive definite).

Calculate the final position of the particle via Hamiltonian dynamics through the system of differential given by,

$$\begin{aligned} \frac{d\boldsymbol{\theta}}{dt} &= +\frac{\partial H}{\partial \mathbf{p}} = \mathbf{p}M^{-1} \\ \frac{d\mathbf{p}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}|D) \end{aligned} \tag{6}$$

Introducing the auxiliary variables $\mathbf{p}$ and using the gradients will lead to a more efficient exploration of the parameter space.

These differential equations cannot be solved analytically. The Störmer-Verlet (or Leapfrog) numerical integrator discretizes the Hamiltonian dynamics as the following steps,

$$\begin{aligned}
\mathbf{p}^{(\tau+\epsilon/2)} &= \mathbf{p}^{(\tau)} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^{(\tau)}) \\
\boldsymbol{\theta}^{(\tau+\epsilon)} &= \boldsymbol{\theta}^{(\tau)} + \epsilon\mathbf{M}^{-1}\mathbf{p}^{(\tau+\epsilon/2)} \\
\mathbf{p}^{(\tau+\epsilon)} &= \mathbf{p}^{(\tau+\epsilon/2)} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}^{(\tau+\epsilon)})
\end{aligned}$$

for some user specified small step-size $\epsilon > 0$. After a given number of time steps this results in a proposal $(\boldsymbol{\theta}^*, \mathbf{p}^*)$.

- Metropolis acceptance probability corrects discretization error and ensures convergence to the invariant distribution.
- The joint distribution of $(\boldsymbol{\theta}, \mathbf{p})$ is our target distribution,
- transition to a proposed value $(\boldsymbol{\theta}^*, \mathbf{p}^*)$ is accepted w.p.

$$\begin{aligned} \alpha[(\boldsymbol{\theta}, \mathbf{p}), (\boldsymbol{\theta}^*, \mathbf{p}^*)] &= \min\left[\frac{f(\boldsymbol{\theta}^*, \mathbf{p}^*)}{f(\boldsymbol{\theta}, \mathbf{p})}, 1\right] \\ &= \min\left[\exp[H(\boldsymbol{\theta}, \mathbf{p}) - H(\boldsymbol{\theta}^*, \mathbf{p}^*)], 1\right]. \end{aligned}$$

- $\mathbf{M}$ is typically diagonal with constant elements, $\mathbf{M} = m\mathbf{I}_d$. The HMC algorithm in its simplest form takes $m = 1$.

The algorithm is as follows.

1. Give an inital position, $\boldsymbol{\theta}^{(0)}$.
2. Iniciate the iterations $i = 1, \ldots, N$ (size of chain).
   1. Draw $\mathbf{p}^* \sim N_d(\mathbf{0}, I_d)$ and $u \sim U(0,1)$,
   2. Do $(\boldsymbol{\theta}^I, \mathbf{p}^I) = (\boldsymbol{\theta}^{(i-1)}, \mathbf{p}^*)$, $H_0 = H(\boldsymbol{\theta}^I, \mathbf{p}^I)$
   3. Repeat the Störmer-Verlet numerical solution in an adequate number of times and choose some step-size ϵ for the discretization of the system.
   4. At the end of the trajectory, do $H_1 = H(\boldsymbol{\theta}^L, \mathbf{p}^L)$
   5. Do, $\alpha[(\boldsymbol{\theta}^L, \mathbf{p}^L), (\boldsymbol{\theta}^I, \mathbf{p}^I)] = \min[\exp(H_0 - H_1), 1]$
   6. Metropolis-hastings acceptance rule,

$$\boldsymbol{\theta}^{(i)} = \begin{cases} \boldsymbol{\theta}^L, \text{ with probability } \alpha[(\boldsymbol{\theta}^L, \mathbf{p}^L), (\boldsymbol{\theta}^I, \mathbf{p}^I)] > u \\ \boldsymbol{\theta}^I, \text{otherwise} \end{cases}$$

Note that the exponential of the negative Hamiltonian function is a joint density function and all parameters must lie on the real line.

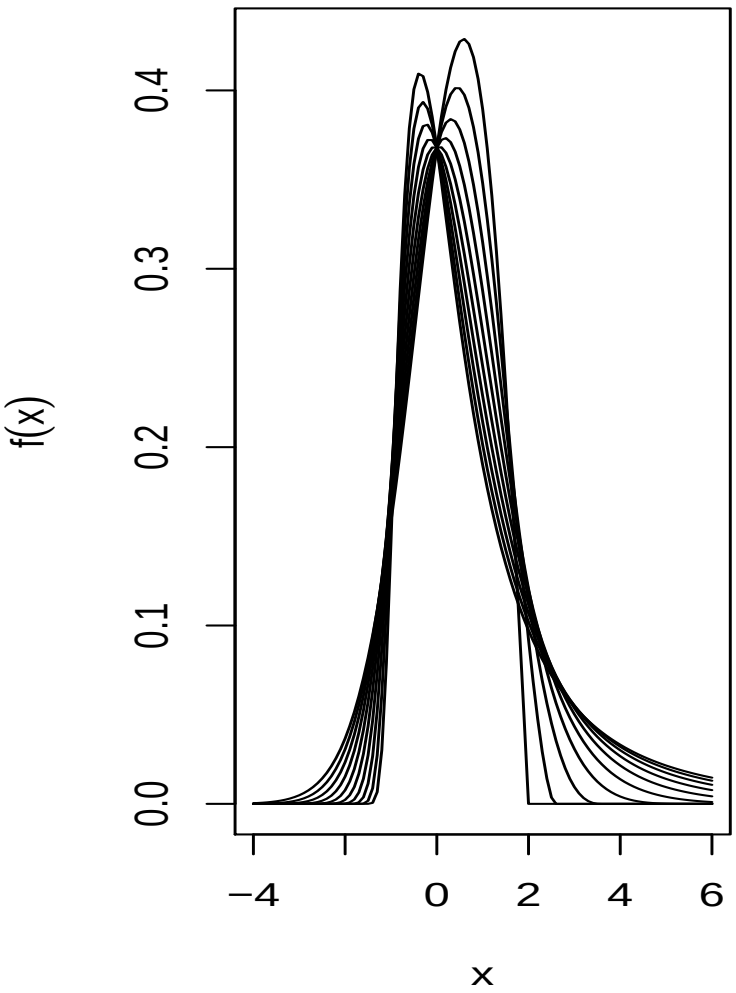
Bayesian inference for extreme value distributions (with M. Hartmann)

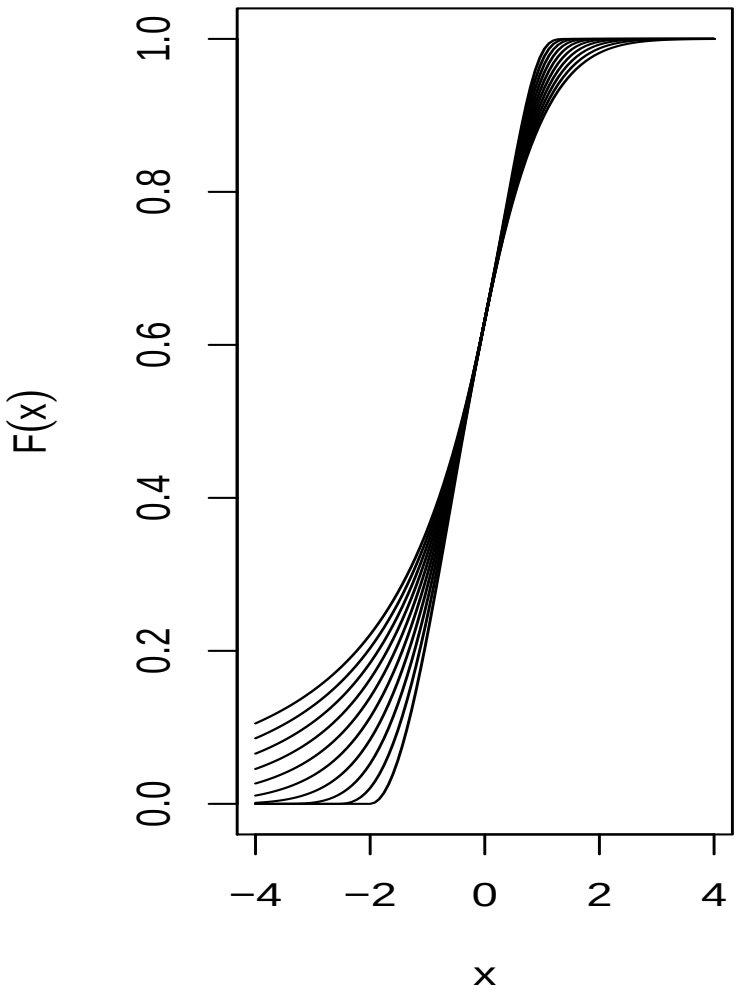
The Generalized Extreme Value (GEV) distribuion function,

$$F(x|\mu,\sigma,\xi) = \exp\left\{-\left(1+\xi\,\frac{x-\mu}{\sigma}\right)_+^{-1/\xi}\right\}, \qquad (7)$$

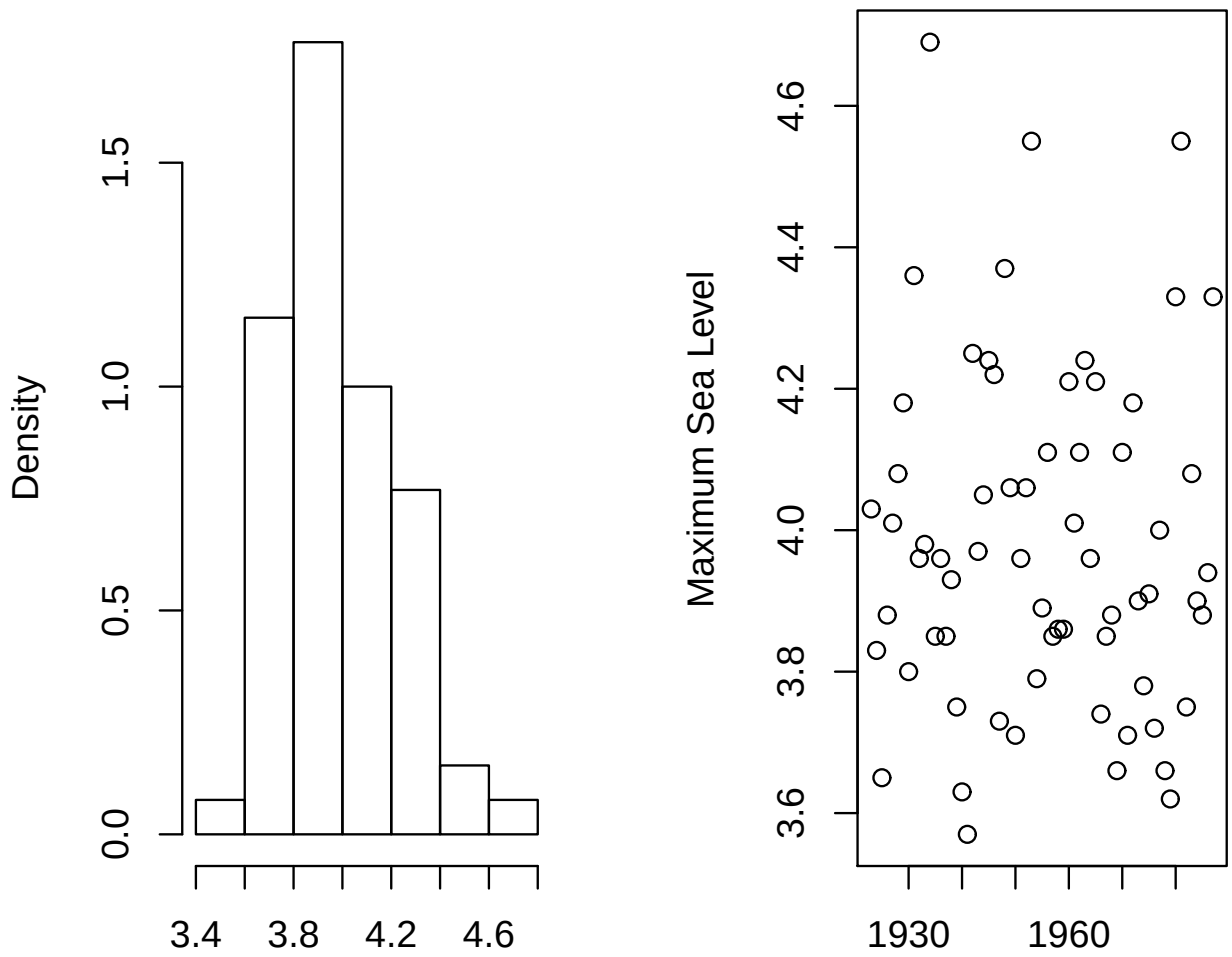
where $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$ are location, scale and shape parameters respectively. The $+$ sign denotes the positive part of the argument.

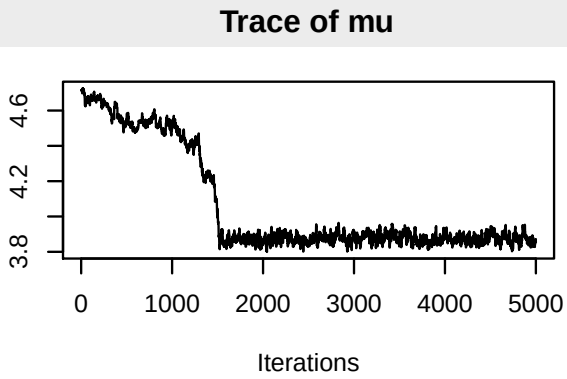
Densities and cumulative distribution functions of a GEV(0,1,ξ) with $\xi \in [-0.5, 0.5]$.

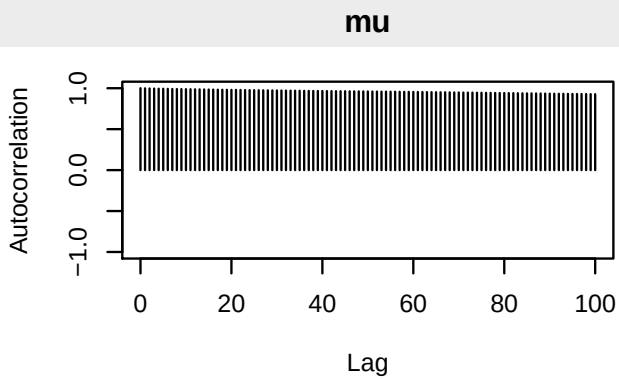


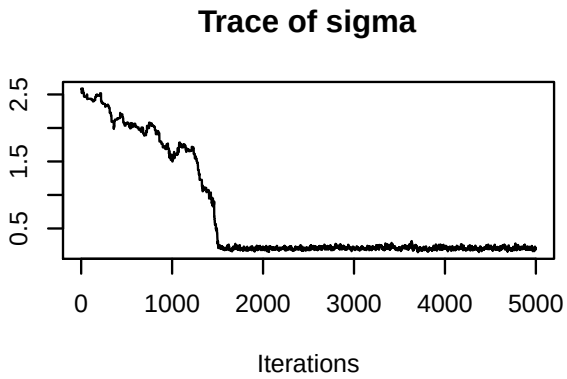


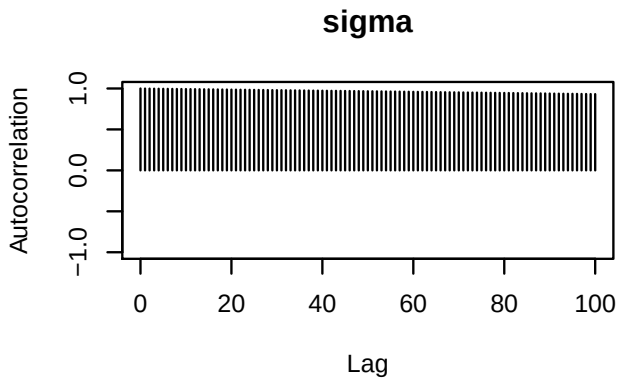
Histogram and plots of maximum sea levels (in metres) from 1923 to 1987 at Port Pirie, South Australia.

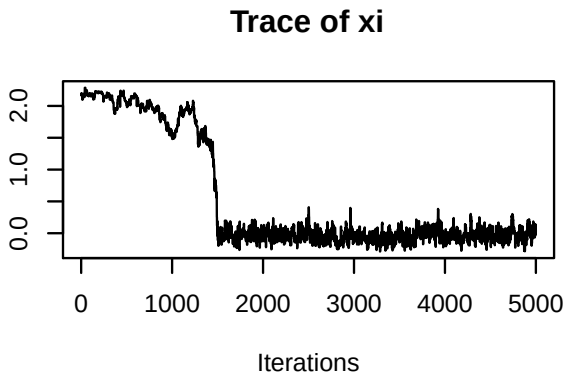


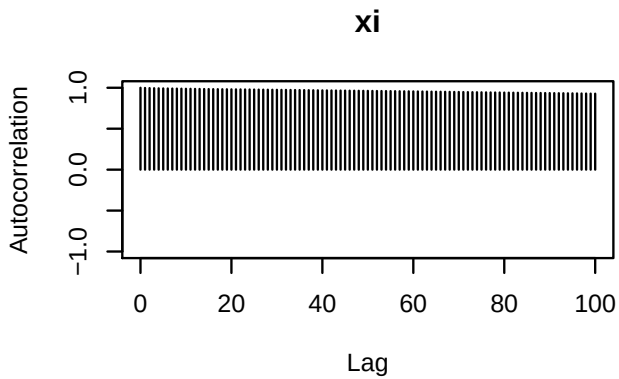

Trace of mu
3.8
4.2
4.6
0
1000
2000
3000
4000
5000
Iterations

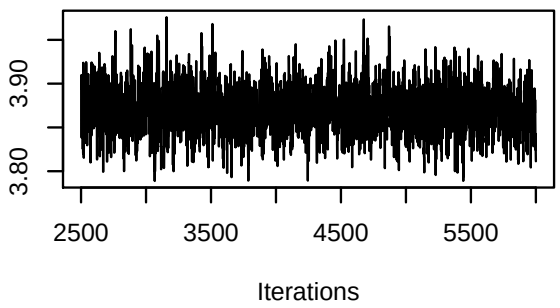

mu
Autocorrelation
−1.0
0.0
1.0
0
20
40
60
80
100
Lag

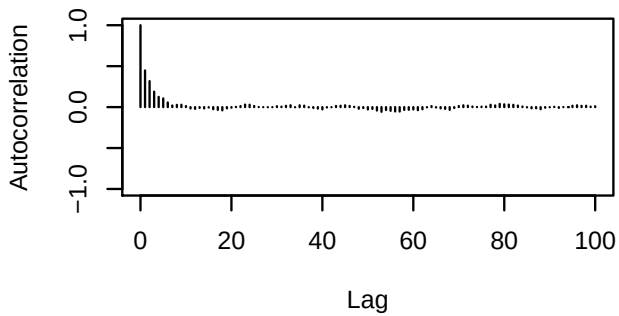

Trace of sigma
0.5
1.5
2.5
0
1000
2000
3000
4000
5000
Iterations

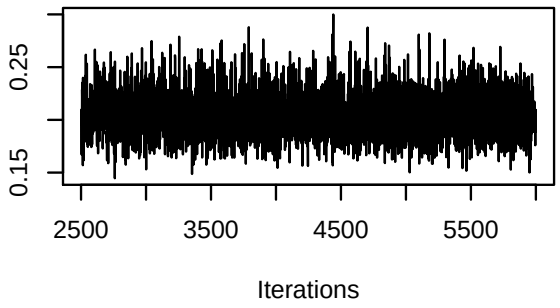

sigma
Autocorrelation
−1.0
0.0
1.0
0
20
40
60
80
100
Lag

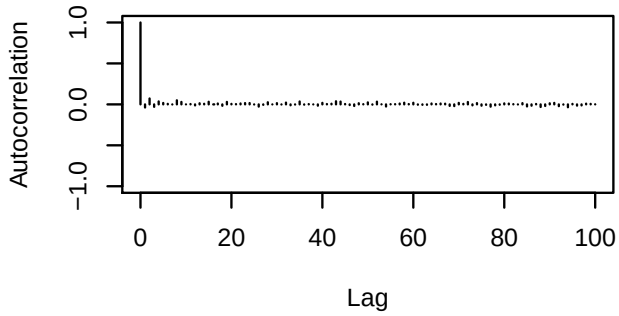

Trace of xi
0.0
1.0
2.0
0
1000
2000
3000
4000
5000
Iterations

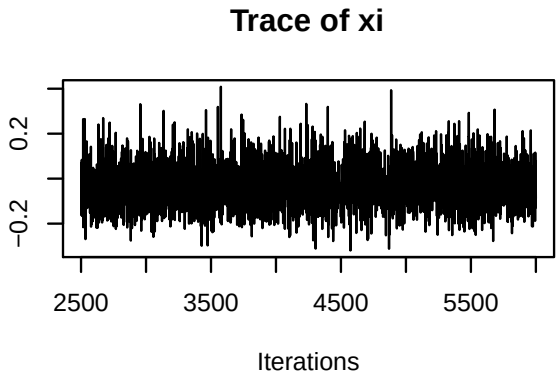

xi
Autocorrelation
−1.0
0.0
1.0
0
20
40
60
80
100
Lag

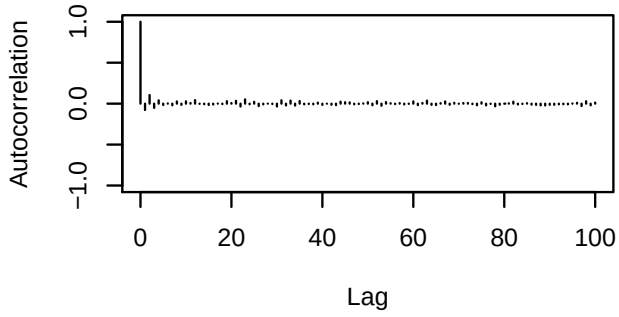

Trace of mu
3.80
3.90
2500
3500
4500
5500
Iterations


mu
Autocorrelation
−1.0
0.0
1.0
0
20
40
60
80
100
Lag


Trace of sigma
0.15
0.25
2500
3500
4500
5500
Iterations


sigma
Autocorrelation
−1.0
0.0
1.0
0
20
40
60
80
100
Lag


Trace of xi
−0.2
0.2
2500
3500
4500
5500
Iterations


xi
Autocorrelation
−1.0
0.0
1.0
0
20
40
60
80
100
Lag

Modelling Time Dependence

- AR-GEV(p) model,

$$Y_t = \mu + \sum_{j=1}^{p} \theta_j Y_{t-j} + e_t, \;\; e_t \sim GEV(0, \sigma, \xi).$$

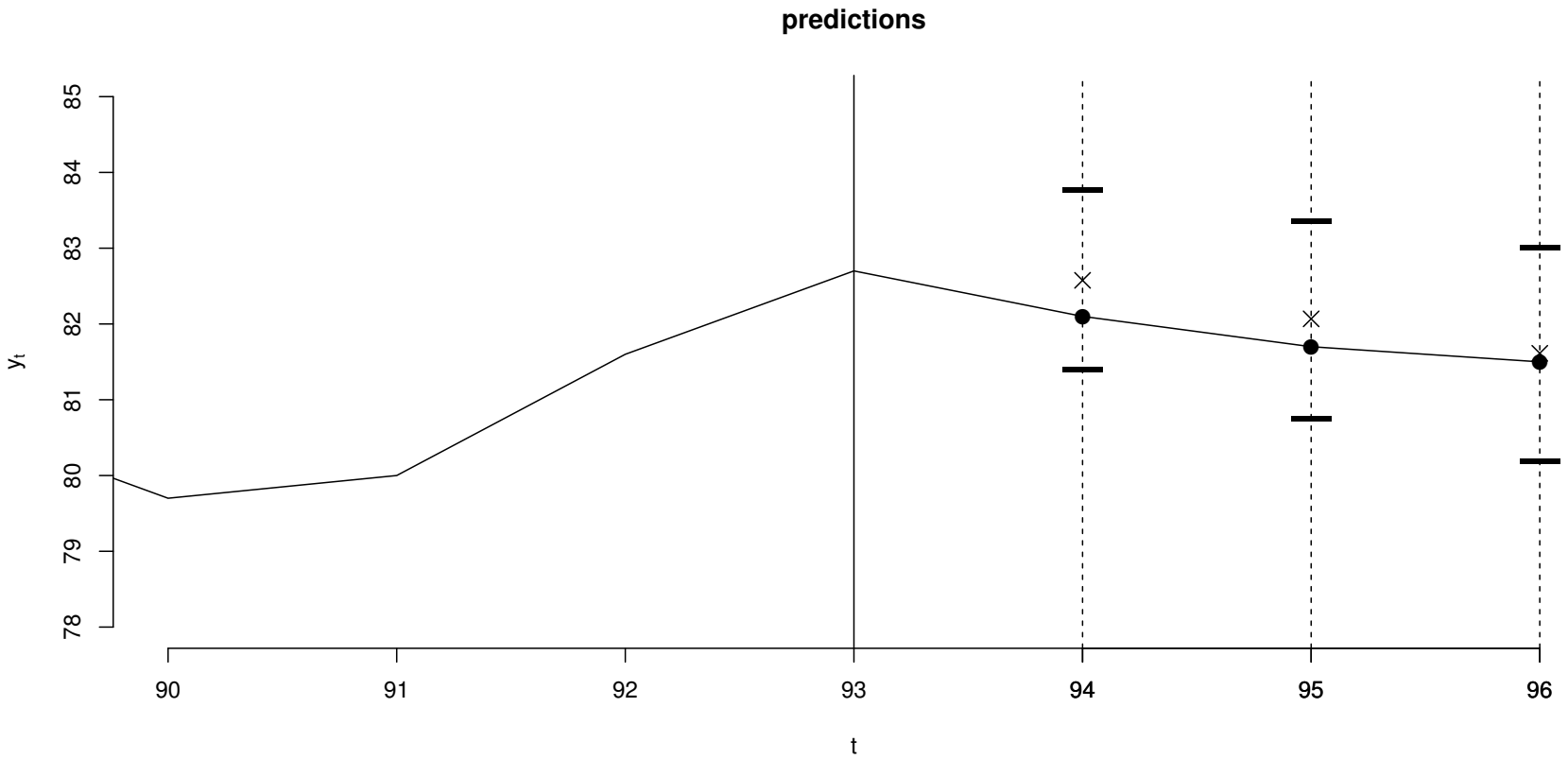
- The likelihood function,

$$l(\mu, \boldsymbol{\theta}, \sigma, \xi) = \prod_{t=p+1}^{n} f(y_t | D_{t-1}, \mu, \boldsymbol{\theta}, \sigma, \xi) I_{\Omega_t}(y_t), \qquad (8)$$

where $D_{t-1} = (y_{t-1}, \ldots, y_{t-p})$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.
Denoting $\mu_t = \mu + \sum_{j=1}^{p} \theta_j Y_{t-j}$ then
$\Omega_t = \{y_t : 1 + \xi(y_t - \mu_t)/\sigma > 0\}$

Maximum annual level of Lake Michigan, 1860 to 1955 (96 observations).
Time Series Data Library https://datamarket.com/data/set/22p3/

Predictions plus 95% credible intervals (last 3 observations removed from estimation) and observed values (circles).



Bayesian Analysis of Clustered Binary Data (with N. Friel and D. Bandyopadhyay)

- Consider a spatial situation where we observe a binary response y_{is} for subject i, at site s within subject i.
- We assume that $Y_{is} \sim \text{Bernoulli}(p_{is})$ with

$$\begin{aligned} P(Y_{is} = 1) = p_{is} &= 1 - F(-(\mathbf{x}_i'\boldsymbol{\beta} + \phi_{is})) \\ &= 1 - \exp\left\{ - \left[1 - \xi(\mathbf{x}_i'\boldsymbol{\beta} + \phi_{is})\right]_+^{-1/\xi} \right\} \end{aligned}$$

Assuming that $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and ξ are a priori independent the joint posterior distribution is given by,

$$p(\boldsymbol{\beta}, \boldsymbol{\phi}, \xi | \mathbf{y}, \mathbf{X}) \propto \prod_{i=1}^{n} p(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\beta}, \boldsymbol{\phi}_i)\ p(\boldsymbol{\phi}_i)\ p(\boldsymbol{\beta})\ p(\xi).$$

Instead of the usual CAR model for the spatial effects we assume,

$$\begin{aligned} \phi_i &\sim N(\mathbf{0}, \Sigma) \\ \Omega = \Sigma^{-1} &\sim \text{G-Wishart}_W(\kappa, S) \end{aligned}$$

degrees of freedom κ and scale matrix S, constrained to have null entries for each zero in the adjacency matrix W,

$$W_{ss'} = \begin{cases} 1, & s \sim s' \\ 0, & \text{otherwise} \end{cases}$$

Its density function is given by,

$$p(\Omega|W) = \frac{1}{Z_W(\kappa, S)} |\Omega|^{(\kappa-2)/2} \exp\left\{-\frac{1}{2} tr(S\Omega)\right\} \; I(\Omega \in M_W), \; \kappa > 2$$

- Start with a hybrid algorithm,
  - sample β and ξ using RMHMC assuming spatial effects as constants,
  - sample spatial effects using traditional MCMC.
- Computing the normalizing constant $Z_W(\kappa, S)$ is not straightforward,

$$Z_W(\kappa, S) = \int |\Omega|^{(\kappa-2)/2} \exp\left\{-\frac{1}{2} tr(S\Omega)\right\} \; I(\Omega \in M_W) d\Omega.$$

Thank you !