Bayesian Statistics

Anna Simoni[2]

[2]CREST, CNRS - Ensae

Outline

Outline

Bayesian estimation

The goal is to compute the full posterior (it gives measure of the uncertainty) . . . but can be hard.

- Finding the centre and/or spread of the posterior may be a substitute.
- Analytical computation of a posterior is rarely possible, but powerful algorithms allow to simulate from it.
- Markov Chain Monte Carlo (MCMC) produces a Markov chain $\theta_1, \theta_2, \ldots$ that has the posterior as its stationary distribution:

  After discarding $\theta_1, \ldots, \theta_k$,

  - the average of $\theta_{k+1}, \ldots, \theta_{k+l}$ is taken as estimate of the posterior mean
  - the fraction of $\theta_{k+1}, \ldots, \theta_{k+l}$ that falls in a set B is taken as estimate of the posterior mass of B.

Bayesian point-estimator.

- Bayesian point-estimator: a point in the model around which the posterior distribution is concentrated in some way.
- As such, any reasonable Bayesian point-estimator should represent the *location* of the posterior distribution . . . but there is no unique definition for the *location* of a distribution.
- Accordingly, there are many different ways to define Bayesian point-estimators.

Posterior mean estimator.

Definition (Posterior Mean Estimator)

Let $\mathscr{P} = \{P_\theta; \theta \in \Theta\}$ be a model parameterized by a closed, convex subset $\Theta \subset \mathbb{R}^d$. Let π be a prior defined on Θ with a regular posterior $\pi(\cdot|X^{(n)})$. If θ is integrable with respect to the posterior, the posterior mean is defined as

$$\widehat{\theta}_n := \int_\Theta \theta\pi(\theta|X^{(n)})d\theta \in \Theta,$$

almost surely.

Posterior median estimator.

Since there are multiple ways of defining the location of a distribution, a point-estimator alternative to $\widehat{\theta}_n$ is given by (which requires that the model is one-dimensional):

Definition (Posterior Median Estimator)

Let $\Theta \subset \mathbb{R}$ with non-empty interior and let $\mathscr{P} = \{P_\theta; \theta \in \Theta\}$ be a parametric model with prior π on Θ and posterior $\pi(\cdot|X^{(n)})$. The posterior median is defined by,

$$\widetilde{\theta}_n := \inf\{s \in \mathbb{R}; \pi(\theta \leq s|X^{(n)}) \geq 1/2\},$$

almost surely.

This definition simplifies in case the posterior has a continuous, (strictly) monotone distribution function: $\widetilde{\theta}_n$ is the unique point in Θ such that

$$\pi(\theta \leq \widetilde{\theta}_n|X^{(n)}) = 1/2.$$

Point Estimators based on expected loss. I

- Suppose that we consider estimation in a metric model $(\mathscr{P}, d)$ and we quantify errors in estimation as follows:

  if the true data distribution is P_* and we estimate that it is P_θ, then we incur a loss L that is a monotone increasing function $\phi : [0, \infty) \to [0, \infty)$ of the distance $d(P_*, P_\theta)$, that is: $L(P_*, P_\theta) = \phi(d(P_*, P_\theta))$.

- Let π be a prior distribution. We define the *integrated risk* (or *Bayes risk*):

$$\begin{aligned} r(\pi, a) &= \mathbf{E}_{\theta, X^{(n)}}\left[L(\theta, a(X^{(n)}))\right] \\ &= \int_\Theta \underbrace{\int_{\mathscr{X}} L(\theta, a(x^{(n)})) f(x^{(n)}|\theta) dx^{(n)}}_{:=R(\theta,a)=\text{Frequentist Risk}} \pi(\theta) d\theta \end{aligned}$$

  and the posterior risk: for $X^{(n)} = (x_1, \ldots, x_n)$

$$\begin{aligned} \rho(\pi, a, X^{(n)}) &= \mathbf{E}_{\theta|X^{(n)}}\left[L(\theta, a(X^{(n)}))|X^{(n)}\right] \\ &= \int_\Theta L(\theta, a(X^{(n)}))\pi(\theta|X^{(n)}) d\theta. \end{aligned}$$

Point Estimators based on expected loss. II

- If we assume that the posterior concentrates its mass around P_* then the expected loss relative to the posterior should provide guidance to construct the following point-estimators:

Definition (Bayes Estimator)

A Bayes estimator associated with a prior distribution π and a loss function L is an estimator $a(X^{(n)})$ that minimizes $r(\pi, a)$. For every $X^{(n)}$, the Bayes estimator is given by

$$a_* = \arg\min_a \rho(\pi, a, X^{(n)}).$$

The value $r(\pi) = r(\pi, a_*)$ is called Bayes risk.

- For a quadratic loss ($L(\theta, \delta) = \|\theta - \delta\|^2$) the Bayes estimator is the posterior mean:

$$\widehat{\theta}_n := \int_\Theta \theta \pi(\theta | X^{(n)}) d\theta \in \Theta.$$

- Bayesian point estimates are optimal from *ex post* standpoint.
- Estimators that minimizes $R(\theta, a)$ $\nexists$ in general.

Point Estimators based on expected loss. III

- Instead, one can

$$\min_a r(\pi, a), \qquad r(\pi, a) := \int_\Theta R(\theta, a)\pi(\theta)d\theta.$$

Remark that

$$\begin{aligned} r(\pi, a) &= \int_\Theta \int_{\mathcal{X}} L(\theta, a(x^{(n)}))f(x^{(n)}|\theta)dx^{(n)}\ \pi(\theta)d\theta \\ &= \int_{\mathcal{X}} \left[\int_\Theta L(\theta, a(x^{(n)}))\frac{f(x^{(n)}|\theta)p(\theta)}{m(x^{(n)})}d\theta\right] m(x^{(n)})dx^{(n)}. \end{aligned}$$

So, the quantity minimizing $r(\pi, a)$ also minimizes the quantity in the square brackets which is $\rho(\pi, a, X^{(n)})$.

- Therefore, any estimator that minimizes weighted risk must be a Bayesian point estimator.
- So, any Bayes estimator based on a proper prior is admissible (i.e. $\nexists$ another estimator with lower risk function at all points in Θ).

The Maximum-A-Posteriori (MAP) estimator. I

- An important estimator of θ based on the posterior $\pi(\theta|x)$ is the *maximum a posteriori* (MAP) estimator, defined as the posterior mode:

$$\delta^{\pi}(x) = \arg\max_{\theta} \pi(\theta|x).$$

  Provided that such a maximizer $\exists!$, the MAP estimator is defined almost surely.

- The MAP estimator maximises also $\ell(\theta|x)\pi(\theta)$ (where $\ell(\theta|x)$ denotes the likelihood function) and so it does not require the computation of the marginal probability.

The Maximum-A-Posteriori (MAP) estimator. II

- The MAP can be expressed as a *penalized MLE*. Consider an i.i.d. experiment with parametric model, the MAP estimator maximizes

$$\theta \mapsto \prod_{i=1}^{n} f(X_i|\theta)\pi(\theta),$$

where it is assumed that the model is dominated and that the prior has a density π with respect to the Lebesgue measure. Differences between ML and MAP estimators are entirely due to non-uniformity of the prior.

Prior non-uniformity has an interpretation in the frequentist setting as well, through penalized ML estimation:

$$\log \pi(\theta|X^{(n)}) = \log f(X^{(n)}|\theta) + \log \pi(\theta) + \log m(X^{(n)}),$$

where $m(X^{(n)})$ is the normalization constant.

Lemma:
Consider a parameterized model $\Theta \to \mathscr{P}$: $\theta \mapsto P_\theta$. If the parameter space Θ is compact and the posterior density $\theta \to \pi(\theta|X^{(n)})$ is upper-semi-continuous, then the MAP estimator exists almost surely.

Predictive Inference. I

- Based on the predictive distribution: for model A

$$m(y_t|y^{(t-1)}, A) = \int_{\Theta} f(y_t|y^{(t-1)}, \theta, A)\pi(\theta|y^{(t-1)}, A)d\theta$$

  where $y^{(t-1)} := (y_1, \ldots, y_{t-1})'$.

- This distribution can be accessed by simulating one value $y_t^{(m)}$ from each of the distributions represented by the density:

$$f(y_t|y^{(t-1)}, \theta^{(m)}, A), \qquad m = 1, \ldots, M,$$

  where $\{\theta^{(m)}\}_{m=1,\ldots,M}$ are drawn from the posterior $\pi(\theta|y^{(t-1)}, A)$.
- The predictive distribution $m(y_t|y^{(t-1)}, A)$ integrates uncertainty about θ and intrinsic uncertainty about the future value y_t, both conditional on the history $y^{(t-1)}$ and the assumptions of model A.
- Advantage of Bayesian predictive distributions: is the combination of the two sources of uncertainty in a coherent framework.

Outline

The normal model: known variance

- Suppose that the model is normal: $x|\theta \sim \mathcal{N}_d(\theta, \Sigma)$ with covariance matrix Σ known.
- The conjugate prior distribution is also normal: $\theta \sim \mathcal{N}_d(\mu, A)$ and the posterior distribution $p(\theta|x)$ is

$$\mathcal{N}_d(\mu + A(\Sigma + A)^{-1}(x - \mu), (A^{-1} + \Sigma^{-1})^{-1}).$$

- With a quadratic loss function, the Bayes estimator is the posterior mean

$$\delta^p(x) = (A^{-1} + \Sigma^{-1})^{-1}(\Sigma^{-1}x + A^{-1}\mu).$$

- For repeated observations of the normal model, $x_1, \ldots, x_n$, the sufficient statistic is $\bar{x} \sim \mathcal{N}_d(\theta, \frac{1}{n}\Sigma)$.

The normal model: unknown variance. I

Variance Estimation.

In many cases, the variance of the model is (partially) unknown $\Rightarrow$ prior on the parameters (θ, Σ).

- If the variance is known up to a multiplicative constant, σ^2, it is possible to go back to the unidimensional framework (i.e. $x_1, \ldots, x_n \sim i.i.d. \mathcal{N}(\theta, \sigma^2)$).
- Let $\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i$ et $s^2 := \sum_{i=1}^n (x_i - \bar{x})^2$, then the Bayes estimator depends only on $\bar{x}$ and σ^2 (sufficient statistic).
- Likelihood:

$$\ell(\theta, \sigma | \bar{x}, s^2) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(s^2 + n(\bar{x} - \theta)^2)\right].$$

The normal model: unknown variance. II

1) Jeffrey's prior: $\pi(\theta, \sigma) = 1/\sigma^2$.

- Posterior:

$$\begin{aligned} \theta|\sigma^2, \bar{x}, s^2 &\sim \mathcal{N}(\bar{x}, \sigma^2/n) \\ \sigma^2|\bar{x}, s^2 &\sim \mathcal{IG}\left(\frac{n-1}{2}, \frac{s^2}{2}\right). \end{aligned}$$

- The marginal posterior of σ^2 is the same as when θ is known. On the other side, the marginal posterior of θ is different:

$$\pi(\theta|\bar{x}, s^2) \propto \{s^2 + n(\bar{x} - \theta)^2\}^{-n/2}$$

i.e. $\theta|\bar{x}, s^2 \sim t_1(n-1, \bar{x}, s^2/n(n-1))$.

The normal model: unknown variance. III

2) Conjugate prior: $\pi(\theta, \sigma^2) = \pi_1(\theta|\sigma^2)\pi_2(\sigma^2)$ where $\pi_1 = \mathcal{N}(\theta_0, \sigma^2/n_0)$ and $\pi_2 = \mathcal{IG}(\nu/2, s_0^2/2)$. Remark: θ and σ^2 are not independent a priori.

- The posterior satisfies:

$$\pi(\theta, \sigma^2|x) \propto \sigma^{-n-\nu-3} \exp\left\{-\frac{1}{2}[s_1^2 + n_1(\theta - \theta_1)^2]/\sigma^2\right\}$$

  where $n_1 = n + n_0$, $\theta_1 = \frac{1}{n_1}(n_0\theta_0 + n\bar{x})$, $s_1^2 = s^2 + s_0^2 + \frac{(\theta_0 - \bar{x})^2}{n_0^{-1} + n^{-1}}$.

- These posterior distributions are conjugate since: $\theta|\sigma, \bar{x}, s^2 \sim \mathcal{N}(\theta_1, \sigma^2/n_1)$ and $\sigma^2|\bar{x}, s^2 \sim \mathcal{IG}(\nu_1/2, s_1^2/2)$, $\nu_1 = \nu + n$.

- The marginal posterior of θ is a Student t-distribution.

The normal model: unknown variance. IV

- n_0/n characterises the precision of the prior distribution, relatively to the precision of the observations.
- If $n_0/n \to 0$, we obtain the limit case: $\theta|\bar{x}, \sigma^2 \sim \mathcal{N}(\bar{x}, \sigma^2/n)$, which corresponds to the posterior associated with the Jeffrey's prior.
- The statistical inference based on the conjugate distributions necessitates a precise determination of the hyperparameters $(\theta_0, s_0^2, n_0, \nu)$.

The normal model: unknown variance. V

Variance estimation : (θ, Σ) **unknown**

- Let $x_1, \ldots, x_n$ be a random sample of $\mathcal{N}_d(\theta, \Sigma)$.
- Sufficient statistic: $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ and $S = \sum_{i=1}(x_i - \bar{x})(x_i - \bar{x})'$.
- Likelihood:

$$\ell(\theta, \Sigma|\bar{x}, S) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\left(n(\bar{x}-\theta)'\Sigma^{-1}(\bar{x}-\theta) + tr(\Sigma^{-1}S)\right)\right\}.$$

- Conjugate prior:

$$\begin{aligned} \theta|\Sigma &\sim \mathcal{N}_d(\theta_0, \Sigma/n_0) \\ \Sigma^{-1} &\sim \mathcal{W}_d(\alpha, W) \end{aligned}$$

where $\mathcal{W}_d$ denotes the Wishart distribution.

The normal model: unknown variance. VI

- Posterior:

$$\begin{aligned}\theta|\Sigma,\bar{x},S &\sim \mathcal{N}_d\left(\frac{n_0\theta_0+n\bar{x}}{n_0+n},\Sigma/(n_0+n)\right)\\ \Sigma^{-1}|\bar{x},S &\sim \mathcal{W}_d(\alpha+n,W_1(\bar{x},S)),\end{aligned}$$

  with $W_1(\bar{x},S)^{-1}=W^{-1}+S+\frac{nn_0}{n+n_0}(\bar{x}-\theta_0)(\bar{x}-\theta_0)'$.

- Jeffrey's prior: limit of the Wishart distribution $\mathcal{W}_d(\alpha,W)$ for Σ^{-1} when $\alpha\to 0$ and $W^{-1}\to 0$. It is given by:

$$\pi_J(\theta,\Sigma^{-1})=\frac{1}{|\Sigma|^{-(d+1)/2}}.$$

The normal model: summary

Let $D_n = (x_1, \ldots, x_n)$

I. $x_i \sim \mathcal{N}(\theta, \sigma^2)$, $\theta \sim \mathcal{N}(\theta_0, \sigma_0^2)$, σ^2 known.

II. $x_i \sim \mathcal{N}(\theta, \sigma^2)$, $\theta \sim \mathcal{N}(\theta_0, \sigma^2/\kappa_0)$, σ^2 known. Remark that κ_0 plays the role of n (equivalent sample size).

III. $x_i \sim \mathcal{N}(\theta, \sigma^2)$, $\theta \sim \mathcal{N}(\theta_0, \infty)$ (non-informative).

IV. $\lambda := \sigma^{-2}$ unknown. $x_i \sim \mathcal{N}(\theta, \sigma^2)$, $\theta|\sigma^2 \sim \mathcal{N}(\theta_0, \sigma^2/\kappa_0)$, $\lambda \sim Ga(\alpha_0, \beta_0)$.

V. $\lambda := \sigma^{-2}$ unknown. $x_i \sim \mathcal{N}(\theta, \sigma^2)$, $\theta|\sigma^2 \sim \mathcal{N}(\theta_0, \sigma^2/0)$, $\lambda \sim Ga(\alpha_0, 0)$ (non-informative, i.e. $\theta, \lambda \propto \lambda^{-1}$).

VI. $\lambda := \sigma^{-2}$ unknown, θ known. $x_i \sim \mathcal{N}(\theta, \sigma^2)$, $\lambda \sim Ga(\alpha_0, \beta_0)$.

VII. Different parametrization: σ^2 unknown. $x_i \sim \mathcal{N}(\theta, \sigma^2)$, $\theta|\sigma^2 \sim \mathcal{N}(\theta_0, \sigma^2 v_0)$, $\sigma^2 \sim IG(\alpha_0, \beta_0)$.

VIII. Σ^{-1} unknown. $x_i \sim \mathcal{N}_d(\theta, \Sigma)$, $\theta|\Sigma \sim \mathcal{N}_d(\theta_0, \Sigma)$, $\Sigma^{-1} \sim \mathcal{W}_d(\alpha, T)$.

IX. Σ^{-1} unknown. $x_i \sim \mathcal{N}_d(\theta, \Sigma)$, $\theta, \Sigma^{-1} \propto |\Sigma|^{(d+1)/2}$ (non-informative).

X. Different parametrization: Σ unknown. $x_i \sim \mathcal{N}_d(\theta, \Sigma)$, $\theta|\Sigma \sim \mathcal{N}_d(\theta_0, \Sigma/\kappa_0)$, $\Sigma \sim \mathcal{IW}_d(\alpha, \Lambda_0^{-1})$.

Outline

Linear Model and G-priors:

Let us consider the linear regression model:

$$y = X\beta + \epsilon, \qquad \epsilon \sim \mathcal{N}_n(0, \Sigma), \qquad y \in \mathbb{R}^n, \quad \beta \in \mathbb{R}^d, \tag{1}$$

which is conditional on X (X is a $n \times d$ matrix with stochastic rows x_i'). This only models the conditional distribution of $y|X$ rather than the joint distribution of (y, X).

1) Σ known:

- Sufficient statistic: $\widehat{\beta} := (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}y$. $\widehat{\beta}$ is the MLE and the OLS. Moreover, $\widehat{\beta}|X \sim \mathcal{N}_d(\beta_*, (X'\Sigma^{-1}X)^{-1})$.
- Conjugate prior: (Lindley & Smith (1972)) $\beta \sim \mathcal{N}_d(\beta_0, \Sigma_0)$ where $\beta_0 \in \mathbb{R}^d$, and $\Sigma_0 \in \mathbb{R}^{d\times d}$. Remark: Σ_0, β_0 may depend on X.

2) Σ unknown :

- Suppose independent observations are available.
- Likelihood:

$$\ell(\beta, \Sigma|y) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} tr\left[\Sigma^{-1} \sum_{i=1}^{n} (y_i - X_i\beta)(y_i - X_i\beta)'\right]\right\}.$$

2.1) **Jeffrey's prior:** $\pi(\beta, \Sigma) = |\Sigma|^{-(n+1)/2}$

2.2) Σ **known up to a finite dimensional parameter** σ^2. Then, it is possible to write the model as $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$.

- The OLS estimator satisfies: $\widehat{\beta}|X \sim \mathcal{N}_d(\beta_*, \sigma^2(X'X)^{-1})$.
- **Prior on σ^2:** $\sigma^2 \sim \mathcal{IG}(\frac{v_0}{2}, \frac{v_0\sigma_0^2}{2})$. The posterior is:

$$\sigma^2|\widehat{\beta}, s^2, \beta, X \sim \mathcal{IG}(\frac{v_0 + n}{2}, \frac{v_0\sigma_0^2 + SSR(\beta)}{2}).$$

Multivariate Regression. III

- **Posterior of $\beta|\widehat{\beta}, s^2, \sigma^2, X$:**

$$\beta|\widehat{\beta}, s^2, \sigma^2, X \sim \mathcal{N}_d(V(\Sigma_0^{-1}\beta_0 + X^t y/\sigma^2), V)$$

  where $V = (\Sigma_0^{-1} + X^t X/\sigma^2)^{-1}$.

- **Gibbs sampler to approximate the joint posterior distribution** $\pi(\beta, \sigma^2|y, X)$**:** Given current values $\{\beta^{(s)}, \sigma^{2(s)}\}$, new values can be generated by

  1. updating β:
     a) compute $V = Var[\beta|y, X, \sigma^{2(s)}]$ and $m = \mathbf{E}[\beta|y, X, \sigma^{2(s)}]$
     b) sample $\beta^{(s+1)} \sim \mathcal{N}(m, V)$
  2. updating σ^2:
     a) compute $SSR(\beta^{(s+1)})$
     b) sample $\sigma^{2(s+1)} \sim \mathcal{IG}\left(\frac{v_0+n}{2}, \frac{v_0\sigma_0^2 + SSR(\beta^{(s+1)})}{2}\right)$.

Multivariate Regression. IV

Bayesian analysis of a regression model requires specification of the prior parameters (β_0, Σ_0) and (v_0, σ_0^2). Finding values of these parameters that represent actual prior information can be difficult.

- One idea is that, if the prior is not representing real prior information about the parameters, then it should be as minimally informative as possible.
  - The resulting posterior would represent the posterior information of someone who began with little knowledge of the population being studied.
  - *Unit information prior* (Kass and Wasserman, 1995): it sets $\Sigma_0^{-1} = (X^T X)/(n\sigma^2)$ and $\beta_0 = \widehat{\beta}_{ols}$. Similarly: $v_0 = 1$ and $\sigma_0^2 = \widehat{\sigma}_{ols}^2$.
- Another principle for constructing a prior distribution for β is based on the idea that the parameter estimation should be invariant to changes in the scale of the regressors. This condition will be met if $\beta_0 = 0$ and $\Sigma_0 = k(X^T X)^{-1}$, for any $k > 0$.
- Conjugate prior (G-prior, Zellner 1971, 1986): $k = \sigma^2/n_0$

$$\begin{aligned} \beta|\sigma^2 &\sim \mathcal{N}_d\left(\beta_0, \frac{\sigma^2}{n_0}(X'X)^{-1}\right) \\ \sigma^2 &\sim \mathcal{IG}(v_0/2, s_0^2/2). \end{aligned}$$

Multivariate Regression. V

- Posterior:

$$\begin{aligned}
\beta|\widehat{\beta}, s^2, \sigma^2, X &\sim \mathcal{N}_d\left(\frac{n_0\beta_0 + \widehat{\beta}}{n_0+1}, \frac{\sigma^2}{n_0+1}(X'X)^{-1}\right) \\
\sigma^2|\widehat{\beta}, s^2, X &\sim \mathcal{IG}\left(\frac{n+\nu_0}{2}, \frac{s^2 + s_0^2 + \frac{n_0}{n_0+1}(\beta_0 - \widehat{\beta})'X'X(\beta_0 - \widehat{\beta})}{2}\right).
\end{aligned}$$

- The marginal distribution of β is a multivariate t distribution.

- The regression model is completely conditional on the explanatory variables X. The G-prior can be seen as a posterior w.r.t. X.

- The G-prior is adequate to take into account problems of multicollinearity, since it allows to assign a larger prior variance to the component affected by multicollinearity.

Multiple Regression. I

Consider a set of regression equations related through common X variables and correlated errors:

$$\begin{aligned} y_1 &= X\beta_1 + \varepsilon_1 \\ &\vdots \\ y_c &= X\beta_c + \varepsilon_c \\ &\vdots \\ y_m &= X\beta_m + \varepsilon_m. \end{aligned} \tag{2}$$

- For $r = 1, \ldots, n$, let y_r and ε_r be m-vectors of the r-th observation on each of the dependent variables and error term.
- Assume

$$p(\varepsilon_1, \ldots, \varepsilon_n|\Sigma) \propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} tr S_\varepsilon \Sigma^{-1}\right\}, \qquad S_\varepsilon = \sum_{r=1}^{n} \varepsilon_r \varepsilon_r'.$$

Multiple Regression. II

- In matrix form we have

$$Y = XB + E, \qquad B = [\beta_1, \ldots, \beta_c, \ldots, \beta_m] \tag{3}$$

  where Y and E are $n \times m$ matrices of observations, X is an $n \times k$ matrix of observations on k common independent variables.

- Remark that: $E'E = S_\varepsilon$.

So, the likelihood is:

$$\begin{aligned} p(Y|X, B, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} tr(Y - XB)'(Y - XB)\Sigma^{-1}\right\} \\ &= |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} tr\left(S + (B - \widehat{B})'X'X(B - \widehat{B})\right)\Sigma^{-1}\right\} \end{aligned}$$

with $S = (Y - X\widehat{B})'(Y - X\widehat{B})$ and $\widehat{B} = (X'X)^{-1}X'Y$.

Multiple Regression. III

- The natural conjugate prior is an inverted Wishart on Σ and a normal prior on B conditional on Σ:

$$\begin{aligned} p(\Sigma, B) &= p(\Sigma)p(B|\Sigma) \\ \Sigma &\sim \mathcal{IW}(\nu_0, V_0) \\ \beta|\Sigma &\sim \mathcal{N}(\beta_0, \Sigma \otimes A^{-1}). \end{aligned}$$

- The posterior is:

$$\begin{aligned} \Sigma|Y, X &\sim \mathcal{IW}(\nu_0 + n, V_0 + S) \\ \beta|Y, X, \Sigma &\sim \mathcal{N}(\widetilde{\beta}, \Sigma \otimes (X'X + A)^{-1}) \\ \widetilde{\beta} &= vec(\widetilde{B}), \qquad \widetilde{B} = (X'X + A)^{-1}(X'X\widehat{B} + AB_0) \\ S &= (Y - X\widetilde{B})'(Y - X\widetilde{B}) + (\widetilde{B} - B_0)'A(\widetilde{B} - B_0). \end{aligned}$$

Outline

Régions α-crédibles. I

- The Bayesian analogs of confidence sets are called credible sets and are derived from the posterior distribution.
- Rationale behind the definition of credible sets: we look for a subset C of the model that is as small as possible while receiving a certain minimal posterior probability.

The Bayesian equivalent of the frequentist confidence interval is the credible region.

Definition

Choose a level $\alpha \in (0,1)$. A level α credible set for θ is a subset $C_\alpha \subset \Theta$ such that

$$1 - \alpha \leq \pi(\theta \in C_\alpha | x). \tag{4}$$

Therefore, we can talk about the probability that θ is in C_α.

To find credible sets in practice: first one has to calculate the posterior distribution and, based on that, to derive a subset C_α that satisfy (4).

Régions α-crédibles. II

Definition

Choose a level $\alpha \in (0,1)$. *If for all* $n \geq 1$, $X^{(n)} = (X_1, \ldots, X_n)$ *is i.i.d.* P_*, *a sequence of subsets* $C_{\alpha,n} \subset \Theta$ *such that*

$$\liminf_{n \to \infty} \pi(\theta \in C_{\alpha,n} | x) \geq 1 - \alpha$$

P_**-a.s., is called an asymptotic credible set of level* α.

Remark:

- In smooth, parametric models for *i.i.d.* data there is a close, asymptotic relation between BCS and FCS centred on the maximum-likelihood estimator: the Bernstein-von Mises theorem implies that level-α BCS coincide with the FCS asymptotically.

- In situations where it is hard to calculate the MLE or to construct the corresponding FCS explicitly, it is sometimes relatively easy to obtain BCS. In such cases, one can calculate BCS and conveniently interpret them as FCS.

Régions α-crédibles. III

For a given $\alpha \in (0,1)$, $\exists$ many sets that satisfy (4): we prefer smaller sets over large ones. Suppose the posterior is dominated with density $\pi(\theta|X)$ and define, for every $k \geq 0$, the level-set

$$C(k) = \{\theta \in \Theta;\ \pi(\theta|X) \geq k\}.$$

Definition

A level α HPD (Highest Posterior Density) credible region for θ is a subset $C_\alpha = C(k_\alpha) \subset \Theta$ where k_α equals

$$k_\alpha = \sup\{k \geq 0;\ \pi(\theta \in C(k)|X) \geq 1-\alpha\}.$$

Outline

Tests d'hypothèses. I

- H_0 et H_1 peuvent être considerées: (1) soit comme deux régions (i.e. une partition) de l'espace paramétrique d'un unique modèle d'échantillonage, (2) soit comme deux modèles d'échantillonage différents.

- Deux points de vue:

  - Une *statistique de test* est un procédé statistique à valeurs dans un espace à deux points: "accepter" et "rejeter" une hypothèse.
  - les tests d'hypothèses peuvent aussi être considerés comme une façon pour les statisticien de gérer ses doutes relatifs à son modèle statistique.

Tests d'hypothèses. II

- Nous avons un espace de décisions avec deux points: $\mathscr{D} = \{\delta_0, \delta_1\}$ et une fonction de perte $L(\theta, \delta)$.
- On peut partitionner en deux classes l'ensemble des états de la nature:

$$\Theta = \Theta_0 \cup \Theta_1$$

où Θ_0 et Θ_1 sont définis:

$$\begin{aligned} \Theta_0 &= \{\theta;\, L(\theta, \delta_0) = 0\} \\ \Theta_1 &= \{\theta;\, L(\theta, \delta_1) = 0\}. \end{aligned}$$

- La specification de la fonction de perte est alors complétée comme suit:

$$L(\theta, \delta) = \begin{cases} L_1(\theta) & \text{si } \delta = \delta_1 \text{ et } \theta \in \Theta_0 \\ L_0(\theta) & \text{si } \delta = \delta_0 \text{ et } \theta \in \Theta_1, \end{cases}$$

c'est-à-dire $L(\theta, \delta_0) = \mathbb{1}_{\Theta_1}(\theta) L_0(\theta)$ et $L(\theta, \delta_1) = \mathbb{1}_{\Theta_0}(\theta) L_1(\theta)$.

Tests d'hypothèses. III

- We obtain

	Θ_0	Θ_1
δ_0	0	$L_0(\theta)$
δ_1	$L_1(\theta)$	0

- Particular case: $L_j(\theta) = L_j, j = 0, 1$ (constant loss function on the elements of the partition of the state of the nature).

- When Θ becomes the parametric space of a statistical model, the element of the partition $\Theta = \Theta_0 \cup \Theta_1$ are called statistical hypothesis.

- Neyman and Pearson approach: H_0 is chosen such that the error of first type is the most important.

- The Bayesian analysis does not require such a specification and treats H_0 and H_1 symmetrically.

Tests d'hypothèses. IV

- Let $x^{(n)} := (x_1, \ldots, x_n)$ be the observation of an i.i.d. sample of $X \in \mathcal{X}$.
- The optimal posterior decision is defined by:

$$\begin{aligned} \delta^*(x^{(n)}) &= \arg\min_{\delta \in \mathcal{D}} \mathbf{E}[L(\theta, \delta)|x^{(n)}] \\ &= \arg\min\{\rho(\pi, \delta_0), \rho(\pi, \delta_1)\} \end{aligned}$$

  where $\rho(\pi, \delta) := \mathbf{E}[L(\theta, \delta)|x^{(n)}]$ is the posterior risk of the decision δ.

- In the particular case $L_j(\theta) = L_j, j = 0, 1$, we define: $\pi(\theta \in \Theta_0|x^{(n)}) = p(x^{(n)})$ and then we can write:

$$\begin{aligned} \mathbf{E}[L(\theta, \delta_0)|x^{(n)}] &= L_0 \times (1 - p(x^{(n)})) \\ \mathbf{E}[L(\theta, \delta_1)|x^{(n)}] &= L_1 \times p(x^{(n)}). \end{aligned}$$

  The optimal decision rule becomes:

$$\delta^*(x^{(n)}) = \delta_0 \quad \Leftrightarrow \quad L_0 \times (1 - p(x^{(n)})) < L_1 p(x^{(n)}).$$

  We can also write the optimal decision rule in terms of the *odds ratio*:

$$\delta^*(x^{(n)}) = \delta_0 \quad \Leftrightarrow \quad \frac{p(x^{(n)})}{1 - p(x^{(n)})} > \frac{L_0}{L_1}.$$

Tests d'hypothèses. V

- For instance: if $L_1 = 19L_0$, then $\delta^*(x^{(n)}) = \delta_1 \quad \Leftrightarrow \quad \frac{p(x^{(n)})}{1-p(x^{(n)})} < \frac{1}{19}$.
- So, in general in Bayesian analyses, we simply compute $\pi(\Theta_0|x^{(n)})$ and $\pi(\Theta_1|x^{(n)})$ and then we take the decision. These probabilities are the subjective probabilities of the hypothesis based on the data and the prior.
- The Bayes rule for a $0-1$ loss consists in choosing the hypothesis with the higher posterior probability.
- Another important tool used in the testing problem is the *Bayes Factor*:

Definition

Let $\{\Theta_0, \Theta_1\}$ be a measurable partition of Θ such that $\pi(\Theta_0) > 0$ and $\pi(\Theta_1) > 0$. Let $\pi(\Theta_0|x^{(n)})/\pi(\Theta_1|x^{(n)})$ be the posterior odds and $\pi(\Theta_0)/\pi(\Theta_1)$ be the prior odds. The quantity

$$B_{01} = \frac{\textit{posterior odds ratio}}{\textit{prior odds ratio}} = \frac{\pi(\Theta_0|x^{(n)})\pi(\Theta_1)}{\pi(\Theta_1|x^{(n)})\pi(\Theta_0)}$$

is called the Bayes factor in favour of Θ_0.

Tests d'hypothèses. VI

- The smaller is the value of B_{01}, the stronger is the evidence against H_0.
- If the hypothesis are simple hypothesis (i.e. $\Theta_j = \{\theta_j\}, j = 0, 1$) and therefore, $\Theta = \{\theta_0, \theta_1\}$, the Bayes factor is exactly equal to the likelihood ratio and so it is independent of the prior.
- In general, B_{01} depends on the prior, but it does not depend on the relative prior weights of Θ_0 and Θ_1. Moreover, $B_{10} = 1/B_{01}$.
- When doing Bayesian hypothesis testing, we have a choice of which ratio to use and that choice will correspond directly with a choice for subjectivist or objectivist philosophies.
- In the **subjectivist**'s view, the posterior odds ratio has a clear interpretation: if

$$\frac{\pi(\Theta_0|x^{(n)})}{\pi(\Theta_1|x^{(n)})} > 1$$

hence, the subjectivist decides to adopt H_0 rather than H_1 (and vice-versa).

Tests d'hypothèses. VII

- The **objectivist** prefers the Bayes factor to make a choice between two hypotheses: if $B_{01} > 1$ the objectivist adopts H_0 rather than H_1; if, on the other hand, $B_{01} < 1$, then the objectivist adopts H_1 rather than H_0.
- If our view of H_0 is like in the frequentist approach (i.e. H_0 should not be rejected except if there is enough evidence for the contrary) then, it is raisonable to assign more prior probability to H_0 than to H_1. An objective choice would be to assign equal prior probabilities.
- We can do this with the following prior specification.
- Prior:

$$\pi(\theta) = \begin{cases} \pi_0 g_0(\theta) & \text{if } \theta \in \Theta_0 \\ \pi_1 g_1(\theta) & \text{if } \theta \in \Theta_1 \end{cases} \tag{5}$$

where $\pi_j = \pi(\Theta_j), j = 0, 1$, $\pi_1 = 1 - \pi_0$ and g_0 and g_1 are proper densities on Θ_0 and Θ_1, respectively. Then,

$$\pi(\theta) = \pi_0 g_0(\theta) \mathbb{1}_{\Theta_0}(\theta) + (1 - \pi_0) g_1(\theta) \mathbb{1}_{\Theta_1}(\theta).$$

- Then, we can write the posterior odds ratio:

$$\frac{\pi(\Theta_0|x^{(n)})}{\pi(\Theta_1|x^{(n)})} = \frac{\int_{\Theta_0}\pi(\theta|x^{(n)})d\theta}{\int_{\Theta_1}\pi(\theta|x^{(n)})d\theta} = \frac{\int_{\Theta_0}f(x^{(n)}|\theta)\pi_0 g_0(\theta)d\theta/m(x^{(n)})}{\int_{\Theta_1}f(x^{(n)}|\theta)\pi_1 g_1(\theta)d\theta/m(x^{(n)})}$$
$$= \frac{\pi_0\int_{\Theta_0}f(x^{(n)}|\theta)g_0(\theta)d\theta}{\pi_1\int_{\Theta_1}f(x^{(n)}|\theta)g_1(\theta)d\theta}$$

and the Bayes factor:

$$B_{01} = \frac{\int_{\Theta_0}f(x^{(n)}|\theta)g_0(\theta)d\theta}{\int_{\Theta_1}f(x^{(n)}|\theta)g_1(\theta)d\theta}$$

- So, the posterior odds ratio is equal to:

$$\frac{\pi_0}{1-\pi_0}B_{01}$$

and becomes equal to B_{01} if $\pi_0 = 1/2$.

Exemple A I

- Consider a blood test conducted for determining the sugar level of a person with diabetes two hours after he had his breakfast.
- We want to see if his medication has controlled his blood sugar levels.
- Assume that the test result X is $\mathcal{N}(\theta, 100)$, where θ is the true level.
- In the appropriate population (diabetic but under this treatment), $\theta \sim \mathcal{N}(100, 900)$.
- Then, marginally $X \sim \mathcal{N}(100, 1000)$, and the posterior distribution is

$$\theta | X = x \sim \mathcal{N}(0.9x + 10, 90).$$

- We want to test:

$$H_0 : \theta \leq 130$$
$$H_1 : \theta > 130.$$

- If the blood test shows a sugar level of 130, what can be concluded?

Exemple A II

- Given this test result, the posterior is $\mathcal{N}(127, 90)$. Consequently:

$$\begin{aligned}\pi(\theta \leq 130 | X = 130) &= \Phi\left(\frac{130 - 127}{\sqrt{90}}\right) = \Phi(.316) = 0.624 \\ \pi(\theta > 130 | X = 130) &= 0.376.\end{aligned}$$

  Therefore, the posterior odds ratio is: $0.624/0.376 = 1.66$.

- Because $\pi_0 = \Phi\left(\frac{130-100}{30}\right) = \Phi(1)$, the prior odds ratio is $\Phi(1)/(1 - \Phi(1)) = 0.8413/0.1587 = 5.3$ and thus the Bayes factor is

$$BF_{01} = \frac{1.66}{5.3} = 0.313.$$

- It can also be noted here that in one-sided testing situations when a continuous prior π can be specified readily for the entire parameter space, there is no need to express it in the form of $\pi(\theta) = \pi_0 g_0 \mathbb{1}_{\Theta_0}(\theta) + (1 - \pi_0) g_1(\theta) \mathbb{1}_{\Theta_1}(\theta)$. However, the problem of testing a point null hypothesis turns out to be quite different.

Exemple B I

- Mister A is interested in determining his true weight from a variable bathroom scale.
- Assume the measurements are $X_i \sim \mathcal{N}(\mu, 9)$.
- Sample (measurements in pounds):
  $182, 172, 173, 176, 176, 180, 173, 174, 179, 175.$
- μ=Mister A's true weight
- Suppose Mister A is interested in assessing if his true weight is more than 175 pounds. He wishes to test the hypotheses

$$H_0 : \mu \leq 175$$
$$H_1 : \mu > 175.$$

- Prior: $\mu \sim \mathcal{N}(170, 5)$.

Exemple B II

- The prior odds of H_0 is given by

$$\frac{\pi_0}{\pi_1} = \frac{P(\mu \leq 175)}{P(\mu > 175)}.$$

```
> pmean=170; pvar=25
> probH=pnorm(175,pmean,sqrt(pvar))
> probA=1-probH
> prior.odds=probH/probA
> prior.odds
[1] 5.302974
```

- So, a priori, H_0 is five times more likely than H_1.
- We enter the ten weight measurements into R and compute the sample mean $\bar{y}$ and the associated sampling variance σ^2/n:

```
> weights=c(182, 172, 173, 176, 176, 180, 173, 174, 179, 175)
> ybar=mean(weights)
> sigma2 = 3^2/length(weights)
```

Exemple B III

- The posterior precision of μ is the sum of the precisions of the data and the prior:

```
> post.precision=1/sigma2+1/pvar
> post.var=1/post.precision
```

- The posterior mean of μ is the weighted average of the sample mean and the prior mean, where the weights are proportional to the respective precisions:

```
>
post.mean=(ybar/sigma2+pmean/pvar)/post.precision
> c(post.mean,sqrt(post.var))
[1] 175.7915058   0.9320547
```

- The posterior density of μ is $\mathcal{N}(175.79, 0.93)$.

Exemple B IV

- Using this normal posterior density, we calculate the odds of H_0:

```
 >
post.odds=pnorm(175,post.mean,sqrt(post.var))/
+ (1-pnorm(175,post.mean,sqrt(post.var)))
> post.odds
[1] 0.2467017
```

- So, the BF_{01} in support of H_0 is

```
 > BF = post.odds/prior.odds
> BF
[1] 0.04652139
```

- From the prior probabilities and the Bayes factor, we can compute the posterior probability of H_0:

```
 > postH=probH*BF/(probH*BF+probA)
> postH
[1] 0.1978835
```

- Based on this calculation, we can conclude that it is unlikely that Mister A's weight is at most 175 pounds.

Test d'une hypothèses nulle ponctuelle I

La loi a priori définie en (5) est utile si on veut tester une hypothèse nulle ponctuelle.

- Une hypothèse nulle ponctuelle $H_0 : \theta = \theta_0$ (contre $H_1 : \theta \neq \theta_0$) ne peut pas être testée sous une loi a priori continue.

- De plus, le facteur de Bayes n'est défini que lorsque $\pi_0 \neq 0$ et $\pi_1 \neq 0$. Cela implique que, si H_0 ou H_1 sont a priori impossibles, les observations ne vont pas modifier cette information absolue: des probabilités nulles a priori le restent a posteriori.

On peut utiliser la priori (5). Cette modification de la loi a priori est surprenante, puisqu'elle revient à mettre un poids a priori sur un ensemble de mesure 0:

- Une probabilité $\pi_0 > 0$ doit être assignée au point θ_0 et $(1 - \pi_0)$ doit être répandue sur $\{\theta \neq \theta_0\}$ utilisant une densité g_1.
- g_0 est alors prise égale à un point masse sur θ_0.

Test d'une hypothèses nulle ponctuelle II

- Alors on a que $\pi(\theta)$ a une partie continue et une parti discrète:

$$\pi(\theta) = \pi_0 \mathbb{1}_{\theta_0}(\theta) + (1-\pi_0) g_1(\theta) \mathbb{1}_{\theta \neq \theta_0}(\theta).$$

- Puisque:

$$\begin{aligned}
\pi(\theta_0|x^{(n)}) &= \frac{\pi_0 f(x^{(n)}|\theta_0)}{\pi_0 f(x^{(n)}|\theta_0) + (1-\pi_0) \underbrace{\int_{\theta \neq \theta_0} f(x^{(n)}|\theta) g_1(\theta) d\theta}_{=: m_1(x^{(n)})}} \\
&= \left(1 + \frac{1-\pi_0}{\pi_0} \frac{m_1(x^{(n)})}{f(x^{(n)}|\theta_0)}\right)^{-1}
\end{aligned}$$

le posterior odds ratio devient

$$\frac{\pi(\theta_0|x^{(n)})}{1-\pi(\theta_0|x^{(n)})} = \frac{\pi_0 f(x^{(n)}|\theta_0)}{(1-\pi_0) m_1(x^{(n)})}$$

Test d'une hypothèses nulle ponctuelle III

et le facteur de Bayes est:

$$B_{01} = \frac{f(x^{(n)}|\theta_0)}{m_1(x^{(n)})}.$$

Markov Chain Monte Carlo (MCMC) Methods

Markov Chain Monte Carlo (MCMC) Methods. I

- Given a model (prior and likelihood), the computational phase of Bayesian inference requires practical methods for summarizing/exploring the posterior distribution.

- In many cases, the posterior distribution is represented by an unnormalized density, $\pi^*(\theta)$, and the problem is to construct simulation-based estimates of various aspects of this distribution.

- The idea behind MCMC methods is to formulate a Markov chain on the parameter space. A Markov chain is a sequence of random variables that can be thought of as evolving over time, with probability of a transition depending on the particular set in which the chain is.
  $\Rightarrow$ define the chain in terms of its transition kernel.

- If care is taken to ensure that this chain has the posterior as its equilibrium (or stationary) distribution, then the chain can be used to construct simulation-based estimates of the required integrals. (A probability distribution π is stationary if $X_n \sim \pi$ implies $X_{n+1} \sim \pi$).

- Starting from some point in the parameter space, we simulate the chain forward.

Markov Chain Monte Carlo (MCMC) Methods. II

- A Markov chain specifies a method for generating a sequence of random variables $\{\theta_1, \theta_2, \ldots, \theta_r, \ldots\}$ starting from initial point θ_0.
- This sequence is created by specifying a way of transitioning from θ_r to θ_{r+1}.
- This transition process is specified by choosing the conditional distribution,

$$\theta_{r+1}|\theta_r \sim F(\theta_r).$$

- Then,
  1. Start from θ_0;
  2. draw $\theta_1 \sim F(\theta_0)$;
  3. replace θ_0 with θ_1 and repeat a total of R times.
- Under some conditions on the conditional distribution F, the distribution of $\theta_r|\theta_0$ will converge to a fixed and unique distribution (stationary distribution) as $r \to \infty$.

Markov Chain Monte Carlo Methods. III

To practically apply MCMC methods we need to provide:

1. methods or algorithms for specifying chains with the right stationary distribution (this amounts to specifying the conditional distribution of $\theta_{r+1}|\theta_r$ using information about the posterior of θ);

2. theoretical assurance that the methods in 1 will produce ergodic chains;

3. practical guidance on convergence.

Échantillonnage de Gibbs (Gibbs Sampler). I

- The Gibbs sampler is a Markov chain obtained by cycling through a set of conditional distributions of the posterior $\pi(\cdot|D_n)$.

- If we break θ into p separate blocks of parameters,

$$\theta' = (\theta_1, \theta_2, \ldots, \theta_p)$$

then the Gibbs sampler is defined by iterative sampling from each of these π conditional distributions:

Échantillonnage de Gibbs (Gibbs Sampler). II

Gibbs Sampler

1. Set $\theta^{(0)}$;
2. Sample from

$$\begin{aligned}
\theta_1^{(1)} &\sim \pi(\theta_1|\theta_2^{(0)},\ldots,\theta_p^{(0)},D_n)\\
\theta_2^{(1)} &\sim \pi(\theta_2|\theta_1^{(1)},\theta_3^{(0)},\ldots,\theta_p^{(0)},D_n)\\
&\vdots\\
\theta_p^{(1)} &\sim \pi(\theta_p|\theta_1^{(1)}\theta_2^{(1)},\ldots,\theta_{p-1}^{(1)},D_n)
\end{aligned}$$

   to obtain the first iterate;
3. Repeat as necessary.

Échantillonnage de Gibbs (Gibbs Sampler). III

- Implementation of the Gibbs sampler requires the ability to sample from the set of conditional posterior distributions.
- In many situations, it is possible to define the groups or blocks of parameters so that the conditional distributions are of known form.
- As a default alternative, one could always define a Gibbs sampler based on the k univariate conditionals implied by π.
- The Gibbs sampler defined by the previous algorithm is a Markov chain.
- It is also easy to verify that the invariant distribution of this chain is π.
- Convergence of the Gibbs sampler justifies the practical use of the Gibbs sampler to start from an arbitrary initial condition and use sample averages to approximate integrals of the posterior.

Example: bivariate normal Gibbs sampler. I

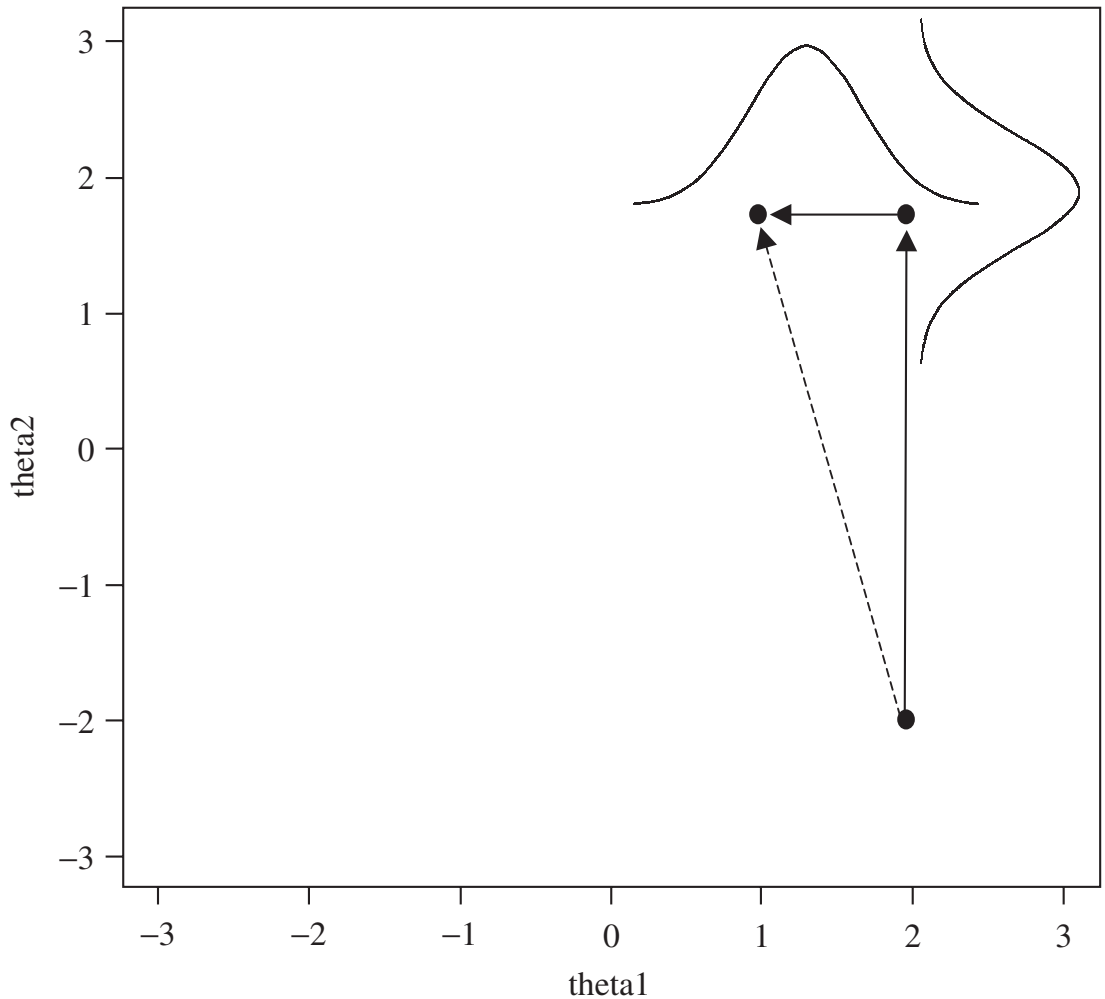
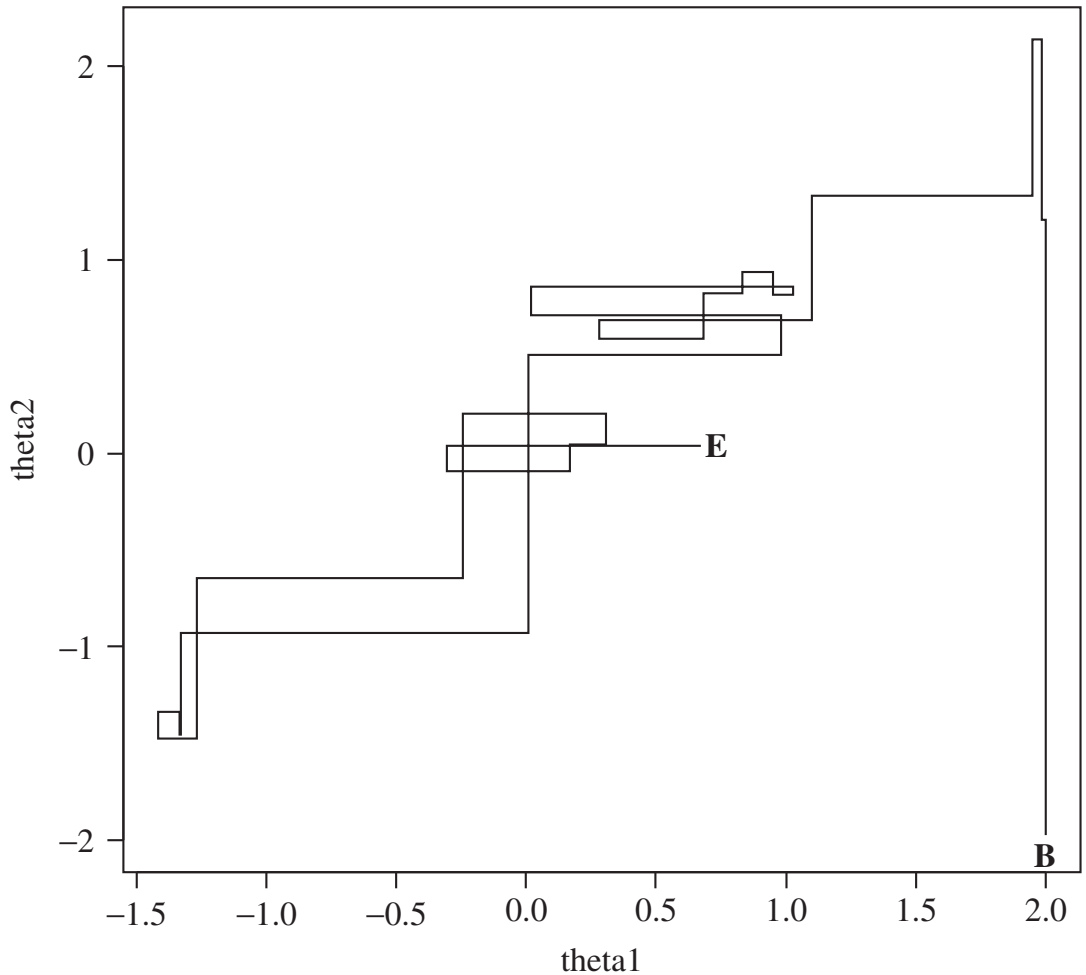


Figure 3.1 Functioning of bivariate normal Gibbs sampler

Example: bivariate normal Gibbs sampler. II



Twenty draws from bivariate Gibbs sampler showing intermediate moves

Example: bivariate normal Gibbs sampler. III

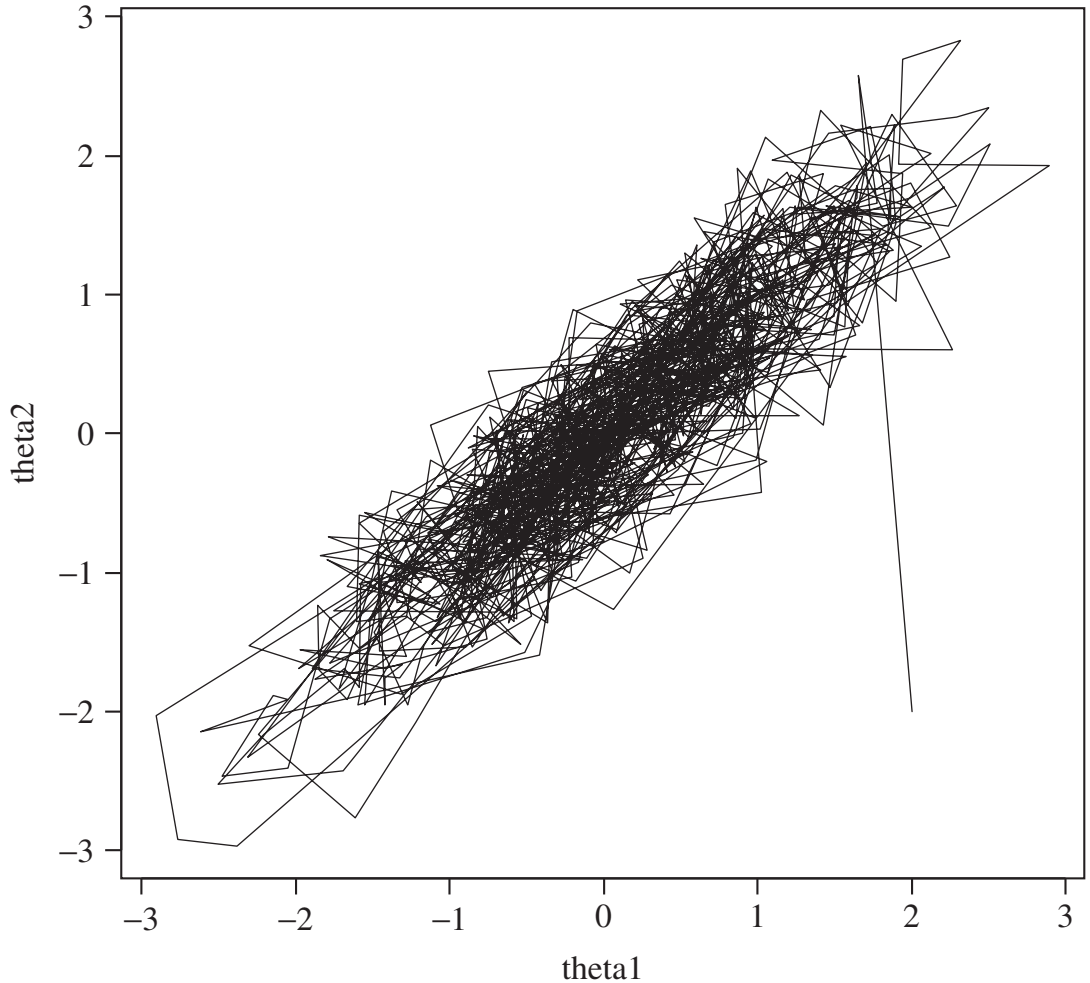


Figure 3.3 One thousand draws from bivariate Gibbs sampler

Example: bivariate normal Gibbs sampler. IV

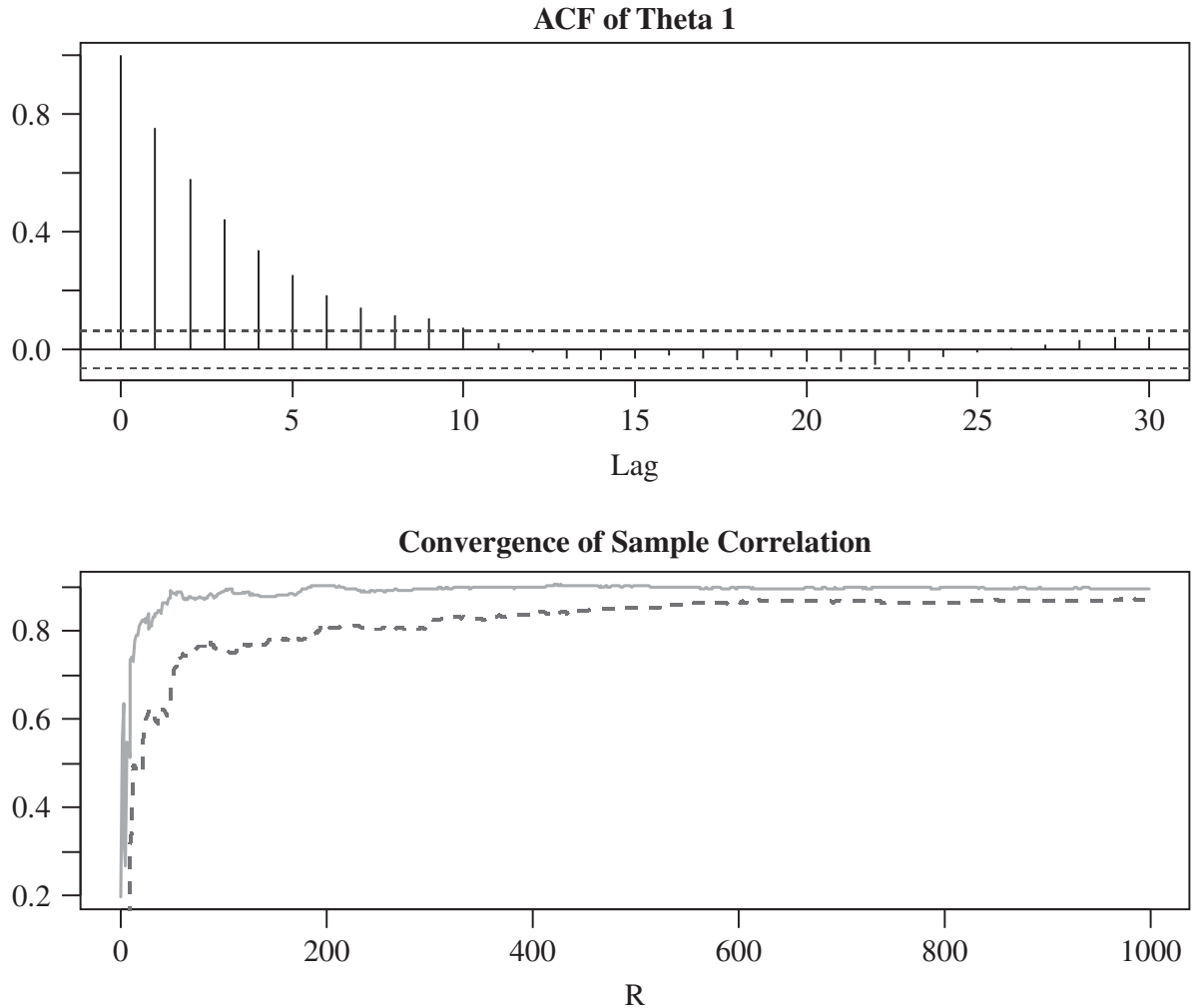


Figure 3.4 Autocorrelation function and illustration of ergodicity

Data Augmentation and Probit Model. I

- The Gibbs sampler can be applied to a much wider class of models once the principle of data augmentation is introduced.
- The idea that missing values are unobserved and, therefore, should properly be considered as part of the parameter vector comes naturally to a Bayesian.
- The idea of data augmentation extends to any situation in which there are unobservable constructs (Tanner and Wong (1987)).
- To illustrate the usefulness of the data augmentation concept for discrete dependent variables models, consider the latent variable formulation of the binary probit model:

$$
\begin{aligned}
z_i &= x_{i1}\beta_1 + \ldots + x_{ik}\beta_k + \epsilon_i, \qquad \epsilon_i \sim i.i.d.\mathcal{N}(0,1) \\
y_i &= \begin{cases} 0, & \text{if } z_i < 0 \\ 1, & \text{otherwise .} \end{cases}
\end{aligned}
$$

- Prior: $\beta \sim \mathcal{N}(\overline{\beta}, A^{-1})$.

Data Augmentation and Probit Model. II

- Data augmentation proceeds by considering the entire n-dimensional vector of z values as part of the parameter vector: $\theta' = (z', \beta')$. So,

$$\pi(z, \beta|X) = \pi(z|\beta, X)\pi(\beta).$$

- The posterior can easily be computed by using a Gibbs sampler:

$$z|\beta, X, y$$
$$\beta|z, X.$$

The Gibbs sampling algorithm is constructed by sampling from the joint posterior distribution of (z, β):

- $\beta|z, \{y_i\} \sim \mathcal{N}_k \left((X'X)^{-1}X'Z, (X'X)^{-1}\right)$;

Data Augmentation and Probit Model. III

- if we are given a value of β, then $(z_1, \ldots, z_n)$ are independent, with (truncated Normal distributions)

$$\begin{aligned} z_i|\beta, \{y_i\} &\sim \mathcal{N}(\beta' x_i, 1) I(z_i > 0), \qquad \text{if } y_i = 1 \\ z_i|\beta, \{y_i\} &\sim \mathcal{N}(\beta' x_i, 1) I(z_i < 0), \qquad \text{if } y_i = 0. \end{aligned}$$

Metropolis Algorithms. I

- The Metropolis class of algorithms is a general-purpose approach to producing Markov chain samplers.
- The idea of the Metropolis approach is to generate a Markov chain with the posterior $\pi(\theta|D_n)$ as its invariant distribution by appropriate modifications to a related Markov chain that is relatively easy to simulate from.
- Let Q be a transition matrix which we want to modify to ensure that the resultant chain has a stationary distribution given by the vector π. The dimension of the state space is d.
- The continuous state space Metropolis algorithm starts with a proposal transition kernel defined by the transition function $q(\theta, \cdot)$. Given θ, $q(\theta, \cdot)$ is a density.

Metropolis Algorithms. II

Continuous State Space Metropolis

- Start at $\theta^{(0)}$.
- Draw $\vartheta \sim q(\theta^{(0)}, \cdot)$.
- Compute $\alpha(\theta^{(0)}, \vartheta) = \min\left\{1, \frac{\pi(\vartheta|D_n)q(\vartheta,\theta^{(0)})}{(\pi(\theta^{(0)}|D_n)q(\theta^{(0)},\vartheta))}\right\}$
- With probability α, $\theta^{(1)} = \vartheta$, else $\theta^{(1)} = \theta^{(0)}$.
- Repeat, as necessary.

Metropolis Algorithms. III

Random-Walk Metropolis Chains:

- Use a random walk to generate proposal values:

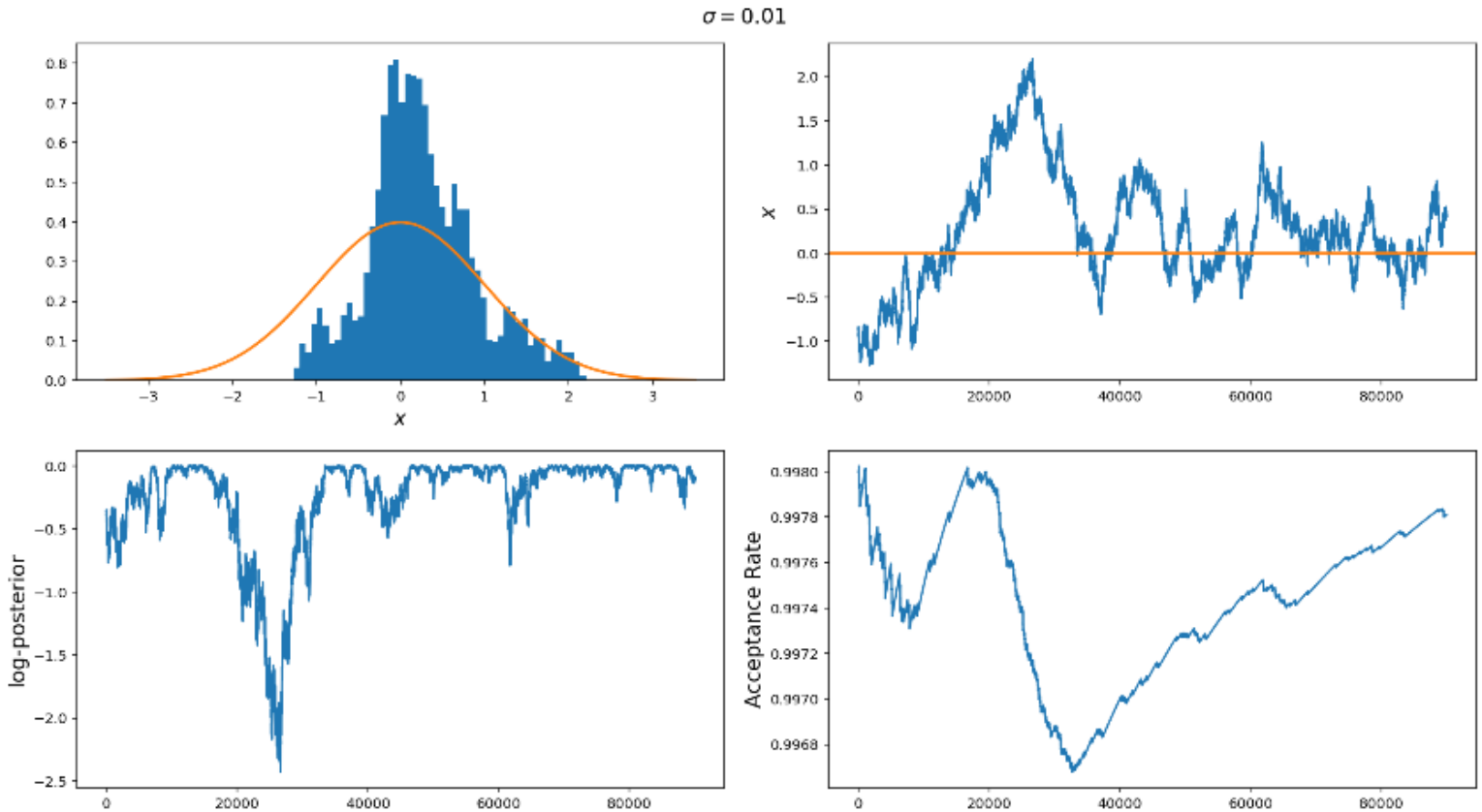
$$\vartheta = \theta + \varepsilon$$

which corresponds to the proposal transition function $q(\theta, \vartheta) = q_\varepsilon(\vartheta - \theta)$ which is symmetric.

Gaussian Random-Walk Metropolis

- Start at $\theta^{(0)}$.
- Draw $\vartheta = \theta^{(0)} + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, s^2\Sigma)$.
- Compute $\alpha(\theta^{(0)}, \vartheta) = \min\left\{1, \frac{p(\vartheta|D_n))}{p(\theta^{(0)}|D_n)}\right\}$
- With probability α, $\theta^{(1)} = \vartheta$, else $\theta^{(1)} = \theta^{(0)}$.
- Repeat, as necessary.

Diagnosing efficiency and convergence



The panels are: Histogram of samples, sample vs iteration (i.e. trace plot), log-posterior vs iteration, and proposal acceptance rate vs iteration.

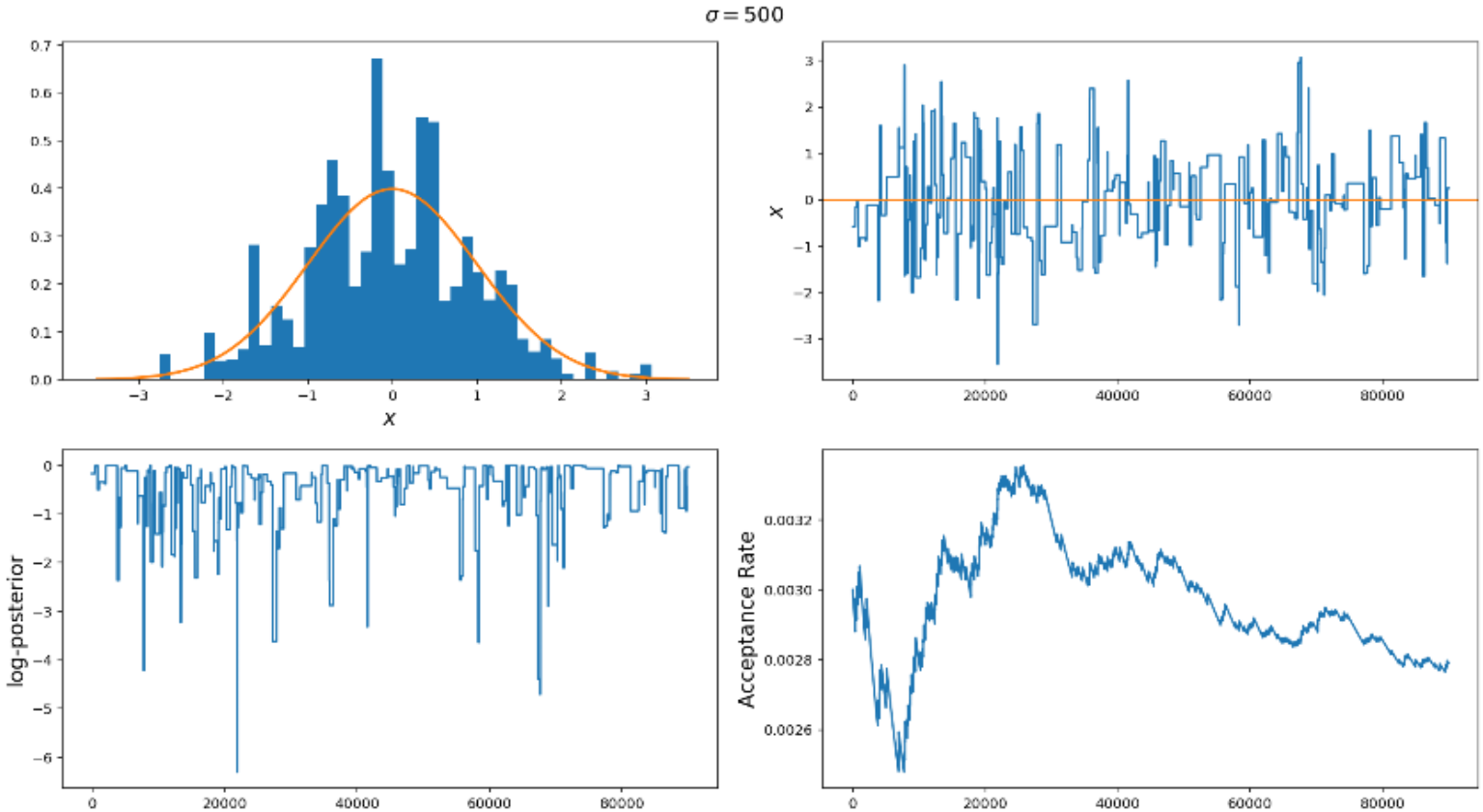
Metropolis Algorithms. V

In an optimally performing MCMC:

- the histogram of samples should converge to the posterior distribution;
- the trace of the chain should sample around the maximum of the posterior such that the samples are close to i.i.d.;
- the log-posterior chain should be smoothly varying around the maximum;
- the acceptance rate depends on the problem but typically for 1-d problems, the acceptance rate should be around 44% (around 23% for more than 5 parameters).

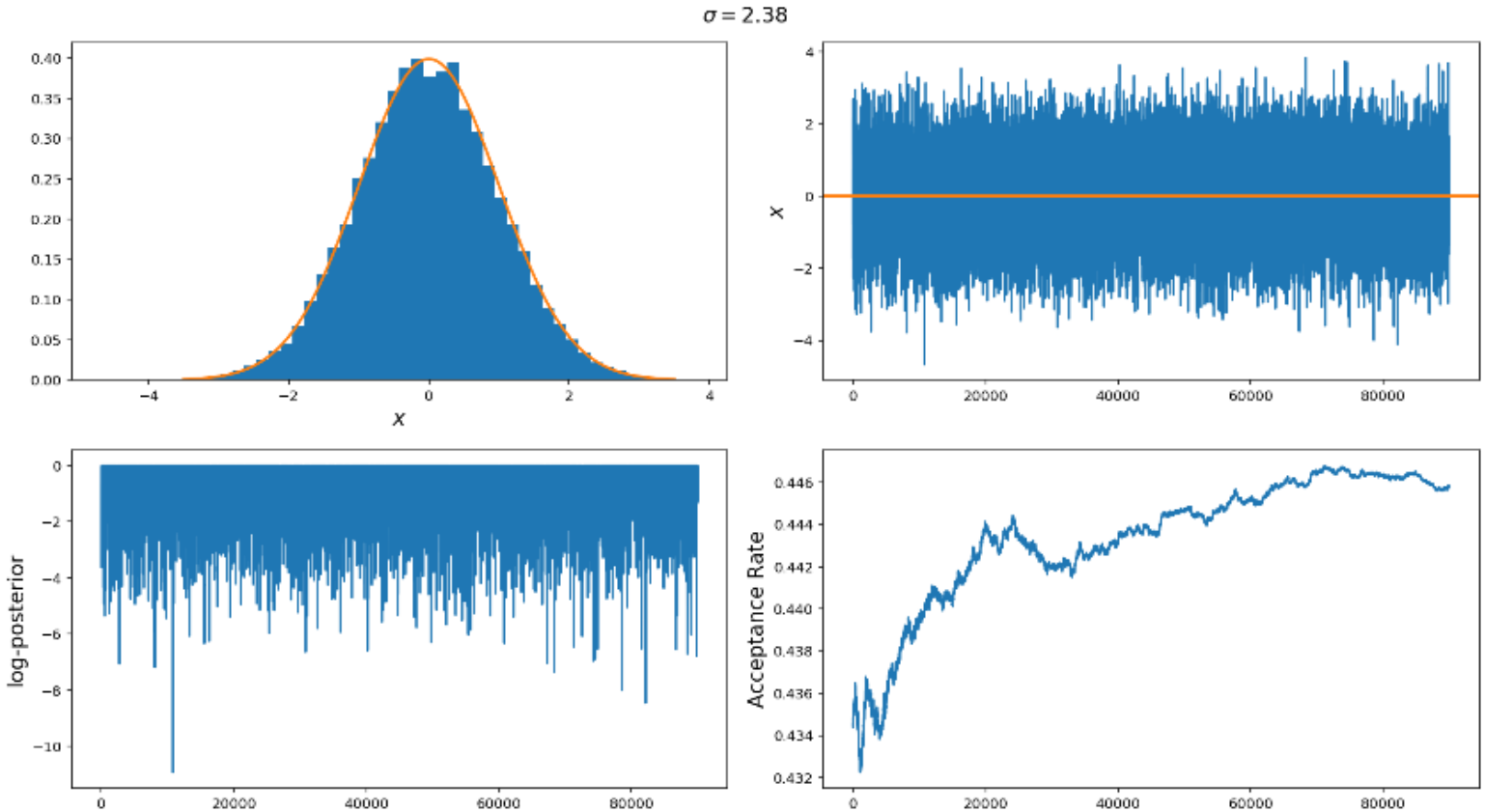
Here: increase the jump proposal size.

Metropolis Algorithms. VI



The jump size here is way too big: better job at recovering the posterior, but choppiness of the trace plots and low acceptance rate: this run is very inefficient. It spends a lot of time at fixed locations in parameter space and rarely moves.
So: decrease the jump proposal size.

Metropolis Algorithms. VII



Good jump proposal size! The samples are near perfect draws from the posterior distribution, the chain trace is nearly i.i.d and the acceptance rate is 44%.