

Point Estimation: definition of estimators



Point estimator: any function $W(X_1, \dots, X_n)$ of a data sample.

The exercise of point estimation is to use particular functions of the data in order to estimate certain unknown population parameters.

Examples: Assume that $X_1, \dots, X_n$ are drawn i.i.d. from some distribution with unknown mean μ and unknown variance σ^2 .

Potential point estimators for μ include: sample mean $\bar{X}_n = \frac{1}{n} \sum_i X_i$; sample median $\text{med}(X_1, \dots, X_n)$.

Potential point estimators for σ^2 include: the sample variance $\frac{1}{n} \sum_i (X_i - \bar{X}_n)^2$.

Any point estimator is a random variable, whose distribution is that induced by the distribution of $X_1, \dots, X_n$.

Example: $X_1, \dots, X_n \sim \text{i.i.d. } N(\mu, \sigma^2)$.

Then sample mean $\bar{X}_n \sim N(\mu_n, \sigma_n^2)$ where $\mu_n = \mu, \forall n$ and $\sigma_n^2 = \sigma^2/n$.

For a particular realization of the random variables $x_1, \dots, x_n$, the corresponding point estimator evaluated at $x_1, \dots, x_n$, i.e., $W(x_1, \dots, x_n)$, is called the *point estimate*.



In these lecture notes, we will consider three types of estimators:

1. Method of moments
2. Maximum likelihood
3. Bayesian estimation



Method of moments:

Assume: $X_1, \dots, X_n \sim \text{i.i.d. } f(x|\theta_1, \dots, \theta_K)$

Here the unknown parameters are $\theta_1, \dots, \theta_K$ ($K \leq N$).

Idea is to find values of the parameters such that the population moments are as close as possible to their “sample analogs”. This involves finding values of the parameters to solve the following K -system of equations:

$$\begin{aligned} m_1 &\equiv \frac{1}{n} \sum_i X_i = EX = \int x f(x|\theta_1, \dots, \theta_K) dx \\ m_2 &\equiv \frac{1}{n} \sum_i X_i^2 = EX^2 = \int x^2 f(x|\theta_1, \dots, \theta_K) dx \\ &\vdots \\ m_K &\equiv \frac{1}{n} \sum_i X_i^K = EX^K = \int x^K f(x|\theta_1, \dots, \theta_K) dx. \end{aligned}$$

Example: $X_1, \dots, X_n \sim \text{i.i.d. } N(\theta, \sigma^2)$. Parameters are θ, σ^2 .

Moment equations are:

$$\begin{aligned} \frac{1}{n} \sum_i X_i &= EX = \theta \\ \frac{1}{n} \sum_i X_i^2 &= EX^2 = VX + (EX)^2 = \sigma^2 + \theta^2. \end{aligned}$$

Hence, the MOM estimators are $\theta^{MOM} = \bar{X}_n$ and $\sigma^{2MOM} = \frac{1}{n} \sum_i X_i^2 - (\bar{X}_n)^2$.

Example: $X_1, \dots, X_n \sim \text{i.i.d. } U[0, \theta]$. Parameter is θ .

MOM: $\bar{X}_n = \frac{\theta}{2} \implies \theta^{MOM} = 2 \cdot \bar{X}_n$.

■■■

Remarks:

- Apart from these special cases above, for general density functions $f(\cdot|\vec{\theta})$, the MOM estimator is often difficult to calculate, because the “population moments” involve difficult integrals. In Pearson’s original paper, the density was a mixture of two normal density functions:

$$f(x|\vec{\theta}) = \lambda \cdot \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) + (1-\lambda) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right)$$

with unknown parameters $\lambda, \mu_1, \mu_2, \sigma_1, \sigma_2$.

- The model assumption that $X_1, \dots, X_n \sim \text{i.i.d. } f(\cdot|\vec{\theta})$ implies a number of moment equations equal to the number of moments, which can be $\gg K$. This leaves room for evaluating the model specification.

For example, in the uniform distribution example above, another moment condition which should be satisfied is that

$$\frac{1}{n} \sum_i X_i^2 = EX^2 = VX + (EX)^2 = \frac{\theta^2}{3} + \frac{\theta}{2}. \quad (1)$$

At the MOM estimator θ^{MOM} , one can see whether

$$\frac{1}{n} \sum_i X_i^2 = \frac{\theta^{MOM2}}{3} + \frac{\theta^{MOM}}{2}.$$

(Later, you will learn how this can be tested more formally.) If this does not hold, then that might be cause for you to conclude that the original specification that $X_1, \dots, X_n \sim \text{i.i.d. } U[0, \theta]$ is inadequate. Eq. (1) is an example is an **overidentifying restriction**.

- While the MOM estimator focuses on using the *sample uncentered moments* to construct estimators, there are other sample quantities which could be useful, such as the sample median (or other sample percentiles), as well as sample minimum or maximum. (Indeed, for the uniform case above, the sample maximum would be a very reasonable estimator for θ .) All these estimators are lumped under the rubric of “generalized method of moments” (GMM).



Maximum Likelihood Estimation

Let $X_1, \dots, X_n \sim \text{i.i.d.}$ with density $f(\cdot|\theta_1, \dots, \theta_K)$.

Define: the likelihood function, for a continuous random variable, is the joint density of the sample observations:

$$L(\vec{\theta}|x_1, \dots, x_n) = \prod_{i=1}^n f(x_i|\vec{\theta}).$$

View $L(\vec{\theta}|\vec{x})$ as a function of the parameters $\vec{\theta}$, for the data observations $\vec{x}$.

From “classical” point of view, the likelihood function $L(\vec{\theta}|\vec{x})$ is a random variable due to the randomness in the data $\vec{x}$. (In the “Bayesian” point of view, which we talk about later, the likelihood function is also random because the parameters $\vec{\theta}$ are also treated as random variables.)

The **maximum likelihood estimator (MLE)** are the parameter values $\vec{\theta}^{ML}$ which maximize the likelihood function:

$$\vec{\theta}^{ML} = \operatorname{argmax}_{\vec{\theta}} L(\vec{\theta}|\vec{x}).$$

Usually, in practice, to avoid numerical overflow problems, maximize the log of the likelihood function:

$$\vec{\theta}^{ML} = \operatorname{argmax}_{\vec{\theta}} \log L(\vec{\theta}|\vec{x}) = \sum_i \log f(x_i|\vec{\theta}).$$

Analogously, for discrete random variables, the likelihood function is the joint probability mass function:

$$L(\vec{\theta}|\vec{x}) = \prod_{i=1}^n P(X = x_i|\vec{\theta}).$$

■■■

Example: $X_1, \dots, X_n \sim \text{i.i.d. } N(\theta, 1)$.

- $\log L(\theta|\vec{x}) = \log(n/\sqrt{2\pi}) - \frac{1}{2} \sum_{i=1}^n (x_i - \theta)^2$
- $\max_{\theta} \log L(\theta|\vec{x}) = \min_{\theta} \frac{1}{2} \sum_i (x_i - \theta)^2$
- FOC: $\frac{\partial \log L}{\partial \theta} = \sum_i (x_i - \theta) = 0 \Rightarrow \theta^{ML} = \frac{1}{n} \sum_i x_i$ (sample mean)

Also should check second order condition: $\frac{\partial^2 \log L}{\partial \theta^2} = -n < 0$: so satisfied.

Example: $X_1, \dots, X_n \sim \text{i.i.d. Bernoulli with prob. } p$. Unknown parameter is p .

- $L(p|\vec{x}) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i}$
-

$$\begin{aligned} \log L(p|\vec{x}) &= \sum_{i=1}^n [x_i \cdot \log p + (1-x_i) \cdot \log(1-p)] \\ &= y \log p + (n-y) \log(1-p) : \quad y \text{ is number of 1's} \end{aligned}$$

- FOC: $\frac{\partial \log L}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p} \implies p^{ML} = \frac{y}{n}$

For $y = 0$ or $y = n$, the p^{ML} is (respectively) 0 and 1: corner solutions.

- SOC: $\left. \frac{\partial \log L}{\partial p^2} \right|_{p=p^{ML}} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2} < 0$ for $0 < y < n$.

When parameter is multidimensional: check that the *Hessian matrix* $\frac{\partial^2 \log L}{\partial \theta \partial \theta'}$ is negative definite.

Example: $X_1, \dots, X_n \sim U[0, \theta]$.

Likelihood function

$$L(\vec{X}|\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \text{if } \max(X_1, \dots, X_n) \leq \theta \\ 0 & \text{if } \max(X_1, \dots, X_n) > \theta \end{cases}$$

which is maximized at $\theta_n^{MLE} = \max(X_1, \dots, X_n)$.

■■■

You can think of ML as a generalized MOM estimator: for $X_1, \dots, X_n$ i.i.d., and K -dimensional parameter vector θ , the MLE solves the FOCs:

$$\begin{aligned} \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_1} &= 0 \\ \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_2} &= 0 \\ &\vdots \\ \frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_K} &= 0. \end{aligned}$$

Under LLN: $\frac{1}{n} \sum_i \frac{\partial \log f(x_i|\theta)}{\partial \theta_k} \xrightarrow{p} E_{\theta_0} \frac{\partial \log f(X|\theta)}{\partial \theta_k}$, for $k = 1, \dots, K$, where the notation E_{θ_0} denote the expectation over the distribution of X at the true parameter vector θ_0 .

Hence, MLE is equivalent to GMM with the moment conditions

$$E_{\theta_0} \frac{\partial \log f(X|\theta)}{\partial \theta_k} = 0, \quad k = 1, \dots, K.$$

■■■

Bayesian inference

Philosophically different view of the world. Model the unknown parameters $\vec{\theta}$ as random variables, and assume that researcher's beliefs about θ are summarized in a *prior distribution* $f(\theta)$.

In this sense, Bayesian approach is “subjective”, because researcher's beliefs about θ are accommodated in inferential approach.

$X_1, \dots, X_n \sim \text{i.i.d.}$ $f(x|\theta)$: the Bayesian views the density of each data observation as a *conditional density*, which is conditional on a realization of the random variable θ .

Given data $X_1, \dots, X_n$, we can update our beliefs about the parameter θ by computing the posterior density (using Bayes Rule):

$$\begin{aligned} f(\theta|\vec{x}) &= \frac{f(\vec{x}|\theta) \cdot f(\theta)}{f(\vec{x})} \\ &= \frac{f(\vec{x}|\theta) \cdot f(\theta)}{\int f(\vec{x}|\theta) f(\theta) d\theta}. \end{aligned}$$

A Bayesian point estimate of θ is some feature of this posterior density. Common point estimators are:

- Posterior mean:

$$E[\theta|\vec{x}] = \int \theta f(\theta|\vec{x}) d\theta.$$

- Posterior median: $F_{\theta|\vec{x}}^{-1}(0.5)$, where $F_{\theta|\vec{x}}$ is CDF corresponding to the posterior density: i.e., $F_{\theta|\vec{x}}(\tilde{\theta}) = \int_{-\infty}^{\tilde{\theta}} f(\theta|\vec{x}) d\theta$.
- Posterior mode: $\max_{\theta} f(\theta|\vec{x})$. This is the point at which the density is highest.

Note that $f(\vec{x}|\theta)$ is just the likelihood function, so that the posterior density $f(\theta|\vec{x})$ can be written as:

$$f(\theta|\vec{x}) = \frac{L(\theta|\vec{x}) \cdot f(\theta)}{\int L(\theta|\vec{x}) f(\theta) d\theta}.$$

But there is a difference in interpretation: in Bayesian world, the likelihood function is random due to both $\vec{x}$ and θ , whereas in classical world, only $\vec{x}$ is random.

Example: $X_1, \dots, X_n \sim \text{i.i.d. } N(\theta, 1)$, with prior density $f(\theta)$.

$$\text{Posterior density } f(\theta|\vec{x}) = \frac{\exp\left(-\frac{1}{2} \sum_i (x_i - \theta)^2\right) f(\theta)}{\int \exp\left(-\frac{1}{2} \sum_i (x_i - \theta)^2\right) f(\theta) d\theta}.$$

Integral in denominator can be difficult to calculate: computational difficulties can hamper computation of posterior densities.

However, note that the denominator is not a function of θ . Thus

$$f(\theta|\vec{x}) \propto L(\theta|\vec{x}).$$

Hence, if we assume that $f(\theta)$ is constant (ie. uniform), for all possible values of θ , then the posterior mode $\text{argmax}_{\theta} f(\theta|\vec{x}) = \text{argmax}_{\theta} L(\theta|\vec{x}) = \theta^{ML}$.

■■■

Example: Bayesian updating for normal distribution, with normal priors

$X \sim N(\theta, \sigma^2)$, assume σ^2 is known.

Prior: $\theta \sim N(\mu, \tau^2)$, assume τ is known.

Then posterior distribution

$$\theta|X \sim N(E(\theta|X), V(\theta|X)),$$

where

$$E(\theta|X) = \frac{\tau^2}{\tau^2 + \sigma^2} X + \frac{\sigma^2}{\sigma^2 + \tau^2} \mu$$

$$V(\theta|X) = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}.$$

This is an example of a *conjugate* prior and conjugate distribution, where the posterior distribution comes from the same family as the prior distribution.

Posterior mean $E(\theta|X)$ is weighted average of X and prior mean μ .

In this case, as $\tau \rightarrow \infty$ (so that prior information gets worse and worse): then $E(\theta|X) \rightarrow X$ (a.s.). These are just the MLE (for just one data observation).

When you observe an i.i.d. sample $\vec{X}_n \equiv (X_1, \dots, X_n)$, with sample mean $\bar{X}_n$:

$$E(\theta|\vec{X}_n) = \frac{n\tau^2}{n\tau^2 + \sigma^2} \bar{X}_n + \frac{\sigma^2}{\sigma^2 + n\tau^2} \mu$$

$$V(\theta|\vec{X}_n) = \frac{\sigma^2 \tau^2}{\sigma^2 + n\tau^2}.$$

In this case, as the number of observations $n \rightarrow \infty$, the posterior mean $E(\theta|\vec{X}_n) \rightarrow \bar{X}_n$. So as $n \rightarrow \infty$, the posterior mean converges to the MLE: when your sample becomes arbitrarily large, you place no weight on your prior information.

■■■

Properties of Bayesian updating typify “rational” behavior (learning) in economics.

Note that, by the law of iterated expectations:

$$E[E(\theta|X_1, X_2)|X_1] = E(\theta|X_1)$$

that is: the sequence of posterior means is a *martingale*. Your best guess of the posterior mean at $t = 2$ is just the posterior mean at $t = 1$.

Equivalently, the “forecast error” has mean zero:

$$E[E(\theta|X_1, X_2) - E(\theta|X_1)|X_1] = 0.$$

This property is being “correct on average” characterizes *rational expectations*.

■■■

Introduction to MCMC simulation

Initially, Bayesian inference was hampered by the difficulties in evaluating the posterior distribution (typically not available in closed form). In recent decades, this has changed due to computational advances, in particular the development of *Markov Chain Monte Carlo* simulation methods. Here I offer a brief (untechnical) introduction. Source: Chib and Greenberg (1995)

Background: First-order Markov chains

- Random sequence $X_1, X_2, \dots, X_n, X_{n+1}, \dots$
- First-order Markov: $P(X_{n+1}|X_n, X_{n-1}, X_{n-2}, \dots) = P(X_{n+1}|X_n) = P(X'|X)$. History-less. Denote this transition distribution as $P(x, dy) = Pr(X' \in dy|X = x)$.
- Invariant distribution: Π is distribution, π is density

$$\Pi(dy) = \int P(x, dy)\pi(x)dx.$$

- Markov chain theory prescribes conditions, under which Markov chain converges to invariant distribution in the following sense: for starting value x , we have

$$\begin{aligned}
p^{(1)}(x, A) &= P(x, A) \\
p^{(2)}(x, A) &= \int_y P^{(1)}(x, dy) P(y, A) \\
p^{(3)}(x, A) &= \int_y P^{(2)}(x, dy) P(y, A) \\
&\dots\dots\dots \\
p^{(n)}(x, A) &= \int P^{(n-1)}(x, dy) P(y, A) \approx \Pi(A)
\end{aligned}$$

That is, for n large enough, each realization of X_n drawn according to $P(x, dy)$ is drawn from the marginal distribution $\Pi(dy)$. (Initial value x does not matter.)

- MCMC simulation goes backwards: given a marginal distribution Π , can we *create* a Markov process with some kernel function, such that Π is the invariant distribution?
- Let $p(x, y)$ denote density function corresponding to kernel function $P(x, dy)$ (ie. $P(x, A) = \int_A p(x, y) dy$). For a given π , if the following relationship is satisfied, then the kernel $P(x, dy)$ achieves π as the invariant distribution:

$$\pi(x)p(x, y) = \pi(y)p(y, x). \quad (2)$$

This is a “reversibility” condition: interpreting $X = x$ and $X' = y$, it (roughly) implies that the probability of transitioning from x to y is the same as transitioning from y to x .

Then π is the invariance density of $P(x, dy)$:

$$\begin{aligned}
\int P(x, A)\pi(x)dx &= \int \int_A p(x, y)dy\pi(x)dx \\
&= \int_A \int p(x, y)\pi(x)dx dy \\
&= \int_A \int p(y, x)\pi(y)dx dy \\
&= \int_A \left[\int p(y, x)dx \right] \pi(y)dy \\
&= \int_A \pi(y)dy = \Pi(A).
\end{aligned}$$

- Can we find such a magic function $p(x, y)$ which satisfies (2)?
- Consider any conditional density function $q(x, y)$. Suppose that

$$\pi(x)q(x, y) > \pi(y)q(y, x)$$

so condition (2) fails. We will “fudge” $q(x, y)$ so that (2) holds:

Metropolis-Hastings approach

- Introduce the “fudge factor” $\alpha(x, y) \leq 1$, such that

$$\pi(x)q(x, y)\alpha(x, y) = \pi(y)q(y, x)\alpha(y, x).$$

When Eq. (2) is violated such that LHS > RHS, you want to set $\alpha(x, y) < 1$ but $\alpha(y, x) = 1$. Vice versa, if LHS < RHS, then you want $\alpha(x, y) = 1$ but $\alpha(y, x) < 1$. We can summarize this as:

$$\alpha(x, y) = \min \left[\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1 \right]. \quad (3)$$

Correspondingly, define the Metropolis-Hastings kernel as:

$$P_{MH}(x, dy) = q(x, y)\alpha(x, y)dy + \underbrace{\left[1 - \int q(x, y)\alpha(x, y)dy \right]}_{r(x)} \mathbb{1}(x \in dy). \quad (4)$$

For a given value x , the Markov chain moves to $y \neq x$ with probability $q(x, y)\alpha(x, y)$, and stays at $y = x$ with probability $[1 - \int q(x, y)\alpha(x, y)dy] = r(x)$.

- Note that this kernel achieves π as the invariant distribution. As above:

$$\begin{aligned} \int P_{MH}(x, A)\pi(x)dx &= \int \int_A q(x, y)\alpha(x, y)dy\pi(x)dx + \int_A r(x)\pi(x)dx \\ &= \int_A \int q(x, y)\alpha(x, y)\pi(x)dx dy + \int_A r(x)\pi(x)dx \\ &= \int_A \int q(y, x)\alpha(y, x)\pi(y)dx dy + \int_A r(x)\pi(x)dx \\ &= \int_A \left[\int q(y, x)\alpha(y, x)dx \right] \pi(y)dy + \int_A r(x)\pi(x)dx \\ &= \int_A [1 - r(y)] \pi(y)dy + \int_A r(x)\pi(x)dx \\ &= \int_A \pi(y)dy = \Pi(A). \end{aligned}$$

- This MH kernel can be implemented via simulation: start with x^0 , then
 - Draw “candidate” y_1 from the conditional density $q(x^0, y_1)$.
 - Set $x^1 = \begin{cases} y_1 & \text{with probability } \alpha(x^0, y_1) \\ x^0 & \text{with probability } 1 - \alpha(x^0, y_1) \end{cases}$
 - Now draw candidate y_2 from the conditional density $q(x^1, y_2)$.
 - Set $x^2 = \begin{cases} y_2 & \text{with probability } \alpha(x^1, y_2) \\ x^1 & \text{with probability } 1 - \alpha(x^1, y_2) \end{cases}$
 - And so on.

According to the Markov chain theory, for N large enough we have approximately:

$$\Pi(A) \approx \frac{1}{N} \sum_{i=\tau+1}^{\tau+N} \mathbb{1}(x^i \in A).$$

Here τ refers to the length of an initial “burn-in” period, when the Markov chain is still converging.

- What is nice here is that this works regardless of the $q(\theta, \theta')$ that you choose. Two common options are:
 - Random walk: $q(x, y) = q_1(y - x)$, where q_1 is a density function, eg. $N(0, \sigma^2)$. According to this approach, the candidate $y = x + \epsilon$, $\epsilon \sim N(0, \sigma^2)$, and so it is called a random walk density.
 - Independent draws: $q(x, y) = q_2(y)$, where q_2 is a density function. The candidate y is independent of x . (However, because the candidate is sometimes rejected, with probability $1 - \alpha(x, y)$, the resulting random sequence still exhibits dependence.)

Application to Bayesian posterior inference

- For Bayesian inference, the desired density is the posterior density $f(\theta|\vec{z})$ (where $\vec{z}$ denotes the data information). We want to construct a Markov-chain, using the MH idea, such that its invariant distribution is the posterior distribution of $\theta|\vec{z}$.
- We use θ and θ' to denote the “current” and “next” draws of θ from the Markov chain.

- Recall that $f(\theta|\vec{z}) \propto f(\vec{z}|\theta)f(\theta)$, the likelihood function times the prior density. Hence, the MH fudge factor is

$$\begin{aligned}\alpha(\theta, \theta') &= \min \left[\frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}, 1 \right] \\ &= \min \left[\frac{f(\vec{z}|\theta')f(\theta')q(\theta', \theta)}{f(\vec{z}|\theta)f(\theta)q(\theta, \theta')}, 1 \right].\end{aligned}$$

So you are more likely to accept a draw θ' relative to θ if the likelihood ratio is higher.

- For a given $q(x, y)$, drawing a sequence $\theta^1, \theta^2, \dots, \theta^n, \dots$ such that, for n large enough, each θ^n can be treated as a draw with marginal distribution equal to $\pi(\theta|\vec{z})$. Hence, for instance, the posterior mean can be approximated as:

$$\mathbb{E}[\theta|\vec{z}] \approx \frac{1}{N} \sum_{i=\tau}^{\tau+N} \theta^i.$$

- From the perspective of maximizing the likelihood function, the MCMC chain can be viewed as a *stochastic optimization* algorithm where one can move to candidate parameters which decrease the likelihood function. (In contrast, most optimization algorithms are hill-climbing procedures which always move to parameter values which successively increase the optimization criterion.) Similar to “simulated annealing”; more effective at finding global optima.



Example: “Data augmentation”

Another computational simplification available in Bayesian contexts is to *treat latent variables in models as parameters* and do joint inference on model parameters and latent variables. In many problems this leads to vast simplification in doing inference on (just) the model parameters. This is called data augmentation.

The important philosophical distinction of the Bayesian approach is that data and model parameters are treated on an “equal footing”. Hence, just as we make posterior inference on model parameters, we can also make posterior inference on unobserved variables in “latent variable” models, which are models where not all the model variables are observed.

Consider a simple example (the “binary probit” model):

$$\begin{aligned} z &= \beta x + \epsilon, \quad \epsilon \sim N(0, 1) \\ y &= \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{if } z \geq 0. \end{cases} \end{aligned} \quad (5)$$

The researcher observes (x, y) , but not (z, β) . He wishes to form the posterior of $z, \beta | x, y$.

We do all inference conditional on x . Therefore the relevant prior is

$$\begin{aligned} (z, \beta | x) &= (z | \beta, x) \cdot \beta | x \\ &= N(\beta x, 1) \cdot \underbrace{N(\bar{\beta}, a^2)}_{\equiv f(\beta)} \\ &= \phi(z - \beta x) \cdot \frac{1}{a} \phi\left(\frac{\beta - \bar{\beta}}{a}\right). \end{aligned} \quad (6)$$

In the above, we assume the marginal prior on β is normal (and doesn’t depend on x). The conditional prior density of $z | \beta, x$ is derived from the model specification (5). ϕ denotes density for standard normal.

The posterior is:

$$\begin{aligned} (z, \beta | y, x) &\propto “L(y | z, \beta, y, x)” \cdot (z, \beta | x) \\ &= (\mathbb{1}(z \geq 0) * y + \mathbb{1}(z < 0) * (1 - y)) \cdot f_0(z, \beta | x). \end{aligned} \quad (7)$$

Note how doing this simplifies the “likelihood”. This is not the likelihood you would use when you do MLE (because you do not observe z !) but this is what you use if you treat the latent z as a parameter. With very complicated likelihoods involving latent variables this approach can computationally simplify the inference on the model parameters β .

Accordingly, this can be marginalized over z to obtain the posterior of the model parameters $\beta | y, x$.

References

CHIB, S., AND E. GREENBERG (1995): “Understanding the Metropolis-Hastings Algorithm,” *American Statistician*, 49, 327–335.