Linear Model

Problem and General Approach

Suppose that we are estimating a linear model of the form

$$y = X\beta + \varepsilon, \tag{1}$$

where y is a $N \times 1$ vector of observed outputs, X is a $N \times k$ matrix of observed inputs, β is a $k \times 1$ vector of slope parameters, and $\varepsilon \sim N(0, \sigma^2 I)$ is a $N \times 1$ vector of error terms. Samples of parameters $\beta|y, X$ and $\sigma|y, X$ can be obtained through application of an MCMC algorithm, given priors on parameters β and σ. Also suppose that X follows some distribution $p(X|\theta_X)$, where θ_X are parameters of that distribution.

Given a scalar output value $\tilde{y}_i$, our goal is to estimate the distribution of inputs that could have resulted in $\tilde{y}_i$, given by $p(\tilde{X}_i|\tilde{y}_i, y, X, \theta_X)$, where $\tilde{y}_i = \tilde{X}_i\beta + \tilde{\varepsilon}_i$. We'll start by considering $p(\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta, \sigma)$. Notice that

$$\begin{aligned} p(\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta, \sigma) &\propto p(\tilde{y}_i|\tilde{X}_i, y, X, \theta_X, \beta, \sigma)p(\tilde{X}_i|y, X, \theta_X, \beta, \sigma) \\ &\equiv p(\tilde{y}_i|\tilde{X}_i, \beta, \sigma)p(\tilde{X}_i|\theta_X). \end{aligned} \tag{2}$$

This relationship immediately reveals a general sampling approach for $\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta, \sigma$: $p(\tilde{X}_i|\theta_X)$ acts as the prior and $p(\tilde{y}_i|\tilde{X}_i, \beta, \sigma)$ acts as the likelihood, resulting in the posterior $p(\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta, \sigma)$. Samples of the posterior can be generated by the application of any appropriate MCMC algorithm, such as Metropolis-Hastings (MH) or Hamiltonian Monte Carlo (HMC).

In practice, of course, parameters β and σ are not known. To address this issue, we can simply integrate those parameters out:

$$p(\tilde{X}_i|\tilde{y}_i, y, X, \theta_X) = \int p(\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta, \sigma)p(\beta|y, X)p(\sigma|y, X)d\beta d\sigma \tag{3}$$

$$\propto \int p(\tilde{y}_i|\tilde{X}_i, \beta, \sigma)p(\tilde{X}_i|\theta_X)p(\beta|y, X)p(\sigma|y, X)d\beta d\sigma. \tag{4}$$

This reveals a general sampling approach for $\tilde{X}_i|\tilde{y}_i, y, X, \theta_X$:

1. Obtain samples of $\beta|y, X$ and $\sigma|y, X$ through the application of an appropriate MCMC algorithm to estimate Equation (1). This results in samples $\beta^{[s]}|y, X$ and $\sigma^{[s]}|y, X$, where $s = 1, ..., S$.

2. For each s, generate samples of $\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta^{[s]}, \sigma^{[s]}$ using Equation (7).

3. Combining samples of $\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta^{[s]}, \sigma^{[s]}$ across $s = 1, ..., S$ results in samples of $\tilde{X}_i|\tilde{y}_i, y, X, \theta_X$.

The final challenge is to accommodate various distributions for $\tilde{X}_i$. We'll start by illustrating the simple case where $\tilde{X}_i$ follows a multivariate normal distribution. We will then relax that assumption to allow $\tilde{X}_i$ to follow a mixture of multivariate normals, which provides adequate approximation to a wide range of empirical distributions and datasets.

Example: $\tilde{X}_i \sim N(\mu_X, \Sigma_X)$

We will begin with an example where $\tilde{X}_i$ follows a multivariate normal distribution. For purposes of estimation, we will separate the constant term from variable inputs, resulting in the linear model

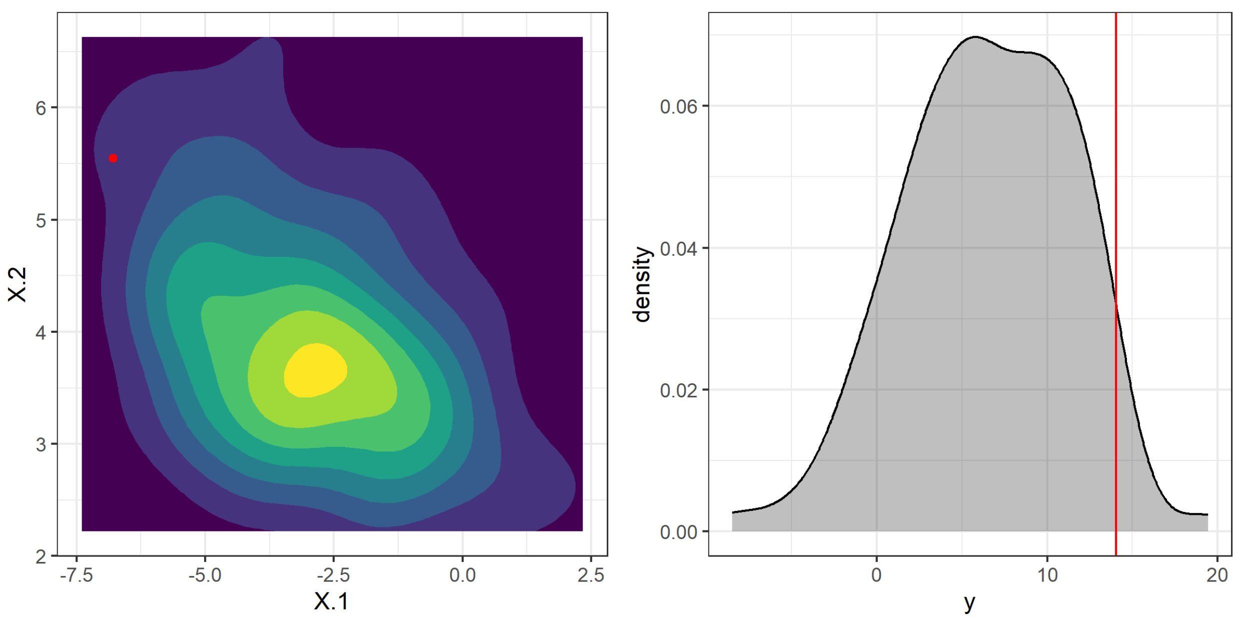
$$y_i = \alpha + X_i\beta + \varepsilon_i, \tag{5}$$

where α is a scalar parameter, β is a $k \times 1$ parameter vector, and $\varepsilon_i \sim$ iid $N(0, \sigma^2)$. Notice that there are $k + 2$ parameters in this model formulation.

For this example, we use population and parameter values shown in Table 1. Values of X_i were generated

Parameter	Value
N	100
k	2
α	4
β	$\begin{bmatrix} -2.5 \\ -1.25 \end{bmatrix}$
σ	0.4
μ_X	$\begin{bmatrix} -3 \\ 4 \end{bmatrix}$
Σ_X	$\begin{bmatrix} 4.5 & -0.75 \\ -0.75 & 1.25 \end{bmatrix}$

Table 1: Parameter values



(a) Density of X (b) Density of y

Figure 1: Densities of observed data

Parameter	Prior
α	$N(0, 100)$
β	$N(0, 100I)$
σ	Cauchy$(0, 1)$

Table 2: Parameter priors

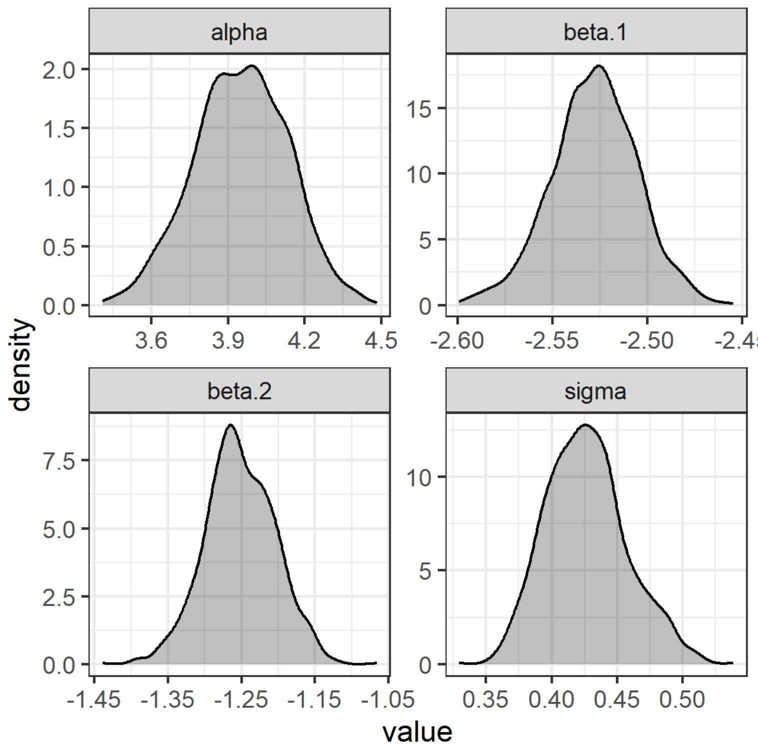


Figure 2: Parameter posterior densities

using $X_i \sim$ iid $N(\mu_X, \Sigma_X)$, values of ε_i were generated using $\varepsilon_i \sim$ iid $N(0, \sigma^2)$, and values of y were generated using $y = X\beta + \varepsilon$. The bivariate density of X is shown in Figure 1a and the density of y is shown in Figure 1b; note that particular values of X and y are singled out in red, which we'll return to later.

Samples of parameters $\alpha, \beta, \sigma | y, X$ were generated using No U-Turn Sampling (NUTS) as implemented in Stan, using priors shown in Table 2. The sampling procedure used a single chain with 1000 warmup iterations and 1000 sampling iterations, and standard convergence diagnostics were applied. Posterior densities are shown in Figure 2.

Next, a particular observation of X and y was singled out to validate the inverse propagation procedure. Specifically, we used an observation of $\tilde{X}_i \approx \begin{bmatrix} -6.79 & 5.55 \end{bmatrix}'$ and $\tilde{y}_i = \alpha + \tilde{X}_i \beta \approx 14.04$, shown in red in Figure 1. The sampling procedure laid out in the General Approach section was used to generate samples of $\tilde{X}_i | \tilde{y}_i, y, X, \theta_X$ using NUTS, and standard convergence checks were again applied. The resulting posterior distribution of $\tilde{X}_i$ and actual value of X are shown in Figure 3.

To analyze this fit, we've recreated Figure 3 with two additional lines in Figure 4. The orange line represents all values of X_i such that the equation $\tilde{y}_i = X_i \beta$ is satisfied. Since $\tilde{y}_i$ was generated without a random error component, the true value of $\tilde{X}_i$ is necessarily on this line. Values of X_i not on the orange line must have some non-zero error component in order for $\tilde{y}_i = X_i \beta + \varepsilon_i$ to be satisfied. The blue line shows the first principal component of the posterior sample of $\tilde{X}_i$.

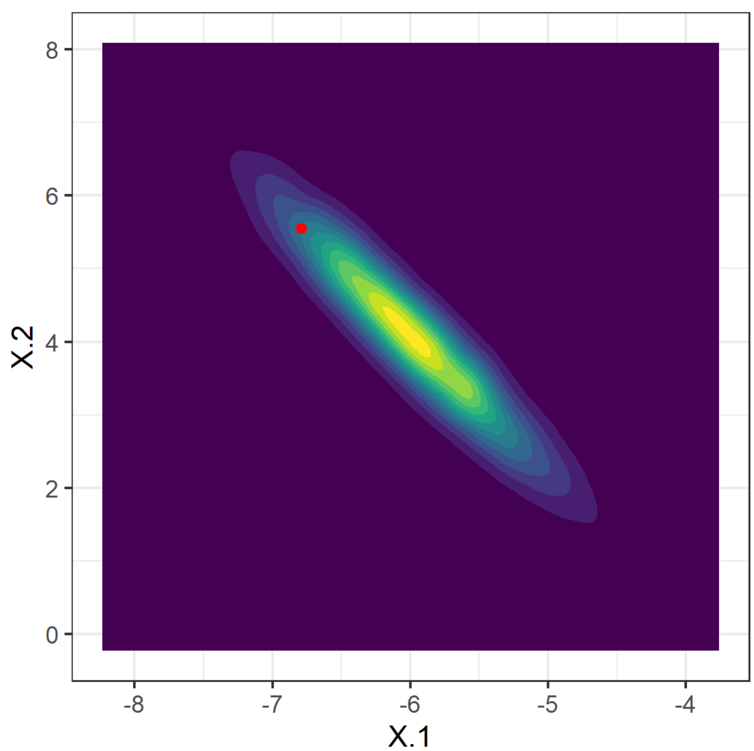


Figure 3: Posterior density of $\tilde{X}_i$

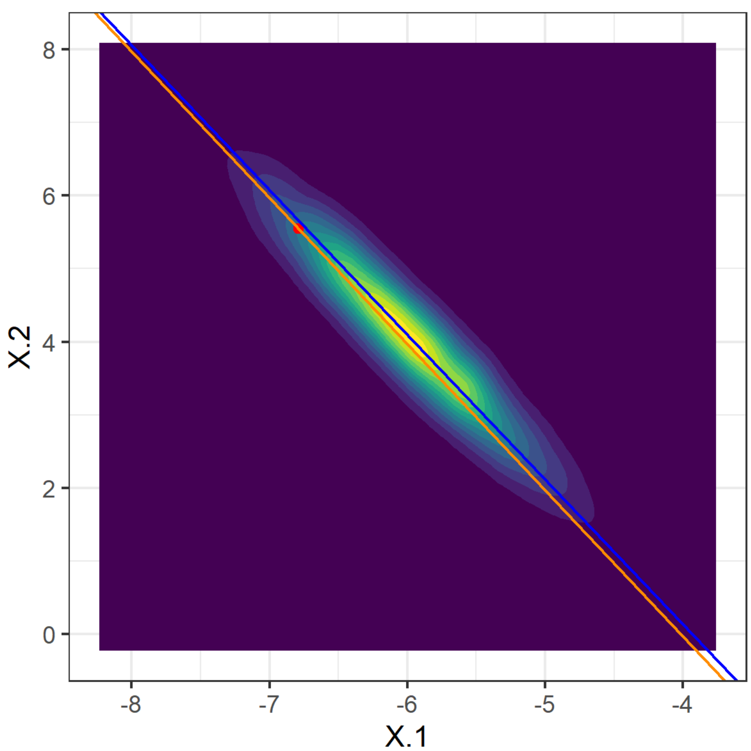


Figure 4: Posterior density of $\tilde{X}_i$ with solution line and PCA

Example: $\tilde{X}_i$ Follows a Mixture of Normals

In this example, we make all the same assumptions as the previous example, except we generalize to the case where $\tilde{X}_i$ follows a mixture of normal distributions. That is, there are mixing probabilities $m_1, ..., m_C$ with $\sum_c m_c = 1$ such that

$$p(X|\theta_X) = \sum_c m_c p(X|\mu_c, \Sigma_c), \tag{6}$$

where $p(X|\mu_c, \Sigma_c)$ is a multivariate normal distribution with known mean μ_c and variance Σ_c. As before, we can note that

$$\begin{aligned} p(\tilde{X}_i|\tilde{y}_i, y, X, \theta_X, \beta, \sigma) &\propto p(\tilde{y}_i|\tilde{X}_i, \beta, \sigma) p(\tilde{X}_i|\theta_X) \\ &\equiv p(\tilde{y}_i|\tilde{X}_i, \beta, \sigma) \sum_c m_c p(\tilde{X}_i|\mu_c, \Sigma_c) \end{aligned} \tag{7}$$