Homework I

Bernardo Flores López

1 Bayesian inference in simple conjugate families

1. Suppose that we take independent observations $x_1, \dots, x_N$ from a Bernoulli sampling model with unknown probability w. That is, the x_i are the results of flipping a coin with unknown bias. Suppose that w is given a Beta(a,b) prior distribution:

$$p(w) = \frac{\Gamma(a+b)}{\Gamma(a) \cdot \Gamma(b)}\, w^{a-1}(1-w)^{b-1}\,,$$

where $\Gamma(\cdot)$ denotes the Gamma function. Derive the posterior distribution $p(w \mid x_1, \dots, x_N)$.

The posterior is proportional to

$$\begin{aligned} p(w|x_1,\dots,x_N) &\propto p(w)\,p(x_1,\dots,x_N|w) \\ &\propto w^{a-1}(1-w)^{b-1} w^{\sum_{i=1}^N x_i}(1-w)^{N-\sum_{i=1}^N x_i} \\ &= w^{a+\sum_{i=1}^N x_i - 1}(1-w)^{b+N-\sum_{i=1}^N x_i - 1} \end{aligned}$$

We recognize the kernel of a $\text{Beta}\Big(a+\sum_{i=1}^N x_i, b+N-\sum_{i=1}^N x_i\Big)$, hence the posterior distribution has that law.

2. The probability density function (PDF) of a gamma random variable, $x \sim \text{Ga}(a,b)$, is

$$p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\}\,.$$

Suppose that $x_1 \sim \text{Ga}(a_1, 1)$ and that $x_2 \sim \text{Ga}(a_2, 1)$. Define two new random variables $y_1 = x_1/(x_1+x_2)$ and $y_2 = x_1 + x_2$. Find the joint density for (y_1, y_2) using a direct PDF transformation (and its Jacobian). Use this to characterize the marginals $p(y_1)$ and $p(y_2)$, and propose a method that exploits this result to simulate beta random variables, assuming you have a source of gamma random variables.

First let us obtain the Jacobian. Let $f_1(x_1,x_2) = x_1/(x_1+x_2)$ and $f_2(x_1,x_2) = x_1+x_2$, then $f_1^{-1}(y_1,y_2) = y_1y_2$ and $f_2^{-1}(y_1,y_2) = y_1+y_2$. Thus

$$J = \begin{bmatrix} \partial_{y_1} f_1^{-1} & \partial_{y_2} f_1^{-1} \\ \partial_{y_1} f_2^{-1} & \partial_{y_2} f_2^{-1} \end{bmatrix} = \begin{bmatrix} y_2 & y_1 \\ -y_2 & 1-y_1 \end{bmatrix},$$

so

$$|J| = y_2$$

Assuming independence, we can then obtain the joint density

$$\begin{aligned} f_{y_1,y_2}(x,y) &= f\,[y_1y_2, y_2(1-y_1)]\ |J| \\ &= f_1(y_1y_2) f_2(y_2(1-y_1)) \left(\frac{x_1}{x_1+x_2}\right)^{a_1} \\ &\propto e^{-y_1y_2}(y_1y_2)^{a_1-1} e^{-y_2+y_2y_1}(y_2-y_2y_1)^{a_2-1} y_2 \\ &= \underbrace{y_1^{a_1-1}(1-y_1)^{a_2-1}}_{\text{Kernel of Beta}(a_1,a_2)} \quad \underbrace{y_2^{a_1+a_2-1} e^{-y_2}}_{\text{Kernel of Gamma}(a_1+a_2,1)} \end{aligned}$$

Hence the joint distribution is the product measure $\text{Beta}(a_1,a_2) \otimes \text{Gamma}(a_1+a_2, 1)$.

3. Suppose that we take independent observations $x_1, \dots, x_N$ from a normal sampling model with unknown mean θ and *known* variance σ^2: $x_i \sim \text{N}(\theta, \sigma^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta \mid x_1, \dots, x_N)$.

Again, the posterior is proportional to

$$\begin{aligned}
p(\theta|x_1,\ldots,x_N) &\propto p(\theta)\,p(x_1,\ldots,x_N|\theta)\\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\theta)^2\right\}\exp\left\{-\frac{1}{2v}\,(\theta-m)^2\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^{N}x_i^2-\frac{2}{\sigma^2}N\theta\bar{x}+\frac{N}{\sigma^2}\theta^2+\frac{1}{v}\theta^2-\frac{2}{v}\theta m+\frac{1}{v}m^2\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{N\nu+\sigma^2}{v\sigma^2}\theta^2-2\frac{Nv\bar{x}+m\sigma^2}{v\sigma^2}\theta\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{N\nu+\sigma^2}{v\sigma^2}\theta^2-2\frac{Nv\bar{x}+m\sigma^2}{v\sigma^2}\theta+\left[\frac{\frac{1}{v}m+\frac{n}{\sigma^2}\bar{x}}{\frac{1}{v}+\frac{n}{\sigma^2}}\right]^2\right)\right\}\\
&\propto \exp\left\{-\frac{1}{2\frac{N\nu+\sigma^2}{v\sigma^2}}\left(\theta-\frac{\frac{1}{v}m+\frac{n}{\sigma^2}\bar{x}}{\frac{1}{v}+\frac{n}{\sigma^2}}\right)^2\right\}
\end{aligned}$$

We recognize the kernel of a $\mathrm{N}\left(\mu_n,\sigma_n^2\right)$, where

$$\mu_n=\frac{\frac{1}{v}m+\frac{n}{\sigma^2}\bar{x}}{\frac{1}{v}+\frac{n}{\sigma^2}}\quad\text{and}\quad\sigma_n^2=\frac{N}{\sigma^2}+\frac{1}{v},$$

making this the posterior distribution of the mean.

4. Suppose that we take independent observations $x_1,\ldots,x_N$ from a normal sampling model with *known* mean θ but *unknown* variance σ^2. (This seems even more artificial than the last, but is conceptually important.) To make this easier, we will re-express things in terms of the precision, or inverse variance $\omega=1/\sigma^2$:

$$p(x_i\mid\theta,\omega)=\left(\frac{\omega}{2\pi}\right)^{1/2}\exp\left\{-\frac{\omega}{2}(x_i-\theta)^2\right\}.$$

Suppose that ω has a gamma prior with parameters a and b, implying that σ^2 has what is called an inverse-gamma prior.

Proceeding as before,

$$\begin{aligned}
p(\omega|x_1,\ldots,x_N) &\propto p(\omega)\,p(x_1,\ldots,x_N|\omega)\\
&\propto \omega^{a-\frac{1}{2}}\exp\{-b\omega\}\,\omega^{N/2}\exp\left\{-\frac{\omega}{2}\sum_{i=1}^{N}(x_i-\theta)^2\right\}\\
&\propto \omega^{a+\frac{N}{2}-1}\exp\left\{-\left(b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2\right)w\right\}.
\end{aligned}$$

We recognize the kernel of a Gamma(a_n, b_n), where

$$a_n=a+\frac{N}{2}\quad\text{and}\quad b_n=b+\frac{1}{2}\sum_{i=1}^{N}(x_i-\theta)^2,$$

making this the posterior distribution for the variance.

The posterior for the precision is then Inverse Gamma with the same parameters.

5. Suppose that, as above, we take independent observations $x_1,\ldots,x_N$ from a normal sampling model with unknown, common mean θ. This time, however, each observation has its own idiosyncratic (but known) variance: $x_i\sim\mathrm{N}(\theta,\sigma_i^2)$. Suppose that θ is given a normal prior distribution with mean m and variance v. Derive the posterior distribution $p(\theta\mid x_1,\ldots,x_N)$. Express the posterior mean in a form that is clearly interpretable as a weighted average of the observations and the prior mean.

The derivation is essentially the same as the one with the common variance. Let

$$\bar{x}_\sigma=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}x_i$$

$$\begin{aligned}
p(\theta|x_1,\ldots,x_N) &\propto p(\theta)\,p(x_1,\ldots,x_N|\theta) \\
&\propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}\left(x_i-\theta\right)^2\right\}\exp\left\{-\frac{1}{2v}\left(\theta-m\right)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{N}\sigma_i^{-2}x_i^2-2N\theta\bar{x}_\sigma+\theta^2\sum_{i=1}^{N}\sigma_i^{-2}+\frac{1}{v}\theta^2-\frac{2}{v}\theta m+\frac{1}{v}m^2\right)\right\} \\
then \qquad &\propto \exp\left\{-\frac{1}{2}\left(\frac{v\sum\sigma_i^{-2}+1}{v}\theta^2-2\frac{Nv\bar{x}_\sigma+m}{v}\theta\right)\right\} \qquad Fromthiskernelwecantelltheposteriordist \\
&\propto \exp\left\{-\frac{1}{2\frac{1}{\sum\sigma_i^{-2}+\frac{1}{v}}}\left(\theta^2-2\frac{Nv\bar{x}_\sigma+m}{v\sum\sigma_i^{-2}+1}\theta+\left[\frac{N\bar{x}_\sigma+\frac{m}{v}}{\sum\sigma_i^{-2}+\frac{1}{v}}\right]^2\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2\frac{1}{\sum\sigma_i^{-2}+\frac{1}{v}}}\left(\theta-\frac{N\bar{x}_\sigma+\frac{m}{v}}{\sum\sigma_i^{-2}+\frac{1}{v}}\right)^2\right\}.
\end{aligned}$$

$\mathrm{N}\left(\mu_n,\sigma_n^2\right)$ with parameters

$$\mu_n=\frac{N\bar{x}_\sigma+\frac{m}{v}}{\sum\sigma_i^{-2}+\frac{1}{v}} \quad \text{and} \quad \sigma_n^2=\frac{1}{\sum\sigma_i^{-2}+\frac{1}{v}}$$

Note how $\bar{x}_\sigma$ is just the mean of the observations each weighted by its variance, so the posterior mean is a weighted average of both the prior mean and the weighted sample mean of the observations.

6. Suppose that $(x \mid \omega) \sim \mathrm{N}\left(m, \omega^{-1}\right)$, and that ω has a Gamma$(a/2, b/2)$ prior, with PDF defined as above. Show that the marginal distribution of x is Student's t with d degrees of freedom, center m, and scale parameter $(b/a)^{1/2}$. This is why the t distribution is often referred to as a *scale mixture of normals*.

$$\begin{aligned}
f_X(x) &= \int f_{X|\omega}(x)f_\omega(\omega)\,\mathrm{d}\omega \\
&\propto \int f_{\omega|X}(\omega)\,\mathrm{d}\omega \\
&\propto \int \underbrace{\omega^{a+\frac{1}{2}-1}\exp\left\{-\left(b+\frac{1}{2}(x-m)^2\right)\omega\right\}}_{\text{Kernel of a Gamma distribution}}\,\mathrm{d}\omega \\
&= \frac{\left(\frac{b+(x-m)^2}{2}\right)^{\frac{a+1}{2}}}{\Gamma\left(\frac{a+1}{2}\right)},
\end{aligned}$$

which is proportional to the density of a Student-t with the required parameters.

1.1 The multivariate normal distribution

1. First, some simple moment identities. The covariance matrix cov(x) of a vector-valued random variable x is defined as the matrix whose (i,j) entry is the covariance between x_i and x_j. In matrix notation, $\text{cov}(x) = E\{(x-\mu)(x-\mu)^T\}$, where μ is the mean vector whose ith component is $E(x_i)$. Prove the following: (1) $\text{cov}(x) = E(xx^T) - \mu\mu^T$; and (2) $\text{cov}(Ax+b) = A\text{cov}(x)A^T$ for matrix A and vector b.

   Proof. Let us begin with the first identity.

$$\begin{aligned}
\text{Cov}(x) &= \mathbb{E}\left[(x-\mu)(x-\mu)'\right] \\
&= \mathbb{E}\left[xx'-x\mu'-\mu x'+\mu\mu'\right] \\
&= \mathbb{E}\left[xx'\right]-\mathbb{E}\left[x\right]\mu'-\mu\mathbb{E}\left[x'\right]+\mu\mu+ \\
&= \mathbb{E}\left[xx'\right]-\mu\mu'-\mu\mu'+\mu\mu' \\
&= \mathbb{E}\left[xx'\right]-\mu\mu'.
\end{aligned}$$

Now for the second one

$$\begin{aligned}
\mathrm{Cov}(Ax+b) &= \mathbb{E}\left[(Ax+b-\mathbb{E}[Ax+b])(Ax+b-\mathbb{E}[Ax+b])'\right] \\
&= \mathbb{E}\left[(Ax+b-A\mathbb{E}[x]+b)(Ax+b-A\mathbb{E}[x]+b)'\right] \\
&= \mathbb{E}\left[(Ax-A\mathbb{E}[x])(Ax-A\mathbb{E}[x])'\right] \\
&= \mathbb{E}\left[A(x-\mathbb{E}[x])(x-\mathbb{E}[x])'A'\right] \\
&= A\mathbb{E}\left[(x-\mathbb{E}[x])(x-\mathbb{E}[x])'\right]A' \\
&= A\mathrm{Cov}(x)A'.
\end{aligned}$$

□

2. Consider the random vector $z = (z_1, \ldots, z_p)^T$, with each entry having an independent standard normal distribution (that is, mean 0 and variance 1). Derive the probability density function (PDF) and moment-generating function (MGF) of z, expressed in vector notation.[1] We say that z has a standard multivariate normal distribution.

   Note that since the x_i are all independent then the joint density factorizes, giving

$$f_{x_1,\ldots,x_p}(y_1,\ldots,y_p) = \prod_{i=1}^{p} f_{x_i}(y_i) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}y_i^2\right\} = (2\pi)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} y_i^2\right\} = (2\pi)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}y'y\right\}$$

   Now, using the independence and the fact that the exponential function is continuous and thus (Borel) measurable,

$$\begin{aligned}
\mathbb{E}\left[\exp\left\{t'x\right\}\right] &= \mathbb{E}\left[\exp\left\{\sum_{i=1}^{p} t_i x_i\right\}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{p} \exp\{t_i x_i\}\right] \\
&= \prod_{i=1}^{p} \mathbb{E}[\exp\{t_i x_i\}] \\
&= \prod_{i=1}^{p} \exp\left\{\frac{t_i^2}{2}\right\} \\
&= \exp\left\{\frac{1}{2}t't\right\}.
\end{aligned}$$

3. A vector-valued random variable $x = (x_1, \ldots, x_p)^T$ has a *multivariate normal distribution* if and only if every linear combination of its components is univariate normal. That is, for all vectors a not identically zero, the scalar quantity $z = a^T x$ is normally distributed. From this definition, prove that x is multivariate normal, written $x \sim \mathrm{N}(\mu, \Sigma)$, if and only if its moment-generating function is of the form $E(\exp\left\{t^T x\right\}) = \exp\left\{t^T\mu + t^T\Sigma t/2\right\}$. Hint: what are the mean, variance, and moment-generating function of z, expressed in terms of moments of x?

   Proof. First note that if x is multivariate normal

$$\mathbb{E}\left[a'x\right] = a'\mathbb{E}[x] = a'\mu \quad \text{and} \quad \mathrm{Var}\left(a'x\right) = a'\mathrm{Var}(x)\,a = a'\Sigma a.$$

   Now suppose that x is multivariate normal. Then for any t $t'x \sim \mathrm{N}(t'\mu, t'\Sigma t)$ the mgf of it is exactly

$$\mathbb{E}\left[\exp\left\{t'x\right\}\right] = \exp\left\{t'\mu + t'\Sigma t/2\right\},$$

   which proves that direction.

   Conversely, assume that the mgf of x has the form $\exp\left\{t^T\mu + t^T\Sigma t/2\right\}$. Then for any vector a

$$\mathbb{E}\left[\exp\left\{t'(a'x)\right\}\right] = \mathbb{E}\left[\exp\{(at\}'x\right] = \exp\left\{(at)'\mu + (at)'\Sigma(at)/2\right\} = \exp\left\{t'(a'\mu) + t'(a'\Sigma a)t)/2\right\},$$

   hence by uniqueness of the Laplace transform $a'x$ is normally distributed, finishing the proof. □

[1] Remember that the MGF of a vector-valued random variable x is the expected value of the quantity $\exp\{()\, t^T x)$, as a function of the vector argument t.

4. Another basic theorem is that a random vector is multivariate normal if and only if it is an affine transformation of independent univariate normals. You will first prove the "if" statement. Let z have a standard multivariate normal distribution, and define the random vector $x = Lz + \mu$ for some $p \times p$ matrix L of full column rank.[2] Prove that x is multivariate normal. In addition, use the moment identities you proved above to compute the expected value and covariance matrix of x.

   Proof. Let us obtain its moment generating function.

$$\begin{aligned}\mathbb{E}\left[\exp\left\{t'x\right\}\right] &= \mathbb{E}\left[\exp\left\{t'(Lz+\mu)\right\}\right] \\ &= \mathbb{E}\left[\exp\left\{(t'L)z\right\}\right]\exp\left\{t'\mu\right\} \\ &= \exp\left\{\frac{1}{2}t'LL't + t'\mu\right\}.\end{aligned}$$

   This implies that $x \sim \text{NMv}(\mu, LL')$. □

5. Now for the "only if." Suppose that x has a multivariate normal distribution. Prove that x can be written as an affine transformation of standard normal random variables. (Note: a good way to prove that something can be done is to do it! Think about a matrix A such that $AA^T = \Sigma$.) Use this insight to propose an algorithm for simulating multivariate normal random variables with a specified mean and covariance matrix.

   Proof. Covariance matrices are positive-definite, which implies that there exists a matrix L such that $L'L = \Sigma = \text{Cov}(x)$. Denote $\mu = \mathbb{E}[x]$ and let $z = (z_i)_{i=1}^p \sim \text{NmV}(0, I)$, where p is the dimension of the space x maps to. Then the mgf of $Lz + \mu$ equals

$$\mathbb{E}\left[\exp\left\{t'(Lz+\mu)\right\}\right] = \mathbb{E}\left[\exp\left\{(t'L)z\right\}\right] + \exp\left\{t'\mu\right\} = \exp\left\{\frac{1}{2}t'LL't + t'\mu\right\} = \exp\left\{\frac{1}{2}t'\Sigma t + t'\mu\right\}.$$

   Note how this is equals the mgf of x, hence their distributions are the same and we can conclude the desired representation holds.

   This also implies that to simulate a normal multivariate with arbitrary mean an covariance matrix we can simply simulate standard normal multivariate random variables and transform them by decomposing the desired covariance matrix and using this representation. □

6. Use this last result, together with the PDF of a standard multivariate normal, to show that the PDF of a multivariate normal $x \sim \text{N}(\mu, \Sigma)$ takes the form $p(x) = C\exp\{-Q(x-\mu)/2\}$ for some constant C and quadratic form $Q(x-\mu)$.[3]

   Proof. Decompose $\Sigma = LL'$. Let z be a standard multivariate normal vector. Since the entries are uncorrelated and Gaussian they are independent, so the joint density is the product of the marginal densities. Furthermore, the positive-definiteness of Σ implies that it is full rank and hence invertible. Then we can use the mapping $x \mapsto Lx + \mu$ with Jacobian L to get

$$\begin{aligned}f_x(x) &= (2\pi)^{p/2}|L^{-1}|\exp\left\{-\frac{1}{2}(L^{-1}(x-\mu))'(L^{-1}(x-\mu))\right\} \\ &= (2\pi)^{p/2}|L^{-1}|\exp\left\{-\frac{1}{2}(x-\mu)'(L'L)^{-1}(x-\mu)\right\} \\ &= (2\pi)^{p/2}|L^{-1}|\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\},\end{aligned}$$

   obtaining the desired form for the pdf. □

7. Let $x_1 \sim N(\mu_1, \Sigma_1)$ and $x_2 \sim N(\mu_2, \Sigma_2)$, where x_1 and x_2 are independent of each other. Let $y = Ax_1 + Bx_2$ for matrices A, B of full column rank and appropriate dimension. Note that x_1 and x_2 need not have the same dimension, as long as Ax_1 and Bx_2 do. Use your previous results to characterize the distribution of y.

   Let us find the mgf of y. Using the independence,

$$\mathbb{E}\left[\exp\left\{t'y\right\}\right] = \mathbb{E}\left[\exp\left\{t'Ax_1 + t'Bx_2\right\}\right] = \mathbb{E}\left[\exp\left\{(t'A)x_1\right\}\right]\mathbb{E}\left[\exp\left\{(t'B)x_2\right\}\right] = \exp\left\{\frac{1}{2}t'A\Sigma_1A't + t'A\mu_1 + \frac{1}{2}t'B\Sigma_2B't + t'B\mu_2\right\},$$

   from where we recognize the mgf of a $\text{N}(A\mu_1 + B\mu_2, A\Sigma_1A' + B\Sigma_2B')$.

[2] The full rank restriction turns out to be unnecessary; relaxing it leads to what is called the *singular normal distribution*.

[3] A useful fact is that the Jacobian matrix of the linear map $x \to Ax$ is simply A.

1.2 Conditionals and marginals

1. Derive the marginal distribution of x_1. (Remember your result about affine transformations.)

   x_1 can be obtained by the transformation $a = (10)'$, where both 1 and 0 are vectors of ones and zeros. Then

   $$x_1 \sim \mathrm{N}\left(a'\mu, a\Sigma a'\right) = \mathrm{N}\left(\mu_1, \Sigma_{22}\right).$$

2. Let $\Omega = \Sigma^{-1}$ be the inverse covariance matrix, or precision matrix, of x, and partition Ω just as you did Σ:

   $$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}^T & \Omega_{22} \end{pmatrix}.$$

   Using (or deriving!) identities for the inverse of a partitioned matrix, express each block of Ω in terms of blocks of Σ.

   Using Wikipedia's identity for inverses of block matrices[4],

   $$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \\ -(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix}^{-1}$$

3. Derive the conditional distribution for x_1, given x_2, in terms of the partitioned elements of x, μ, and Σ. There are several keys to inner peace: work with densities on a log scale, ignore constants that don't affect x_1, and remember the cute trick of completing the square from basic algebra.[5] Explain briefly how one may interpret this conditional distribution as a linear regression on x_2, where the regression matrix can be read off the precision matrix.

   First write

   $$\Sigma^{-1} = \begin{pmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \\ -(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{pmatrix} = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}' & \Omega_{22} \end{pmatrix}$$

   Taking a Bayesian approach we get,

   $$\begin{aligned}
   p(x_1|x_2) &\propto p(x_1, x_2) \\
   &\propto \exp\left\{(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \\
   &\propto \exp\left\{(x-\mu)'\begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{12}' & \Omega_{22} \end{pmatrix}(x-\mu)\right\} \\
   &= \exp\left\{-\frac{1}{2}\left((x_1-\mu_1)'\Omega_{11}(x_1-\mu_1) - (x_2-\mu_2)'\Omega_{21}(x_1-\mu_1) - (x_1-\mu_1)'\Omega_{12}(x_2-\mu_2)\right)\right\} \\
   &= \exp\left\{-\frac{1}{2}\left(x_1'\Omega_{11}x_1 - 2\mu_1'\Omega_{11}x_1 - x_2'\Omega_{21}x_1 + \mu_2'\Omega_{21}x_1 - x_1'\Omega_{12}x_2 + x_1'\Omega_{12}\mu_2\right)\right\} \\
   &\propto \exp\left\{-\frac{1}{2}\left(x_1'\Omega_{11}x_1 - 2\mu_1'\Omega_{11}x_1 - 2x_2'\Omega_{21}x_1 - 2\mu_2'\Omega_{21}x_1\right)\right\} \\
   &= \exp\left\{-\frac{1}{2}\left(-2(\mu_1'\Omega_{11} + x_2'\Omega_{21} - \mu_2'\Omega_{21})x_1 + x_1'\Omega_{11}x_1\right)\right\} \\
   &= \exp\left\{-\frac{1}{2}\left(-2\tilde{\mu} + x_1'\tilde{\Sigma}^{-1}x_1\right)\right\} \\
   &\propto \exp\left\{-\frac{1}{2}(x_1-\tilde{\mu})'\tilde{V}^{-1}(x_1-\tilde{\mu})\right\},
   \end{aligned}$$

   where

   $$\tilde{\Sigma}^{-1} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \Omega_{11}^{-1}$$

   and

   $$\tilde{\mu} = \Omega_{11}'\mu_1 + \Omega_{21}'x_2 - \Omega_{21}'\mu_2.$$

[4] c.f. `https://en.wikipedia.org/wiki/Block_matrix`

[5] In scalar form:

$$\begin{aligned} x^2 - 2bx + c &= x^2 - 2bx + b^2 - b^2 + c \\ &= (x-b)^2 - b^2 + c\,. \end{aligned}$$

1.3 Multiple regression: three classical principles for inference

1. Show that all three of these principles lead to the same estimator. What is the variance of this estimator under the assumption that each ε_i is independent and identically distributed with variance σ^2?

   Proof. We can begin with the least squares estimator.

   $$\begin{aligned}\frac{\partial}{\partial \beta}(Y - X\beta)'(Y - X\beta) &= \frac{\partial}{\partial \beta}(Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta) \\ &= -2X'Y + 2X'X\beta,\end{aligned}$$

   which, assuming full rank of X, implies that

   $$\hat{\beta}_{LSE} = (X'X)^{-1}X'Y.$$

   Now let us get the MLE.

   $$\begin{aligned}\hat{\beta}_{MLE} &= \arg\max_{\beta}\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right\} \\ &= \arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - x_i'\beta)^2\right\}\end{aligned}$$

   This conicides then to the LSE estimator.

   For the method of moments, we need to set $\operatorname{Cov}\left(\varepsilon_i, x_j, =\right)0$. Note that $\varepsilon = Y - X\beta$, and since $\mathbb{E}[\varepsilon] = 0$, this is equivalent to

   $$0 = X'(Y - X\beta) = X'Y - X'X\beta,$$

   so we again obtain the same estimators. □

2. What happens if we isntead postulate that $y \sim \mathrm{N}(X\beta, \Sigma)$, where Σ is an arbitrary known covariance matrix, not necessarily proportional to the identity? What is the mle for β now, and what is the variance of this estimator?

   Note that

   $$(y - x'\beta)'\Sigma^{-1}(y - x'\beta) = (y - x'\beta)'L'^{-1}L^{-1}(y - x'\beta) = (L^{-1}y - L^{-1}x'\beta)'(L^{-1}y - L^{-1}x'\beta),$$

   then

   $$\hat{\beta}_{W-MLE} = (X'L^{-1\prime}L^{-1}X)^{-1}X'L^{-1\prime}L^{-1}Y = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y.$$

   The variance is given by

   $$\operatorname{Var}\left(\hat{\beta}_{W-MLE}\right) = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\operatorname{Var}(Y)\left[(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}\right]' = (X'\Sigma^{-1}X)^{-1}$$

3. Take Σ to be a diagonal matrix and relate the above estimators to the weighted linear regression.

   Note that if the diagonal of Σ takes values σ_i for $i = 1, \ldots, n$, the Σ^{-1} is also a diagonal matrix with diagonal $\sigma_1 {-} 1$, so by taking $W = \Sigma^{-1}$, we get the normal equations

   $$(X'WX)\beta = X'WY,$$

   which correspond to a weighted linear regression with the inverse of each variance as the weights.

1.4 Some practical details

1. Numerically speaking, is the inversion method the fastest and most stable way to actually solve the normal equations?

   No. As said in the referenced blog post, inverting a matrix requires approximately $\frac{2}{3}n^3 + 2n^2$ floating points operations, whereas factorizing it and solving the system by backward-forward substitution takes only $\frac{14}{3}n^3$, making the first choice slower.

   Moreover, inverting an ill-conditioned matrix can be numerically unstable, specially comparing it to factorization methods that require less operations and so the numerical error has less chance to propagate. This reinforces the idea of not inverting a matrix and instead using factorizations.

   We can sketch an algorithm to solve the normal equations using the LU decomposition as follows:

Algorithm 1 Solution of the normal equations

```
Decompose $X'\Sigma^{-1}X$ as $LU$ where $L$ is lower triangular and $U$ an upper triangular.
Solve $Lz = X'\Sigma^{-1}Y$ using backwards substitution.
Solve $U\hat{\beta} = z$.
```

2. Code the above algorithm and implement it for simulated data. Benchmark its performance against an inversion solver.

   The results are presented on table 1,

n	p	Inv time	LU time
100	60	145.53	13.21
200	1600	279.21	24.12
500	400	532.43	39.87

Table 1: Caption

   Clearly the LU factorization improved the computing time by several orders of magnitude.

2 Generalized Linear Models

1. Starting from the standard from of these pdf/pmf, show that the following distributions are in an exponential families, and find the corresponding b, c, θ and $a(\phi)$.

   (a) $Y \sim \mathrm{N}\left(\mu, \sigma^2\right)$ for known σ^2.

$$\begin{aligned} f(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(y^2-2y\mu+\mu^2)\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}(y^2-2y\mu+\mu^2)-\log\sqrt{2\pi\sigma^2}\right\} \\ &= \exp\left\{\frac{y\mu-\frac{1}{2}\mu^2}{\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2+\frac{y^2}{\sigma^2}\right)\right\} \end{aligned}$$

   Then by making

$$b(\mu)=\mu^2/2,\ a(\sigma^2)=\sigma^2 \text{ and } c(y,\sigma^2)=-\frac{1}{2}\log\left(2\pi\sigma^2+\frac{y^2}{\sigma^2}\right),$$

   we obtain that the normal distribution belongs to the exponential family.

   (b) $Y = Z/N$, where $Z \sim \text{Binom}(N, p)$ for known N.

$$\begin{aligned} \mathbb{P}(y=xN|p) &= \binom{N}{Nx}(1-p)^{N-x/N}p^{x/N} \\ &= \exp\left\{Nx\log\left(\frac{p}{1-p}\right)-(-N\log(1-p))+\log\binom{N}{NX}\right\} \end{aligned}$$

   Taking

$$\theta=\log\frac{p}{1-p},\ b(\theta)=-N\log(1-p),\ a(\phi)=1 \text{ and } c(y,\sigma^2)=\log\binom{N}{Nx},$$

   we see that the binomial distribution belongs to the exponential family.

   (c) $Y \sim \text{Poisson}(\lambda)$.

$$\begin{aligned} \mathbb{P}(Y=y|\lambda) &= e^{-\lambda}\frac{\lambda^y}{y!} \\ &= \exp\left\{y\log\lambda-\lambda-\log(y!)\right\} \end{aligned}$$

   Taking

$$\theta=\log\lambda,\ b(\theta)=\lambda,\ a(\phi)=1 \text{ and } c(y,\sigma^2)=-\log(y!),$$

   we see that the Poisson distribution belongs to the exponential family.

2. Prove the score equations.

Proof. Assume the conditions for interchanging integrals and derivatives hold for the density. Then

$$\begin{aligned}
\mathbb{E}[s(\theta)] &= \int f(y|\theta)\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y)\,\mathrm{d}y \\
&= \int f(y|\theta)\frac{1}{f(y|\theta)}\frac{\partial}{\partial\theta}f(y|\theta)\,\mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta}f(y|\theta)\,\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}\int f(y|\theta)\,\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}1 \\
&= 0.
\end{aligned}$$

For the variance, start by taking the derivative of the expectation, getting

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta'}\mathbb{E}[s(\theta)] \\
&= \frac{\partial}{\partial\theta'}\int f(y|\theta)\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y)\,\mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta'}\left(f(y|\theta)\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y)\right)\,\mathrm{d}y \\
&= \int \frac{\partial}{\partial\theta'}f(y|\theta)\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y) + f(y|\theta)\frac{\partial^2}{\partial\theta'\partial\theta}\mathscr{L}(\theta|y)\,\mathrm{d}y \\
&= \int f(y|\theta)\frac{\partial}{\partial\theta'}\log\mathscr{L}(\theta|y)\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y)\,\mathrm{d}y + \int f(y|\theta)\frac{\partial^2}{\partial\theta'\partial\theta}\mathscr{L}(\theta|y)\,\mathrm{d}y \\
&= \mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log\mathscr{L}(\theta|y)\right] + \mathbb{E}\left[\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y)\left(\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y)\right)'\right].
\end{aligned}$$

This implies that

$$\mathbb{E}\big[s(\theta)s(\theta)'\big] = -\mathbb{E}\left[\frac{\partial}{\partial\theta}\log\mathscr{L}(\theta|y\right]$$

□

3. Use the score equations to show that if $y \sim f(y|\theta)$ is in an exponential family then

$$\mathbb{E}[y] = b'(\theta) \quad \text{and} \quad \mathrm{Var}(y) = a(\phi)b''(\theta).$$

Proof. Note that

$$\frac{\partial}{\partial\theta}\mathscr{L}(\theta|y) = \frac{y-\theta-b'(\theta)}{a(\phi)}$$

Then

$$\mathbb{E}[s(\theta)] = \mathbb{E}\left[\frac{y-\theta-b'(\theta)}{a(\phi)}\right] = 0,$$

which implies that

$$\mathbb{E}[y] = b'(\theta).$$

Similarly,

$$\frac{\partial^2}{\partial\theta^2}L(\theta|y) = -\frac{b''(\theta)}{a(\phi)},$$

so

$$\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}L(\theta|y)\right] = -\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}L(\theta|y)\right)^2\right] = .\mathbb{E}\left[\left(\frac{y-b'(\theta)}{a(\phi)}\right)^2\right] = -\frac{b''(\theta)}{a(\phi)}.$$

which shows that

$$b''(\theta) = \mathrm{Var}(y)/a(\phi).$$

□

4. Use these results to compute the mean and variance of the $\mathrm{N}\left(\mu, \sigma^2\right)$ distribution.

$$\mathbb{E}[y] = \frac{1}{2}\mu^{2'} = \mu \text{ and } \mathrm{Var}(y) = \sigma^2 * 1 = \sigma^2.$$

2.1 Generalized Linear Models

1. Deduce that in a GLM,

$$\theta_i = (b')^{-1}\left(g^{-1}(x_i'\beta)\right)$$
$$\mathrm{Var}(Y_i) = \frac{\phi}{w_u}V(\mu_1).$$

Proof. The first identity comes from the fact that

$$g(\mu_i) = g(\mathbb{E}[Y_i]) = g(b'(\theta_i)) = x_i'\beta,$$

so inverting the functions we get the result.

For the variance note that

$$\mathrm{Var}(Y_i) = a(\phi)b''(\theta) = \frac{\phi}{w_i}b''(b'^{-1}(\mu_i)),$$

so take $V = b'' \circ b'^{-1}$. □

2. Take two special cases.

   (a) Take a Poisson glm. Prove that $V(\mu) = \mu$.

   Proof. Note that since $b(\theta) = e^\theta$, $b''(\theta) = e^\theta$ and $b'^1(\theta) = \log\theta$, so $V = b'' \circ b'^{-1} = Id$. □

   (b) Use the Binomial glm obtained from scaling. Show that $V(\mu) = \mu(1-\mu)$.

   Proof. Here $\theta = \log(\mu/(1-\mu)$ and $b(\theta) = \log(1+e^\theta)$, so

$$b'(\theta) = \frac{e^\theta}{1+e^\theta} \quad \text{and} \quad b''(\theta) = \frac{e^\theta}{(1+e^\theta)^2}$$

   From here we get that $V(\mu) = \mu(1-\mu)$. □

3. Show that the canonical link functions of the following glms are:

   (a) Poisson glm - the link is $g(\mu) = \log\mu$.
   Since $b(\mu) = e^\mu$, $g(\mu) = \log\mu$.

   (b) Binomial glm - $g(\mu) = \log\{\mu(1-\mu)\}$. From the form of $b'(\theta)$ we can invert the function and get the desired result.

3 Fitting GLMs

(a) Using the chain rule

$$\frac{\partial}{\partial\beta} = \frac{\partial}{\partial\theta} \times \frac{\partial\theta}{\partial\mu} \times \frac{\partial\mu}{\partial\beta},$$

show that

$$s(\beta,\phi) \equiv \nabla_\beta \ \log L(\beta,\phi) = \sum_{i=1}^{n} \frac{w_i(Y_i-\mu_i)x_i}{\phi V(\mu_i)g'(\mu_i)}$$

where x_i is the vector of predictors for case i (i.e. row i of the predictor matrix X, transposed to be a column vector).

Proof.

$$\frac{\partial \mu}{\partial \beta} = \frac{x_i'}{g'(g^{-1}(x_i'\beta))} = \frac{x_i'}{g'(\mu_i)}$$

Also,

$$\frac{\partial \theta}{\partial \mu} = \frac{1}{b''(\theta)}$$

Finally

$$\frac{\partial}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\phi/w_i} = \frac{y_i - \mu_i)}{\phi/w_i}$$

So by taking the product we get the desired result. □

(b) Show that under the canonical link, $g'(\mu) = 1/V(\mu)$, so that the score function simplifies to:

$$s(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i(Y_i - \mu_i)x_i}{\phi}.$$

Proof. If $g(\mu) = (b')^{-1}(\mu)$, then

$$g'(\mu) = \frac{1}{b''((b')^{-1}(\mu))} = \frac{1}{V(\mu)}.$$

□

(c) Fit a logistic regression using gradient descent.
Note that the gradient descent iteration can be stated as

$$\beta^{(i+1)} = \beta^{(i)} + \gamma \nabla_\beta \log L(\beta) = \beta^{(i)} + \gamma s(\beta, \phi) = \beta^{(i)} + \gamma \sum_{i=1}^{n} (Y_i - g^{-1}(x_i'\beta))x_i$$

If we use the canonical link, then

$$\beta^{(i+1)} = \beta^{(i)} + \gamma \sum_{i=1}^{n} \left(Y_i - \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}} \right) x_i.$$

Implementing the algorithm in R using linesearch for optimizing the step size we obtain the loglikelihood values shown in figure 1. We compare it with the value obtained with the pre-built glm() function in R, which shows that this method, while slower glm(), converges to the right value of the MLE.

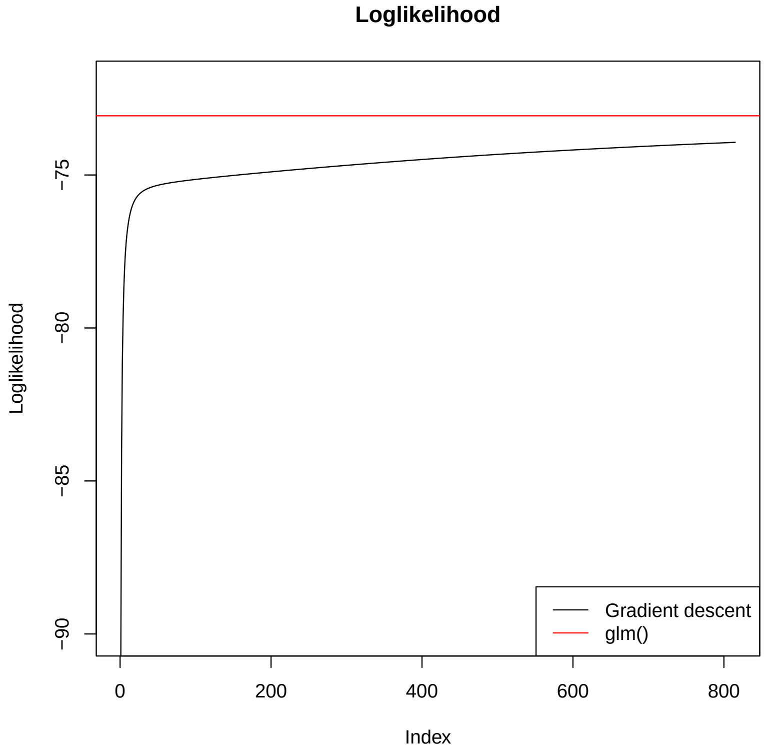


Figure 1: Iterations vs loglikelihood values.

(d) Now consider a point $\beta_0 \in \mathscr{R}^P$, which serves as an intermediate guess for our vector of regression coefficients. Show that, for any GLM, the second-order Taylor approximation of $\log L(\beta, \phi)$, around the point β_0, can be expressed in the form

$$q(\beta; \beta_0) = -\frac{1}{2}(\tilde{y} - X\beta)^T W(\tilde{y} - X\beta) + c\,,$$

where $\tilde{y}$ is a vector of "working responses" and W is a diagonal matrix of "working weights," and c is a constant that doesn't involve β. Give explicit expressions for the diagonal elements W_{ii} and for $\tilde{y}$ (which will necessarily involve the point β_0, around which you're doing the expansion). Again, we're assuming the canonical link to make the algebra a bit simpler.

Proof. First note that

$$\begin{aligned}\nabla_\beta L(\beta)'|_{\beta=\beta_0} &= \sum_{i=1}^{n} \frac{w_i(y_i - \mu_i)x_i'}{\phi} \\ &= \sum_{i=1}^{n} \frac{w_i b''(x_i'\beta)(y_i - \mu_i)x_i'}{b''(x_i'\beta)\phi} \\ &= z'WX,\end{aligned}$$

where $z = \left(\frac{y_i - \mu_i}{b''(x_i'\beta)}\right)_i$ and $W_{ii} = \frac{w_i}{\phi} b''(x_i'\beta)$.

Now,

$$\begin{aligned}q(\beta, \beta_0) &= L(\beta_0) + \nabla_\beta L(\beta)'|_{\beta=\beta_0}(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)'H(\beta_0)(\beta - \beta_0) \\ &= L(\beta_0) + z_{\beta_0}'WX(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)'X'WX(\beta - \beta_0) \\ &= z_{\beta_0}'WX\beta - \frac{1}{2}\beta'X'WX\beta + 2\beta_0'X'WX\beta + K \\ &= -\frac{1}{2}\beta'X'WX\beta + \tilde{y}'WX\beta + K \\ &= -\frac{1}{2}(\tilde{y} - X\beta)'W(\tilde{y} - X\beta) + K,\end{aligned}$$

where $\tilde{y} = \beta_0'X' + z_{\beta_0}'$. □

(e) Read up on Newton's method for optimizing smooth functions (e.g. in Nocedal and Wright, Chapter 2). Implement it for the logistic regression model model and test it out on the same data set you just used to test out gradient descent. Note: while you could do line search, there is a "natural" step size of 1 in Newton's method. Verify that your solution replicates the β estimate you get when using a package solver, e.g. the glm function in R, up to minor numerical differences.

Using Newton's method the iterations are

$$\beta^{(i+1)} = \beta^{(i)} + H^{-1}_{L(\hat{\beta}} \nabla L(\beta)$$

The results are shown in figure 2. We can tell the convergence was much faster than with the first order algorithm in the first order approximation. This was expeceted since Taylor series approximate better as you increase the number of terms.

The (L_1) differences in the betas between our method and glm() is shown in following table. We can tell the method was precise.

β
-2.34E-05
3.35E-04
-1.53E-05
-1.21E-04
-2.90E-04
-1.09E-05
7.49E-07
-7.42E-07
-2.06E-05
-5.44E-06
8.35E-06

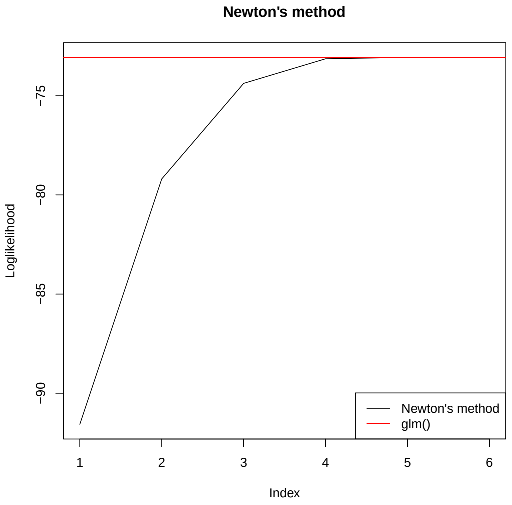


Figure 2: Iterations vs loglikelihood values

(f) Standard asymptotic theory, which we won't go into here, implies that the maximum likelihood estimator is consistent and asymptotically normal around the true value β_0:

$$\hat{\beta} \sim N(\beta_0, I(\beta_0, \phi)^{-1}),$$

where $\mathcal{I}(\beta_0, \phi)$, called the *Fisher information matrix,* is the same $\mathcal{I}$ you met all the way back when you proved the score equations:

$$\mathcal{I}(\beta_0, \phi) \equiv \text{var}\{s(\beta_0, \phi)\} = -E\left\{H(\beta_0, \phi)\right\} .$$

The fact that Fisher information is the negative of the expected Hessian motivates the following idea: use the inverse of the negative Hessian matrix at the MLE to approximate the inverse Fisher information, i.e. the covariance matrix of the estimator. Happily, you get this Hessian matrix for free when fitting by Newton's method.

For your logistic regression on the WDBC data fit via Newton's method, compute the square root of each diagonal element of the inverse Hessian matrix, evaluated at the MLE.[6] Compare these to the standard errors you get when using a package solver, e.g. the glm function in R.

The results are shown in table 2. Again, the differences in absolute value are quite small, showing the approximation was good.

-9.973e-4	-2.129e-2	-6.24e-4	-1.656e-2	-1.104e-2	
-5.740e-4	-1.223e-3	-5.481e-4	-1.511e-3	-2.787e-4	-7.791e-4

Table 2: Standard errors for all the betas.

4 Bayes and the Gaussian linear model

4.1 A simple Gaussian location model

1. By construction, we know that the marginal prior distribution $p(\theta)$ is a gamma mixture of normals. Show that this takes the form of a centered, scaled t distribution:

$$p(\theta) \propto \left(1 + \frac{1}{\nu} \cdot \frac{(x-m)^2}{s^2}\right)^{-\frac{\nu+1}{2}}$$

with center m, scale s, and degrees of freedom ν, where you fill in the blank for m, s^2, and ν in terms of the four parameters of the normal-gamma family. Note: you did a problem just like this on a previous exercise! This shouldn't be a lengthy re-derivation.

[6] These are your standard errors for each coefficient, i.e. the square root of the variance of each coefficient's (approximate) sampling distribution.

Proof.

$$\begin{aligned}
p(\theta) &= \int p(x|\omega)p(\omega)\,\mathrm{d}\omega \\
&\propto \int f_{\omega|X}(\omega)\,\mathrm{d}\omega \\
&\propto \int \underbrace{\omega^{a+\frac{1}{2}-1}\exp\left\{-\left(b+\frac{1}{2}(x-m)^2\right)\omega\right\}}_{\text{Kernel of a Gamma distribution}}\mathrm{d}\omega \\
&= \frac{\left(\frac{b+(x-m)^2}{2}\right)^{\frac{a+1}{2}}}{\Gamma\left(\frac{a+1}{2}\right)},
\end{aligned}$$

which is proportional to the density of a Student-t with the required parameters. □

2. Assume the normal sampling model in Equation **??** and the normal-gamma prior in Equation **??**. Calculate the joint posterior density $p(\theta,\omega \mid \mathbf{y})$, up to constant factors not depending on ω or θ. Show that this is also a normal/gamma prior in the same form as above:

$$p(\theta,\omega \mid y) \propto \omega^{(d^*+1)/2-1}\exp\left\{-\omega\cdot\frac{\kappa^\star(\theta-\mu^*)^2}{2}\right\}\cdot\exp\left\{-\omega\cdot\frac{\eta^\star}{2}\right\} \tag{1}$$

From this form of the posterior, you should able to read off the new updated parameters, by pattern-matching against the functional form in Equation **??**:

- $\mu \longrightarrow \mu^\star = ?$
- $\kappa \longrightarrow \kappa^\star = ?$
- $d \longrightarrow d^\star = ?$
- $\eta \longrightarrow \eta^\star = ?$

You may notice that my parameterization of the normal-gamma in Equation **??** differs from, say, the one you might find in textbooks or on websites. I've chosen this parameterization in order to make these four updates for the parameters, above, as simple-looking and intuitive as possible.

Tip: this one is a bit of an algebra slog, with a lot of completing the square, collecting common terms, and cancelling positives with negatives. For example, to make the calculations go more easily, you might first show (or recall, from a previous exercise) that the likelihood can be written in the form

$$p(y \mid \theta,\omega) \propto \omega^{n/2}\exp\left\{-\omega\cdot\left(\frac{S_y+n(\bar{y}-\theta)^2}{2}\right)\right\},$$

where $S_y=\sum_{i=1}^{n}(y_i-\bar{y})^2$ is the sum of squares for the y vector. This expresses the likelihood in terms of the two statistics $\bar{y}$ and S_y, which you may recall from your math-stat course are sufficient statistics for (θ,σ^2).

Take care in ignoring constants here: some term that is constant in θ may not be constant in ω, and vice versa. You're focusing on the joint posterior, so you can't ignore anything that has a θ or ω in it.

Proof. Start by noticing that

$$\sum(x_i-\mu)^2=\sum(x_i+\bar{x}-\bar{x}-\mu)^2=\sum(x_i-\bar{x})^2+n(\bar{x}-\mu)^2+\sum(x_i-\bar{x})(\bar{x}-\mu)=\sum(x_i+\bar{x}-\bar{x}-\mu)^2=nS^2+n(\bar{x}-\mu)^2$$

Then

$$\begin{aligned}
p(\theta,\omega|y) &\propto p(\theta,\omega)p(y|\theta,\omega) \\
&\propto \omega^{(d+1)/2-1}\exp\left\{-\omega\cdot\frac{\kappa(\theta-\mu)^2}{2}\right\}\cdot\exp\left\{-\omega\cdot\frac{\eta}{2}\right\}\omega^{n/2}\exp\left\{-\omega\cdot\left(\frac{S_y+n(\bar{y}-\theta)^2}{2}\right)\right\}
\end{aligned}$$

Now, expanding the quadratic exponents we get

$$\begin{aligned}
S_y + n(\bar{y}-\theta)^2 + \kappa(\theta-\mu)^2 &= S_y + n\bar{y}^2 - 2n\bar{y}\theta + n\theta^2 + \kappa\theta^2 - 2\kappa\theta\mu + \kappa\mu^2 \\
&= (n+\kappa)\theta^2 - 2(n\bar{y}+\kappa\mu)\theta + \left(S_y + n\bar{y}^2 + \kappa\mu^2\right) \\
&= (n+\kappa)\left(\theta^2 - 2\frac{n\bar{y}+\kappa\mu}{n+\kappa}\theta + \frac{1}{n+\kappa}\left(S_y + n\bar{y}^2 + \kappa\mu^2\right)\right) \\
&= (n+\kappa)\left(\theta^2 - \frac{n\bar{y}+\kappa\mu}{n+\kappa}\right)^2 + \left(S_y + n\bar{y}^2 + \kappa\mu^2 - \frac{(n\bar{y}+\kappa\mu)^2}{n+\kappa}\right) \\
&= (n+\kappa)\left(\theta^2 - \frac{n\bar{y}+\kappa\mu}{n+\kappa}\right)^2 + \left(S_y + \kappa n(\bar{y}-\mu)^2\right)
\end{aligned}$$

Let

- $\mu^\star = \dfrac{n\bar{y}+\kappa\mu}{n+\kappa}$
- $\kappa^\star = n + \kappa$
- $d^\star = d + n$
- $\eta^\star = \eta + S_y + \dfrac{n\kappa(\bar{y}-\mu)^2}{\kappa+n}$

Then we obtain the posterior distribution

$$p(\theta,\omega|y) \equiv \text{NormalGamma}\left(\mu^*, \kappa^*, d^*, \eta^*\right).$$

□

3. From the joint posterior you just derived, what is the conditional posterior distribution $p(\theta \mid y, \omega)$?

   This is just simply a normal distribution with mean μ^* and precision κ^*.

4. From the joint posterior you calculated in (A), what is the marginal posterior distribution $p(\omega \mid y)$?

$$\begin{aligned}
p(\omega|y) &= \int \omega^{\frac{d^*+1}{2}-1} \exp\left\{-\omega\frac{\kappa^*(\theta-\mu^*)^2}{2}\right\} \exp\left\{-\omega\frac{\eta^*}{2}\right\}\, \mathrm{d}\theta \\
&= \omega^{\frac{d^*+1}{2}-1} \exp\left\{-\omega\frac{\eta^*}{2}\right\} \int \underbrace{\exp\left\{-\omega\frac{\kappa^*(\theta-\mu^*)^2}{2}\right\}}_{\text{Kernel of N}(\mu^*,\omega\kappa^*)}\, \mathrm{d}\theta \\
&\propto \omega^{\frac{d^*+1}{2}-1} \exp\left\{-\omega\frac{\eta^*}{2}\right\} \frac{1}{\sqrt{\omega\kappa^*}},
\end{aligned}$$

so $\omega|y \sim \text{Gamma}\left(\dfrac{d^*}{2}, \dfrac{\eta^*}{2}\right)$.

5. Show that the marginal posterior $p(\theta \mid y)$ takes the form of a centered, scaled t distribution and express the parameters of this t distribution in terms of the four parameters of the normal-gamma posterior for (θ, ω).

   Integrating with respect to ω we get a Normal-gamma mixture, which by the first exercise we know it is a t-distribution with parameters d^*, μ^* and $(\eta^*/d^*)^{1/2}$.

6. True or false: in the limit as the prior parameters κ, d, and η approach zero, the priors $p(\theta)$ and $p(\omega)$ are valid probability distributions.

   False. When the parameters are set to zero then we obtain singular σ-finite measures that cannot be normalized.

7. True or false: in the limit as the prior parameters κ, d, and η approach zero, the posteriors $p(\theta \mid y)$ and $p(\omega \mid y)$ are valid probability distributions.

   True. The posterior parameters would not be zero (unless we do not observe anything), so we keep the distribution.

8. Your result in (E) implies that a Bayesian credible interval for θ takes the form

$$\theta \in m \pm t^{\star} \cdot s\,,$$

where m and s are the posterior center and scale parameters from (E), and $t^{\star}$ is the appropriate critical value of the t distribution for your coverage level and degrees of freedom (e.g. it would be 1.96 for a 95% interval under the normal distribution).

True or false: In the limit as the prior parameters κ, d, and η approach zero, the Bayesian credible interval for θ becomes identical to the classical (frequentist) confidence interval for θ at the same confidence level.

True, since the posteriors would become the sample statistics.

4.2 The conjugate Gaussian linear model

1. Derive the conditional posterior $p(\beta|\gamma, y)$.

Proof.

$$\begin{aligned}
p(\beta|\omega, y) &\propto p(y|\beta, \sigma^2)\, p(\beta|\omega)\, p(\omega) \\
&\propto \exp\left\{-\frac{1}{2}(y - X\beta)'\omega\Lambda(y - X\beta)\right\} \exp\left\{-\frac{1}{2}(\beta - m)'\omega K(\beta - m)\right\} \omega^{d/2-1} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2y'\omega\Lambda X\beta + \beta' X'\omega\Lambda X\beta + \beta'\omega K\beta - 2m'\omega K\beta\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(-2(y'\omega\Lambda X + m'\omega K) + \beta'(X'\omega\Lambda X + \omega K)\beta\right)\right\}.
\end{aligned}$$

Let $\tilde{b} = y'\omega\Lambda X + m'\omega K$ and $\tilde{\Lambda} = X'\omega\Lambda X + \omega K$. Then

$$p(\beta|\omega, y) \equiv \mathrm{N}\left(\tilde{\Lambda}^{-1}\tilde{b}, \tilde{\Lambda}^{-1}\right).$$

□

2. Derive the marginal posterior $p(\omega|y)$.

Proof.

$$\begin{aligned}
p(\omega|y) &\propto \omega^{\frac{n+d}{2}-1} \exp\left\{-\frac{\omega}{2}(n + y'\Lambda y + m'Km)\right\} \int \exp\left\{-\frac{1}{2}\left(-2(y'\omega\Lambda X\beta + m'\omega K\beta\right) + \beta'(X'\omega\Lambda X + \omega K)\beta\right\}\, \mathrm{d}\beta \\
&\propto \omega^{\frac{n+p+d}{2}-1} \exp\left\{-\frac{\omega}{2}(n + y'\Lambda y + m'Km)\right\} |\omega^{-1}\tilde{\Lambda}^{-1}|^{1/2} \exp\left\{-\frac{1}{2}(b'\tilde{\Lambda} b)\right\} \\
&\propto \omega^{\frac{n+p+d}{2}-1} \exp\left\{-\frac{\omega}{2}(n + y'\Lambda y + m'Km)\right\} \exp\left\{\frac{\omega}{2}(y'\Lambda X + m'K)'(X'\Lambda X + K)^{-1}(y'\Lambda K + m'K)\right\},
\end{aligned}$$

so

$$p(\omega|y) \equiv \mathrm{Gamma}\left(\frac{n+p+d}{2}, \frac{\eta^*}{2}\right),$$

where

$$\eta^* = \eta + y'\Lambda y + m'Km - (y'\Lambda X + m'K)'(X'\Lambda X + K)^{-1}(y'\Lambda X + m'K)$$

□

3. Putting these together, derive the conditional posterior $p(\beta|y)$.

Proof.

$$\begin{aligned}
p(\beta|y) &\propto \int p(\omega, \beta|y)\, \mathrm{d}\omega \\
&\propto \int p(\beta|\omega, y) p(\omega|y)\, \mathrm{d}\omega \\
&\propto \int \omega^{(n+p+d)/2-1} \exp\left\{-\frac{1}{2}(y - X\beta)'\omega\Lambda(y - X\beta)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2}(\beta - m)'\omega K(\beta - m)\right\} \exp\left\{-\frac{\omega\eta}{2}\right\}\, \mathrm{d}\omega.
\end{aligned}$$

The integrand is the the kernel of a Gamma distribution, so

$$
\begin{aligned}
p(\beta|y) &\propto \left(\frac{(\beta^*)^{a^*}}{\Gamma(a^*)}\right)^{-1} \\
&\propto \frac{\Gamma\left(\frac{n+p+d}{2}\right)}{\left(\frac{1}{2}\left[(y-X\beta)\Lambda(y-X\beta)+(\beta-m)'K(\beta-m)+\eta\right]\right)^{\frac{n+p+d}{2}}} \\
&\propto \left(\frac{1}{2}\left[(y-X\beta)'\Lambda(y-X\beta)+(\beta-m)'K(\beta-m)+\eta\right]\right)^{-\frac{n+p+d}{2}} \\
&\propto \left(\frac{1}{2}\left[-2(y'\Lambda X+m'K)\beta+\beta'(X'\Lambda X+K)\beta\right]\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left(\frac{1}{2}\left[-2(y'\Lambda X+m'K)\beta+\beta'\Lambda^*\beta\right]\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left((\beta-\mu^*)'\Lambda^*(\beta-\mu^*)+\eta^*\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left(\frac{1}{n+d}(\beta-\mu^*)'\frac{n+d}{\eta^*}\Lambda^*(\beta-\mu^*)+1\right)^{-\frac{\nu^*+p}{2}} \\
&\propto \left(\frac{1}{\nu^*}(\beta-\mu^*)'\Sigma^*(\beta-\mu^*)+1\right)^{-\frac{\nu^*+p}{2}}
\end{aligned}
$$

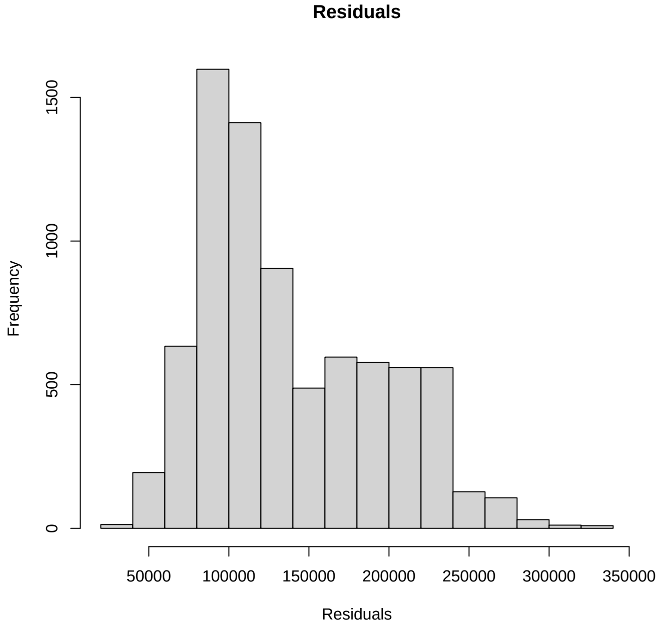
where $\nu^* = n+d$, $\Lambda^* = X'\Lambda X+K$, $\mu^* = (\Lambda^*)^{-1}(X^T\Lambda y+K'm)$ and $\Sigma^* = \frac{\nu^*}{\eta^*}\Lambda^*$.

This implies that the marginal posterior for β is a t-distribution with ν^* degrees of freedom, location μ^*, and covariance Σ^*. □

4. What is a 95% credible interval for the coefficient associated to the green rating? How does it compare to the one obtained using lm? What does the residual plot reveal?

   A 95% credible interval is given by [-0.03, 1.71] vs [-0.04, 1.7], the latter obtained via the `lm` function. Both are similar which shows the estimation was satisfactory.

   For the residuals plot in figure 1 we see that they are not well approximated by a normal distribution, which might be due to the unresolved spatial correlation of the data. This can be corrected by changing the prior for the variance to a more flexible distribution like a mixture.



4.3 A heavy tailed error model

1. Under this model, what is the implied conditional distribution $p(y_i|X,\beta,\omega)$?

$$p(y_i|X,\beta,\omega) = \int p(y_i|X,\beta,\omega,\Lambda)p(\Lambda)\,\mathrm{d}\Lambda$$

Now, $p(\Lambda)$ is a Gamma distribution and $p(y_i|X,\beta,\omega,\Lambda)$ is a normal distribution with variance Λ, so the conditional is a mixture of normal with mixing measure Gamma, *i.e.*, a t distribution with parameters $(\nu = h, m = x_i'\beta.s^2 = 1/\omega)$.

2. What is the conditional posterior $p(\lambda_i|y,\beta,\omega)$?

$$\begin{aligned}
p(\lambda_i|y,\beta,\omega) &\propto p(y|\beta,\lambda_i,\omega)\,p(\lambda_i) \\
&\propto (\omega\lambda_i)^{1/2}e^{-\frac{1}{2}\omega\lambda_i(y-x_i'\beta)^2}\lambda_i^{\frac{h}{2}-1}e^{-\lambda_i\frac{h}{2}} \\
&\propto e^{-\frac{1}{2}\lambda_i(\omega(y-x_i'\beta)^2+h)}\lambda_i^{\frac{h+1}{2}-1},
\end{aligned}$$

which is the kernel of a Gamma$\left(\dfrac{h+1}{2},\dfrac{1}{2}\left(\omega(y_i - x_i'\beta)+h\right)\right)$.

3. Code up a Gibbs sampler that repeatedly cycles through sampling the following three sets of conditional distributions:

   - $p(\beta|y,\omega,\Lambda)$.
   - $p(\omega|y,\Lambda)$.
   - $p(\lambda_i|y,\omega,\beta)$.

   We already know the last full conditional. For the one for β note that

$$p(\beta|y,\omega,\Lambda) \propto \mathrm{N}\left(y|X\beta,(\omega\Lambda)^{-1}\right)\mathrm{N}\left(\beta|m,(\omega K)^{-1}\right) \sim \mathrm{N}\left(\mu^*,(\omega\Lambda^*)^{-1}\right),$$

   where

$$\Lambda^* = X'\Lambda X + K \quad \text{and} \quad \mu^* = \left(\Lambda^{*\prime}\right)^{-1}(X'\Lambda y + K'm)$$

   For ω_i we run into a simmilar situation of a regression model with a NormalGamma distribution, so we have

$$p(\omega|y,\Lambda) \equiv \mathrm{Gamma}\left(\frac{n+d}{2},\frac{\eta^*}{2}\right),$$

   where

$$\eta^* = \eta + y'\Lambda y + m'Km - (y'\Lambda X + m'K)'(X'\Lambda X + K)^{-1}(y'\Lambda X + m'K).$$

5 Intro to hierarchical models

5.1 Math tests

1. Show (somewhat trivially) that the maximum likelihood estimate for θ is just the vector of sample means: $\hat{\theta}_{\mathrm{MLE}} = (\bar{y}_1,\ldots,\bar{y}_P)$.

   Proof. Since the observations of each school come from a normal distribution, the individual MLE is $\bar{y}_i$, so the (vector) MLE is the vector of sample means. □

2. Make a plot that illustrates the following fact: extreme school-level averages $\bar{y}_i$ (both high and low) tend to be at schools where fewer students were sampled. Explain briefly why this would be.

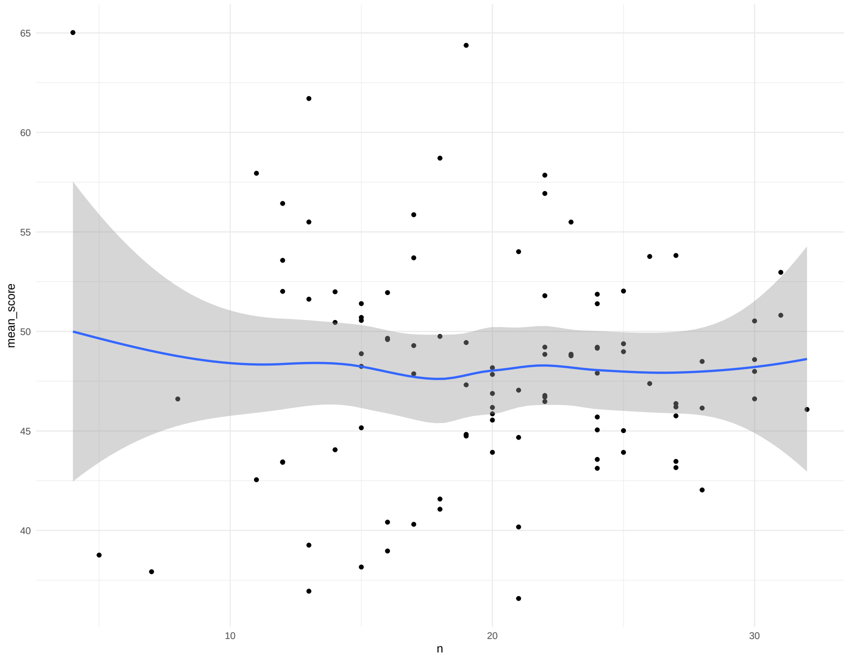


Figure 3: Number of students vs average score

In figure 3 we can tell a weak correlation between the number of students in each school vs the average score. The premise of the exercise can be seen form the boomerang-like shape the data cloud makes. This can be due to the fact that as the number of students grows the law of large numbers pulls the average towards the global mean, obtaining a regression effect.

3. Fit the following two-level hierarchical model to these data via Gibbs sampling:

$$\begin{aligned}(y_{ij} \mid \theta_i, \sigma^2) &\sim \mathrm{N}(\theta_i, \sigma^2)\\ (\theta_i \mid \tau^2, \sigma^2) &\sim \mathrm{N}(\mu, \tau^2\sigma^2)\end{aligned}$$

As a starting point, use a flat prior on μ, Jeffreys' prior on σ^2, and an inverse-gamma(1/2, 1/2) prior on τ^2. Your Gibbs sampler should cycle between the complete conditional posterior distributions for each of these parameters in turn, as well as θ (the vector of means). While you could update each θ_i individually, I encourage you to think about it as a vector whose conditional distribution is multivariate normal, and whose covariance matrix happens to be diagonal. This view will generalize more readily to future problems.

Solución.

To do this we need to first get the full conditionals. The joint posterior is given by

$$p(\mu, \sigma^2, \tau^2, \theta|y) \propto \frac{1}{\sigma^2}(\tau^2)^{-1.5}\exp\left\{-\frac{1}{2\tau^2}\right\}\left(\tau^2\sigma^2\right)^{-\frac{p}{2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta-\mu)'(\theta-\mu)\right\}(\sigma^2)^{-\frac{p\times n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\theta_i)^2\right\}$$

Let us obtain the full conditionals from easier to (arguably) hardest. We have

$$p(\mu|\cdot) \propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta-\mu)'(\theta-\mu)\right\} = \exp\left\{-\frac{1}{2\tau^2\sigma^2}\left(p\mu^2 - 2\mu\sum_{i=1}^{p}\theta_i + \sum_{i=1}^{p}\theta_i^2\right)\right\} \propto \mathrm{N}\left(\bar{\theta}, p\tau^2\sigma^2\right).$$

For τ^2,

$$p(\tau|\cdot) \propto (\tau^2)^{-1.5}\exp\left\{-\frac{1}{2\tau^2}\right\}\left(\tau^2\right)^{-\frac{p}{2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta-\mu)'(\theta-\mu)\right\} \equiv \mathrm{InvGamma}\left(\frac{1+p}{2}, \frac{(\theta-\mu)'(\theta-\mu)}{2\sigma^2}\right).$$

For σ^2

$$
\begin{aligned}
p(\sigma^2|\cdot) &\propto \frac{1}{\sigma^2}\left(\sigma^2\right)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2\tau^2\sigma^2}(\theta-\mu)'(\theta-\mu)\right\}(\sigma^2)^{-\frac{p\times n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\theta_i)^2\right\} \\
&= \left(\sigma^2\right)^{-\frac{p}{2}-1-\frac{p\times n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{\tau^2}(\theta-\mu)'(\theta-\mu)+\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\theta_i)^2\right]\right\} \\
&\equiv \text{InvGamma}\left(\frac{p(n+1)}{2}, \frac{1}{\tau^2}(\theta-\mu)'(\theta-\mu)+\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\theta_i)^2\right).
\end{aligned}
$$

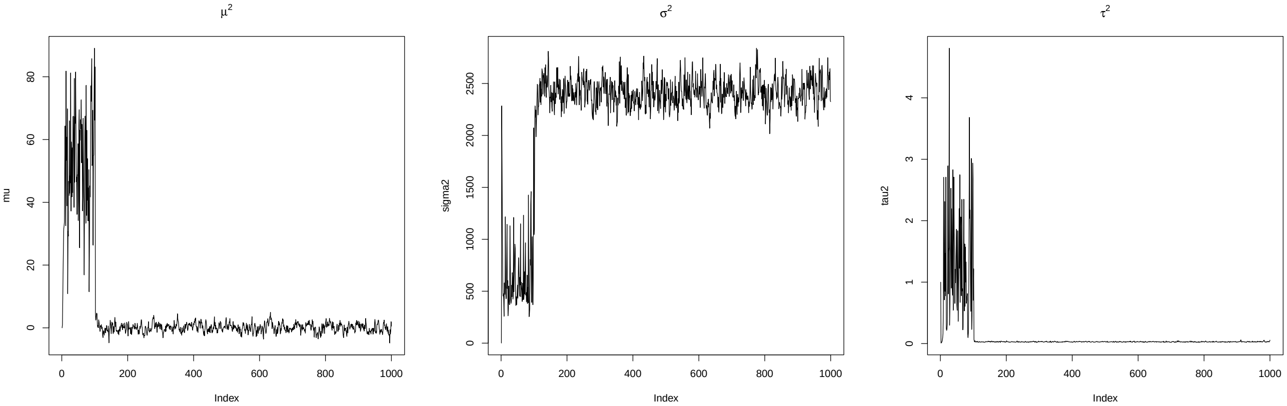
Finally note that for each θ_i we have a Normal-Normal model, so

$$p(\theta_i|\cdot) \equiv \text{N}\left(\mu_\theta, \sigma_\theta\right),$$

where if n_i is the number of students in the school i

$$\mu_{\theta_i} = \frac{\frac{1}{\tau^2\sigma^2}\mu + \frac{n_i}{\sigma^2}\bar{y}_i}{\frac{1}{\tau^2\sigma^2}+\frac{n_i}{\sigma^2}} \quad \text{and} \quad \sigma_n^2 = \frac{n_i}{\sigma^2}+\frac{1}{\sigma^2\tau^2}.$$

We can the run the Gibbs sampler 1,000 iterations, obtaining the traceplots shown in figure **??**



We can tell the chains appear to have reached stationarity.

4. Express the conditional posterior mean for θ_i in the following form:

$$E(\theta_i \mid y, \tau^2, \sigma^2, \mu) = \kappa_i\mu + (1-\kappa_i)\bar{y}_i\,,$$

i.e. a convex combination of prior mean and data mean. Here κ_i is a *shrinkage coefficient* whose form you should express in terms of the model hyperparameters. In the extreme case where κ_i is 1, then the data ($\bar{y}_i$) are essentially ignored, and the posterior mean is "shrunk" all the way back to the prior mean. In the other extreme where κ_i is 0, the prior mean is ignored, and the posterior mean is entirely "un-shrunk" compared to the MLE for θ_i.

For each draw of your MCMC, calculate κ_i for each school, and save the posterior draws. Average these MCMC samples to calculate $\bar{\kappa}_i$, the posterior mean of this shrinkage coefficient. Plot $\bar{\kappa}_i$ for each school as a function of that school's sample size, and comment on what you see.

Solution.

From the full conditional of each θ_i we can tell that

$$E(\theta_i \mid y, \tau^2, \sigma^2, \mu) = \kappa_i\mu + (1-\kappa_i)\bar{y}_i,$$

where

$$\kappa_i = \frac{\tau^2 n_i}{\tau^2 n_i + 1}$$

The resulting plot is shown in figure 4. A clear lorgarithmic trend is shown in which the shrinkage coefficient grows as n does but it does it more dramatically for small ns.

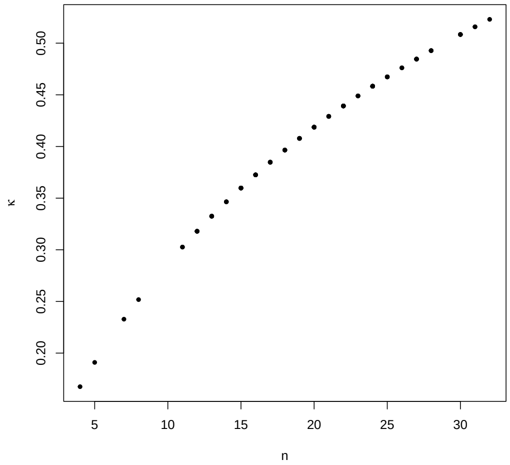


Figure 4: Number of students vs 1-kappa

5. Observe that an equivalent way to write your model involves the following decomposition:

$$y_{ij} = \mu + \delta_i + e_{ij}$$

where $\delta_i \sim N(0, \tau^2\sigma^2)$ and $e_{ij} \sim N(0, \sigma^2)$. (In the paper by Gelman that I've asked you to read, he writes it this way, where the school-level "offsets" are centered at zero, although he doesn't scale these offsets by σ the way I prefer to do.) To translate between the two parameterizations, just observe that in the previous version, $\theta_i = \mu + \delta_i$.

Conditional on the "grand mean" μ, but *marginally* over both δ_i and e_{ij}, compute the following two covariances:

- $\text{cov}(y_{i,j}, y_{i,k})$, $j \neq k$.

$$\begin{aligned}\text{Cov}\left(y_{ij}\,,\,y_{ik}\right) &= \text{Cov}\left(\mu + \delta_i + e_{ij}\,,\,\mu + \delta_i + e_{ik}\right)\\ &= \text{Cov}\left(\delta_i + e_{ij}\,,\,\delta_i + e_{ik}\right)\\ &= \mathbb{E}\left[\delta_i^2\right]\\ &= \tau^2\sigma^2\end{aligned}$$

- $\text{cov}(y_{i,j}, y_{i',k})$, $i \neq i'$ and $j \neq k$

$$\begin{aligned}\text{Cov}\left(y_{i'j}\,,\,y_{ik}\right) &= \text{Cov}\left(\mu + \delta_{i'} + e_{i'j}\,,\,\mu + \delta_i + e_{ik}\right)\\ &= \text{Cov}\left(\delta_{i'} + e_{i'j}\,,\,\delta_i e_{ik}\right)\\ &= 0.\end{aligned}$$

Does this make sense to you? Why or why not?

Yes?

6. Does the assumption that σ^2 is common to all schools look justified in light of the data?

Yes and no. The plot of κ shows that the model needs to effectively shrink the variance as the number of students increase. This is possible in this Bayesian model but from the pure frequentist approach you would need different variances.

5.2 Blood pressure

1. Is the experimental medication effective at reducing blood pressure? Do the naive thing and perform a t-test for a difference of means, pooling all the data from treatment 1 into group 1, and all the data from treatment 2 into group 2. What does this

t-test say about the difference between these two group means, and the standard error for the difference? Why is the t-test (badly) wrong?

Performing a one-tailed t-test we obtain a p-value lower than 2.2×10^{-16}, so based on this we can conclude with 95% confidence that the systolic pressure of the patients treated with the medication is in average lower than the ones in the control group. This of course is all based on the assumption that the measurements are independent, which is not since this is a longitudinal study.

Also, the standard error is of 1.004414.

2. Now do something better, but still less than ideal. Calculate $\bar{y}_i$, the mean blood pressure measurement for each patient. Now treat each person-level mean as if it were just a single data point, and conduct a different t-test for mean blood pressure between treatment 1 and treatment 2. (If you're doing this correctly, you should have only ten "observations" in each group, where each observation is actually a person-level mean.) What does this t-test say about the difference between these two group means, and the standard error for the difference? Why is the standard error so much bigger, and why is this appropriate? Even so, why is this approach (subtly) wrong?

   Doing this now we get a p-value of 0.05925, which lays in the uncomfortable border so the evidence for the efficacy of the treatment is reduced. Also, the standard value is 4.511, which might be due to the fact that we are using less data points.

3. Now fit a two-level hierarchical model to this data, of the following form:

$$
\begin{aligned}
(y_{ij} \mid \theta_i, \sigma^2) &\sim \mathrm{N}(\theta_i, \sigma^2) \\
(\theta_i \mid \tau^2, \sigma^2) &\sim \mathrm{N}(\mu + \beta x_i, \tau^2\sigma^2)
\end{aligned}
$$

   where y_{ij} is blood pressure measurement j on person i, and x_i is a dummy (0/1) variable indicating whether a patient received treatment 2, the experimental medication. Apply what you learned on the previous problem about sampling, hyperparameters, etc, but account for the extra wrinkle here, i.e. the presence of the βx_i term that shifts the mean between the treatment and control groups.

   Write our your model's complete conditional distributions, and fit it. Make a histogram of the posterior distribution for β, which represents the treatment effect here. In particular, what are the posterior mean and standard deviation of β? How do these compare to the estimates and standard errors from the approaches in (A) and (B)?

   Using the same priors as in the same exercise we get

$$
\begin{aligned}
y_{ij} \mid \theta_i, \sigma^2 &\sim \mathrm{N}(\theta_i, \sigma^2) \\
\theta_i \mid \tau^2, \sigma^2 &\sim \mathrm{N}(\mu + \beta x_i, \tau^2\sigma^2) \\
\tau &\sim \mathrm{InvGamma}(0.5, 0.5) \\
p(\mu) &\propto 1 \\
\beta &\sim \mathrm{N}(0, 1) \\
p(\sigma^2) &\propto 1/\sigma^2.
\end{aligned}
$$

   The full conditionals are derived in a similar way as in the last exercise. We have

$$
\begin{aligned}
p(\mu|\cdot) &\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{p}(\theta_i - \mu - \beta x_i)^2\right\} \\
&\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{p}(\mu^2 - 2\theta_i\mu + 2\mu\beta x_i)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\frac{\sum\theta_i - \beta\sum x_i}{\tau^2\sigma^2}\right)\mu + \left(\frac{p}{\tau^2\sigma^2}\right)\mu^2\right\} \\
&\equiv \mathrm{N}\left(\frac{\sum_{i=1}^{p}\theta_i - \beta\sum_{i=1}^{p}x_i}{p}, \frac{\tau^2\sigma^2}{p}\right)
\end{aligned}
$$

For τ^2,

$$\begin{aligned} p(\tau|\cdot) &\propto (\tau^2)^{-p/2} \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{p}(\theta_i-\mu-\beta x_i)^2\right\}(\tau^2)^{-1.5}\exp\left\{-\frac{1}{2\tau^2}\right\} \\ &\propto (\tau^2)^{-(p+1)/2-1}\exp\left\{-\frac{1}{\tau^2}\left(\frac{\sum_{i=1}^{p}(\theta_i-\mu-\beta x_i)^2}{2\sigma^2}+\frac{1}{2}\right)\right\} \\ &\equiv \text{InvGamma}\left(\frac{1+p}{2},\frac{\sum_{i=1}^{p}(\theta_i-\mu-\beta x_i)^2}{2\sigma^2}\right). \end{aligned}$$

For σ^2

$$\begin{aligned} p(\sigma^2|\cdot) &\propto \frac{1}{\sigma^2}\left(\sigma^2\right)^{-\frac{p}{2}}\exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{p}(\theta_i-\mu-\beta x_i)^2\right\}(\sigma^2)^{-\frac{p\times n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\theta_i)^2\right\} \\ &= \left(\sigma^2\right)^{-\frac{p}{2}-1-\frac{p\times n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\frac{1}{\tau^2}\sum_{i=1}^{p}(\theta_i-\mu-\beta x_i)^2\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\theta_i)^2\right]\right\} \\ &\equiv \text{InvGamma}\left(\frac{p(n+1)}{2},\frac{1}{\tau^2}\sum_{i=1}^{p}(\theta_i-\mu-\beta x_i)^2+\sum_{i=1}^{p}\sum_{j=1}^{n}(y_{ij}-\theta_i)^2\right). \end{aligned}$$

Finally, for β,

$$\begin{aligned} p(\beta|\cdot) &\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{p}(\theta_i-\mu-\beta x_i)^2\right\}\exp\left\{-\frac{1}{2}\beta^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{p}(\beta^2x_i^2+2\mu x_i\beta-2\theta_i x_i\beta)\right\}\exp\left\{-\frac{1}{2}\beta^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(-2\frac{\sum(\theta_i-\mu)x_i}{\tau^2\sigma^2}\beta+\beta^2\left(\frac{\sum x_i^2}{\tau^2\sigma^2}+1\right)\right)\right\} \\ &\equiv \text{N}\left(\left(\frac{\sum x_i^2}{\tau^2\sigma^2}+1\right)^{-1}\frac{\sum(\theta_i-\mu)x_i}{\tau^2\sigma^2},\quad \frac{\sum x_i^2}{\tau^2\sigma^2}+1\right). \end{aligned}$$

Finally note that for each θ_i we have a Normal-Normal model again, so

$$p(\theta_i|\cdot) \equiv \text{N}\left(\mu_\theta,\sigma_\theta\right),$$

where if n_i is the number of students in the school i

$$\mu_{\theta_i} = \frac{\frac{1}{\tau^2\sigma^2}(\mu+\beta x_i)+\frac{n_i}{\sigma^2}\bar{y}_i}{\frac{1}{\tau^2\sigma^2}+\frac{n_i}{\sigma^2}} \quad \text{and} \quad \sigma_n^2 = \frac{n_i}{\sigma^2}+\frac{1}{\sigma^2\tau^2}.$$

Implementing the algorithm in R we get the traceplots in figure 5. The chains seem to have converged. We also plot the histogram of beta after a burn in of 1,000. The posterior mean and standard deviation are -0.38 and 0.96, respectively.

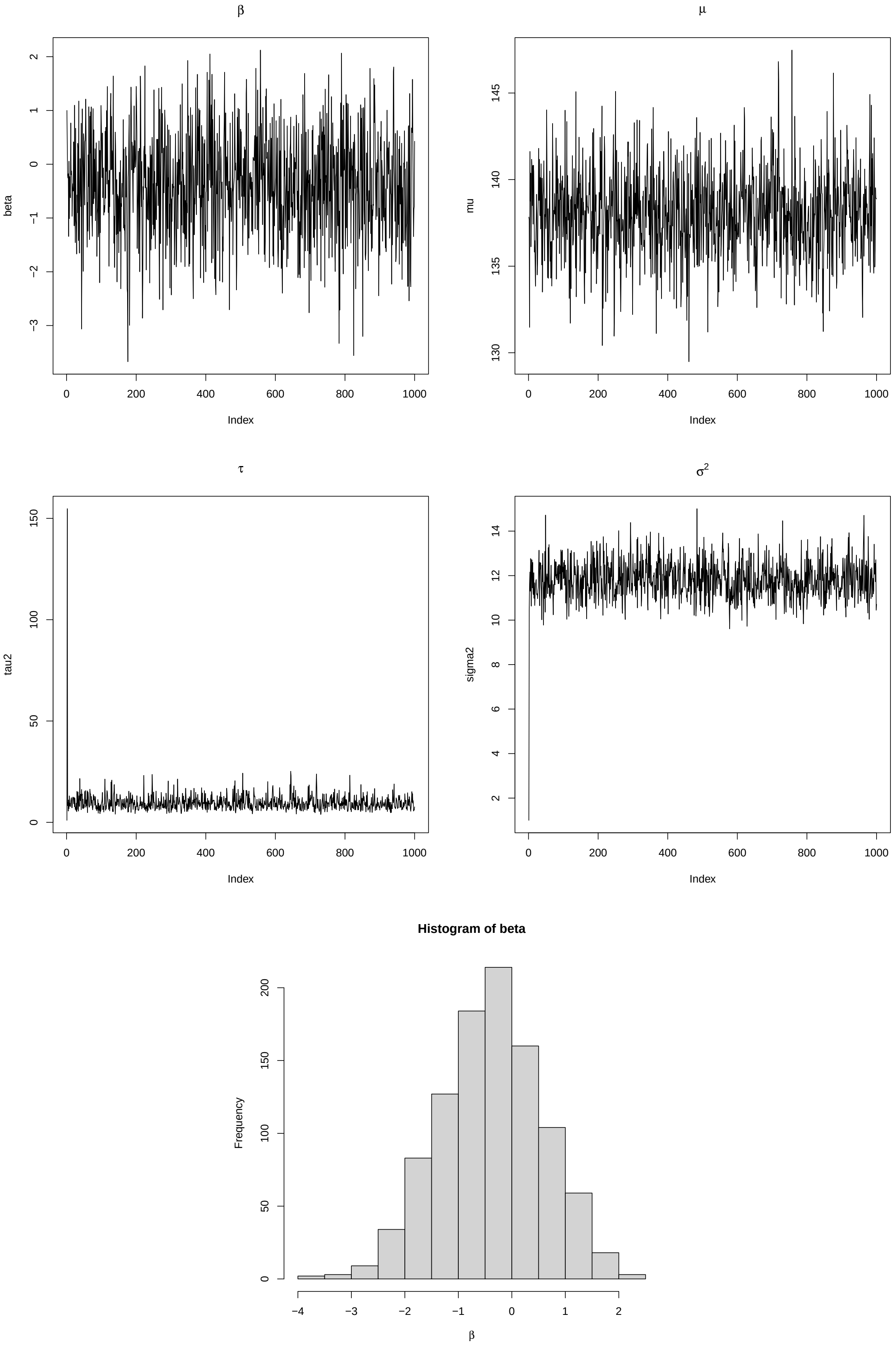


Figure 5

4. Your two-level model assumes that, conditional on θ_i, the y_{ij} are independent. Written concisely: $(y_{ij} \perp y_{ik} \mid \theta_i)$ for $j \neq k$. There are many ways this assumption could break down. So check! Does this assumption look (approximately) sensible in

light of the data? Provide evidence one way or another.

It looks plausible since comparing the ACFs of all the subjects (figure 6)show low to moderate autocorrelations, all less or equal than 0.5 in absolute value, so while it is not an unreasonable assumption; personally, I would still try to change the model, for instance, assuming an ARMA covariance structure for the likelihood.

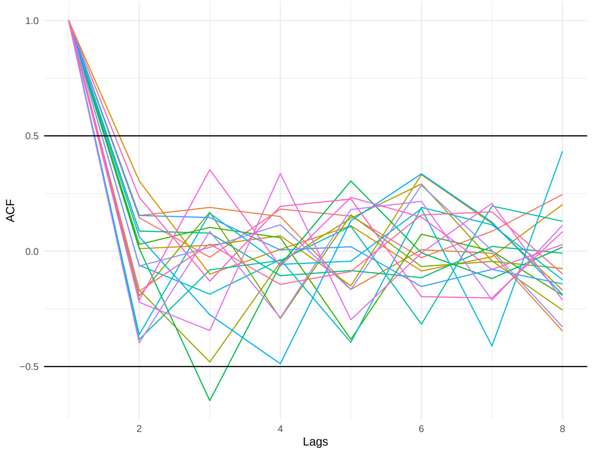


Figure 6: ACF plots of all the subjects in the study.

6 Exercise 5

6.1 Price elasticity of demand

The data in “cheese.csv” are about sales volume, price, and advertisting display activity for packages of Borden sliced “cheese.” The data are taken from Rossi, Allenby, and McCulloch’s textbook on *Bayesian Statistics and Marketing*. For each of 88 stores (store) in different US cities, we have repeated observations of the weekly sales volume (vol, in terms of packages sold), unit price (price), and whether the product was advertised with an in-store display during that week (disp = 1 for display).

Your goal is to estimate, on a store-by-store basis, the effect of display ads on the demand curve for cheese. A standard form of a demand curve in economics is of the form $Q = \alpha P^{\beta}$, where Q is quantity demanded (i.e. sales volume), P is price, and α and β are parameters to be estimated. You’ll notice that this is linear on a log-log scale,

$$\log Q = \log \alpha + \beta \log P$$

which you should feel free to assume here. Economists would refer to β as the price elasticity of demand (PED). Notice that on a log-log scale, the errors enter multiplicatively.

There are several things for you to consider in analyzing this data set.

- The demand curve might shift (different α) and also change shape (different β) depending on whether there is a display ad or not in the store.
- Different stores will have very different typical volumes, and your model should account for this.
- Do different stores have different PEDs? If so, do you really want to estimate a separate, unrelated β for each store?
- If there is an effect on the demand curve due to showing a display ad, does this effect differ store by store, or does it look relatively stable across stores?
- Once you build the best model you can using the log-log specification, do see you any evidence of major model mis-fit?

Propose an appropriate hierarchical model that allows you to address these issues, and use Gibbs sampling to fit your model.

We will use a mixture approach to this problem to take into account the intrinsic variability within each store. Let $j = 1, \ldots, 88$ index the stores and t the times at which the measurements were taken. Take

$$\log Q_{ti} \sim \mathrm{N}\left(\alpha_i + \log \alpha^* \mathbb{I}_{\text{disp}=1} + \beta^* \mathbb{I}_{\text{disp}=1} \log P_{it} + \beta_i \log P_{it}, \sigma_i^2\right) \quad \text{for all t,i.}$$

To use a Bayesian model we need to put appropriate priors over the parameters. For simplicity let us use

$$\begin{aligned} \beta = (\alpha_i, \alpha^*, \beta^*, \beta_i) &\sim \mathrm{N}\left(\mu, \mathrm{diag}(s_1, \ldots, s_4)\right) \\ \mu &\propto 1 \\ s_i &\sim \mathrm{InvGamma}(a_0/2, b_0/2) \\ \sigma_\beta^2 &\sim \mathrm{InvGamma}(1/2, 1/2) \end{aligned}$$

The explanation for this choice is simple. We included interactions between the intercept and the coefficient on P_i, and since hyperparameters are difficult to choose here we tried to use empirical Bayes the most possible.
Let us write now the full conditionals. We can take advantage of the conjugacy. Let X be the design matrix containing all the necessary variables, then

- $\beta_i|\cdot$.

  This is a simple Bayesian linear regression, so this full conditional must be

  $$p(\beta_i|\cdot) \equiv \mathrm{N}\left(\Sigma^{-1}b, \Sigma^{-1}\right),$$

  where $\Sigma^{-1} = X'X/\sigma_i^2 + \mathrm{diag}(s_1, \ldots, s_4)$ and $b = \Sigma^{-1}(\mathrm{diag}(s_1, \ldots, s_4)\mu + X' \log Q/\sigma_i^2)$
- $s_k|\cdot$. Again, this is the variance for a linear regression, so we get

  $$p(s_k|\cdot) \equiv \mathrm{InvGamma}\left(\frac{n+1}{2}, \frac{1}{2}\left(1 + \sum_{i=1}^{n}(\beta k - \mu_k)^2\right)\right).$$

- $\mu_i|\cdot$.

  Here we simply get

  $$p(\mu_i|\cdot) \propto \mathrm{N}\left(\bar{\beta}, \frac{1}{n}\mathrm{diag}(s_1, \ldots, s_4)\right).$$

- $\sigma_\beta^2|\cdot$.

  $$p(\sigma_\beta^2|\cdot) \equiv \mathrm{InvGamma}\left(\frac{1+n_i}{2}, \frac{1}{2}\left(1 + (\log Q_i - X\beta_i)'(\log Q_i - X\beta_i)\right)\right),$$

Let's code the model. We set $a, b = 2$ and run the Gibbs sampler for 20,000 iterations, using a burn-in of 5,000. There are too many traceplots to fit in this document, so we will skip them.

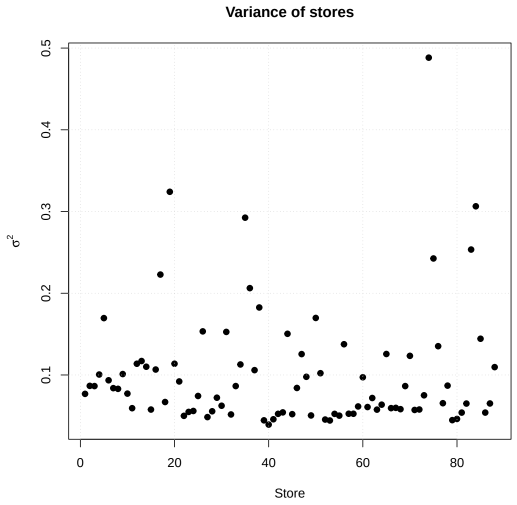


Figure 7: Estimated variances for each store

In figure 7 we show the estimated variances for each store. As we can discern the heteroskedisticity assumption is justified as several stores have values that do not match the average.

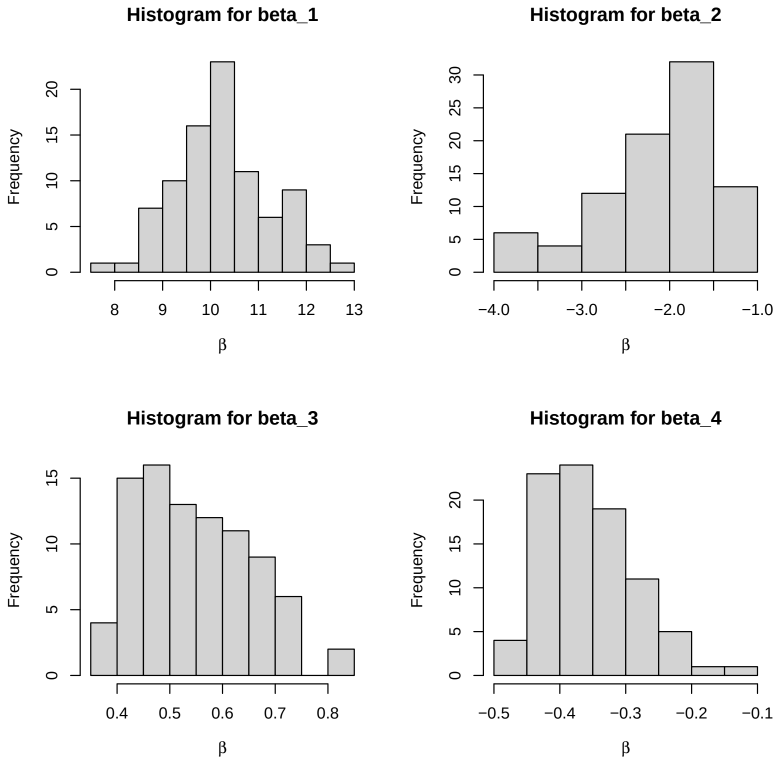


Figure 8: Histograms for each β.

Finally, on figure 8 we plotted the histograms of each β along each store, that is, estimated the regression coefficients for each store and plotted the histograms. Again. the variance seems to be wide enought to justify the inclusion of different coefficients for each store.
Also, note that β_3, the interaction between `disp` and the log-price has a very large variance and has positive support, which can serve as evidence that putting cheese into display indeed shifts positively the demand.

6.2 A hierarchical probit model

In "polls.csv" you will find the results of several political polls from the 1988 U.S. presidential election. The outcome of interest is whether someone plans to vote for George Bush (senior, not junior). There are several potentially relevant demographic predictors here, including the respondent's state of residence. The goal is to understand how these relate to the probability that someone will support Bush in the election. You can imagine this information would help a great deal in poll re-weighting and aggregation (ala Nate Silver).
Use Gibbs sampling, together with the Albert and Chib trick, to fit a hierarchical probit model of the following form:

$$\begin{aligned} \Pr(y_{ij} = 1) &= \Phi(z_{ij}) \\ z_{ij} &= \mu_i + x_{ij}^T \beta_i \,. \end{aligned}$$

Here y_{ij} is the response (Bush=1, other=0) for respondent j in state i; $\Phi(\cdot)$ is the probit link function, i.e. the CDF of the standard normal distribution; μ_i is a state-level intercept term; x_{ij} is a vector of respondent-level demographic predictors; and β_i is a vector of regression coefficients for state i.
Notes:

1. There are severe imbalances among the states in terms of numbers of survey respondents. Following the last problem, the key is to impose a hierarchical prior on the state-level parameters.

2. The data-augmentation trick from the Albert and Chib paper above is explained in many standard references on Bayesian analysis. If you want to get a quick introduction to the idea, you can consult one of these. A good presentation is in Section 8.1.1 of "Bayesian analysis for the social sciences" by Simon Jackman, available as an ebook through lib.utexas.edu.

3. You are welcome to use the logit model instead of the probit model. If you do this, you'll need to read the following paper, rather than Albert and Chib: Polson, N.G., Scott, J.G. and Windle, J. (2013). Bayesian inference for logistic models using Polya-Gamma latent variables. J. Amer. Statist. Assoc. 108 1339–1349. You can find a routine for simulation Polya-Gamma random variables in the BayesLogit R package and the pypolyagamma python library.

Let us use the model in Albert & Chib by considering conjugate hyperpriors on the hyperparameters.

$$\begin{aligned}
\mathbb{P}\left(Y_{ij} = 1 | Z_i\right) &= \Phi(x_{ij}'\beta_i) \\
Z_i | \lambda_i &\sim \mathrm{N}\left(x_i'\beta_i, \lambda_i^{-1}\right) \\
p(\beta_i) &\propto 1 \\
\lambda &\sim \mathrm{Gamma}\left(\nu/2, 2/\nu\right) \\
\nu &\sim \pi(\nu)
\end{aligned}$$

This yields the following full conditionals:

- $$\beta | \cdot \sim \mathrm{N}\left(x_i'\beta, \lambda_i^{-1}\right)$$
- $$\lambda_i | \cdot \sim \mathrm{Gamma}\left(\frac{\nu+1}{2}, \frac{2}{\nu + (Z_i - x_i'\beta_i)^2}\right)$$
- $$p(\nu|\cdot) \propto \pi(\nu) \prod_{i=1}^{n} \prod_{j=1}^{n_i} (c(\nu)\, \lambda_i^{\nu/2-1} e^{-\nu\lambda_i/2}),$$

  where $c(\nu) = \left(\Gamma(\nu/2)(\nu/2)^{\nu/2}\right)^{-1}$
- $$Z_i \sim \begin{cases} \mathrm{N}\left(z_i | x_i'\beta, \lambda_i^{-1}\right) \mathbb{I}(z_i \geq 0) & \text{if } Y_i = 1 \\ \mathrm{N}\left(z_i | x_i'\beta, \lambda_i^{-1}\right) \mathbb{I}(z_i \leq 0) & \text{if } Y_i = 0 \end{cases} \tag{2}$$

We will implement the associated Gibbs sampler by sampling from (2) on a discrete grid `seq(1,10,by=0.1)`, which makes it easy to implement.
We used 20,000 iterations with a burn in of 2,000. In figure 9 we show the estimated probability per state, which seems to align with intuition.

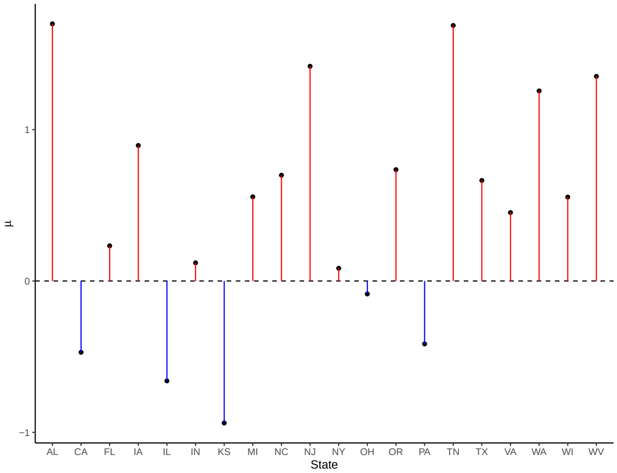


Figure 9

In figure 10 we show the estimated coefficients for each covariate. Note that the second beta, corresponding to the change in voting propensity for black Americans we can see how they tend to vote more Democrat, an interesting observation.

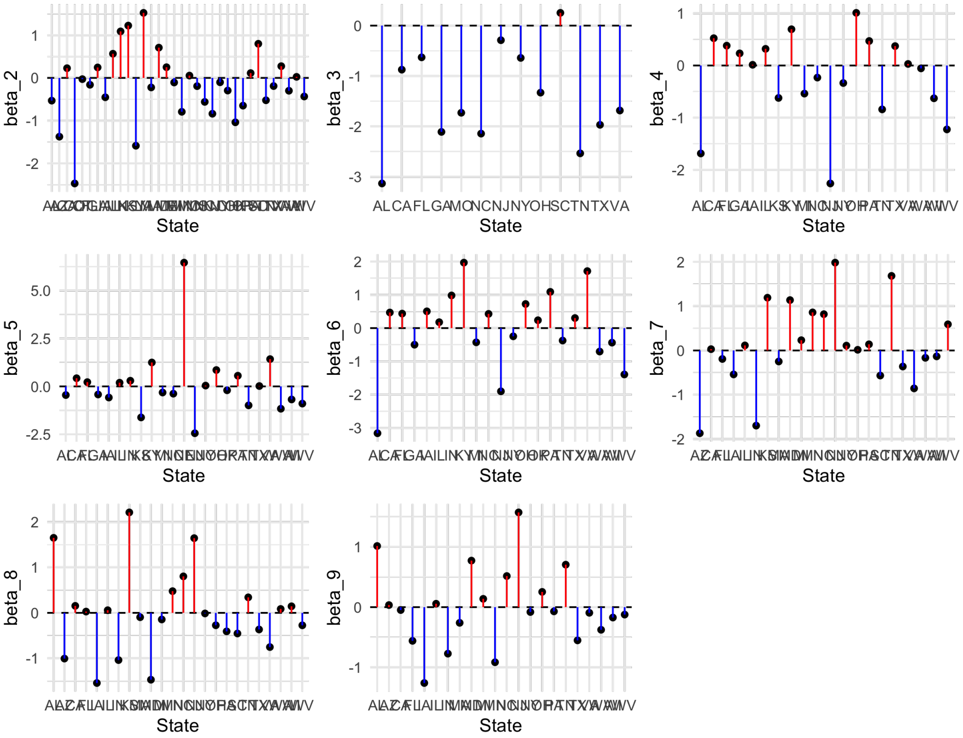


Figure 10

7 Curve fitting by linear smoothing

Consider a nonlinear regression problem with one predictor and one response: $y_i = f(x_i) + \epsilon_i$, where the ϵ_i are mean-zero random variables.

1. Suppose we want to estimate the value of the regression function $y^\star$ at some new point $x^\star$, denoted $\hat{f}(x^\star)$. Assume for the moment that $f(x)$ is linear, and that y and x have already had their means subtracted, in which case $y_i = \beta x_i + \epsilon_i$.

   Return to your least-squares estimator for multiple regression. Show that for the one-predictor case, your prediction $\hat{y}^\star = f(x^\star) = \hat{\beta} x^\star$ may be expressed as a *linear smoother*[7] of the following form:

$$\hat{f}(x^\star) = \sum_{i=1}^{n} w(x_i, x^\star) y_i$$

   for any $x^\star$. Inspect the weighting function you derived. Briefly describe your understanding of how the resulting smoother behaves, compared with the smoother that arises from an alternate form of the weight function $w(x_i, x^\star)$:

$$w_K(x_i, x^\star) = \begin{cases} 1/K, & x_i \text{ one of the } K \text{ closest sample points to } x^\star\,, \\ 0, & \text{otherwise.} \end{cases}$$

   This is referred to as *K-nearest-neighbor smoothing*.

   Solution.

   Remember the normal equations

$$(X'X)\beta = X'Y$$

   Taking X to be just a one dimensional column vector we get

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

   Hence

$$\hat{y}^* = \hat{\beta} x^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} x^* = \sum_{i=1}^{n} \frac{x_i x^*}{\sum_{j=1}^{n} x_j^2} y_i,$$

[7] This doesn't mean the weight function $w(\cdot)$ is linear in its arguments, but rather than the new prediction is a linear combination of the past outcomes y_i.

so take

$$w(x_i, x^*) = \frac{x_i x^*}{\sum_{j=1}^n x_j^2}$$

This essentially weights y by projecting x^* onto each x_i; in contrast with knn, where it scales it uniformly considering only the nearest vector instead of weighting with all of them.

2. A *kernel function* $K(x)$ is a smooth function satisyfing

$$\int_{\mathbb{R}} K(x)dx = 1\,, \quad \int_{\mathbb{R}} xK(x)dx = 0\,, \quad \int_{\mathbb{R}} x^2K(x)dx > 0\,.$$

A very simple example is the uniform kernel,

$$K(x) = \frac{1}{2}I(x) \quad \text{where} \quad I(x) = \begin{cases} 1, & |x| \leq 1 \\ 0, & \text{otherwise}\,. \end{cases}$$

Another common example is the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,.$$

Kernels are used as weighting functions for taking local averages. Specifically, define the weighting function

$$w(x_i, x^\star) = \frac{1}{h} K\left(\frac{x_i - x^\star}{h}\right)\,,$$

where h is the bandwidth. Using this weighting function in a linear smoother is called *kernel regression*. (The weighting function gives the unnormalized weights; you should normalize the weights so that they sum to 1.)

Write your own R function that will fit a kernel smoother for an arbitrary set of x-y pairs and arbitrary choice of (positive real) bandwidth h.[8] You choose the kernel. Set up an R script that will simulate noisy data from some nonlinear function, $y = f(x) + \epsilon$; subtract the sample means from the simulated x and y; and use your function to fit the kernel smoother for some choice of h. Plot the estimated functions for a range of bandwidths wide enough to yield noticeable differences in the qualitative behavior of the prediction functions.

We will use the Gaussian kernel (because it is easier) to try to approximate the function

$$f(x) = x + \cos x\,,$$

and used the bandwidths $h = 0.1, 0.5, 1$ and 2. The results are shown in figure 11. We can tell that as the bandwidth grows smaller we get better results.

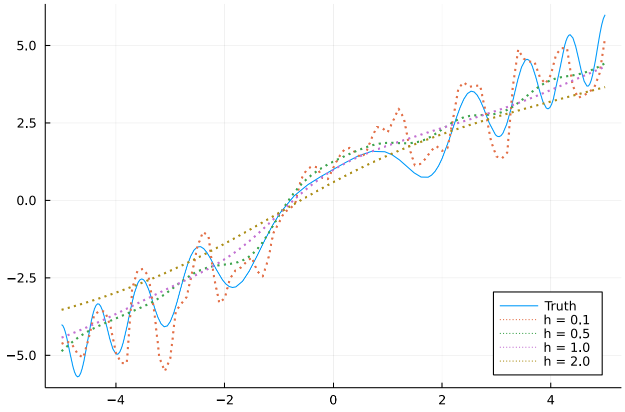


Figure 11: Kernel smoothing for $x + \cos x$.

[8] Coding tip: write things in a modular way. A kernel function is a function accepting a distance and a bandwidth and returning a nonnegative real. A weighting function is a function that accepts a vector of previous x's, a new x, and a kernel function; and that returns a vector of weights. Et cetera. It's much, much easier to debug modular code.

8 Cross validation

Left unanswered so far in our previous study of kernel regression is the question: how does one choose the bandwidth h used for the kernel? Assume for now that the goal is to predict well, not necessarily to recover the truth. (These are related but distinct goals.)

1. Presumably a good choice of h would be one that led to smaller predictive errors on fresh data. Write a function or script that will: (1) accept an old ("training") data set and a new ("testing") data set as inputs; (2) fit the kernel-regression estimator to the training data for specified choices of h; and (3) return the estimated functions and the realized prediction error on the testing data for each value of h. This should involve a fairly straightforward "wrapper" of the function you've already written.

   After implementing it for a grid of `h=0.1:0.05:4` we obtained that the optimum is given at $h = 0.15$. The results are shown in figure 12.

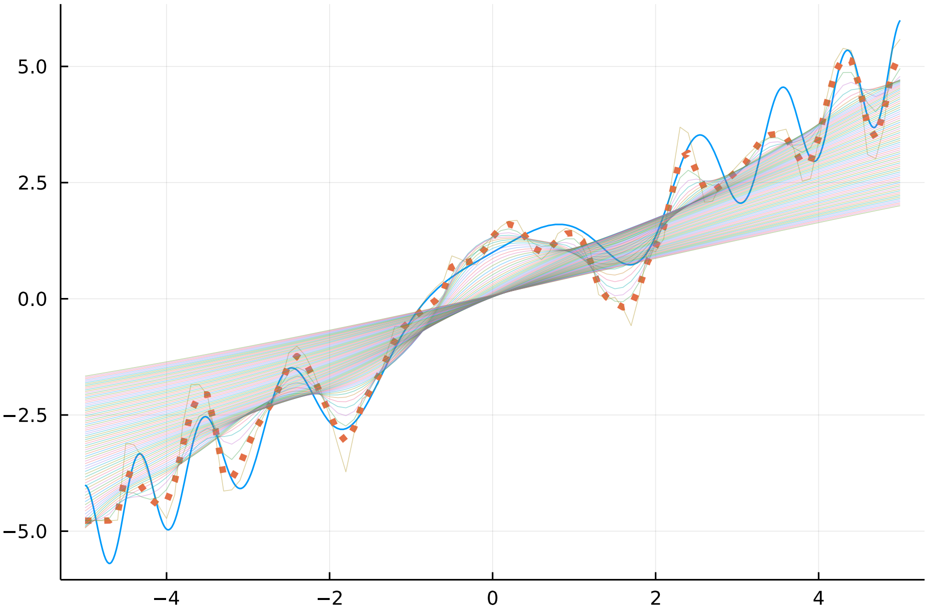


Figure 12: The truth is the solid line and the pointed line corresponds to the best approximation.

2. Imagine a conceptual two-by-two table for the unknown, true state of affairs. The rows of the table are "wiggly function" and "smooth function," and the columns are "highly noisy observations" and "not so noisy observations." Simulate one data set (say, 500 points) for each of the four cells of this table, where the x's take values in the unit interval. Then split each data set into training and testing subsets. You choose the functions.[9] Apply your method to each case, using the testing data to select a bandwidth parameter. Choose the estimate that minimizes the average squared error in prediction, which estimates the mean-squared error:

$$L_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n^\star}(y_i^\star - \hat{y}_i^\star)^2 ,$$

   where $(y_i^\star, x_i^\star)$ are the points in the test set, and $\hat{y}_i^\star$ is your predicted value arising from the model you fit using only the training data. Does your out-of-sample predictive validation method lead to reasonable choices of h for each case?

   We kept the previous functions and changed the variance of the normal noise, obtaining

$\cos(x) + x + \mathrm{N}(0, 0.3)$	$\cos(x) + x + \mathrm{N}(0, 3)$
$\cos(5x) + x + \mathrm{N}(0, 0.3)$	$\cos(5x) + x + \mathrm{N}(0, 3)$

   The results are shown in figure 13. There are two observations to be made here: higher variance results in a flatter estimation, since the point cloud becomes more uniform; moreover, when the function becomes «wigglier» the data concentrates more and so the estimation becomes more stable.

[9] Trigonometric functions, for example, can be pretty wiggly if you make the period small.

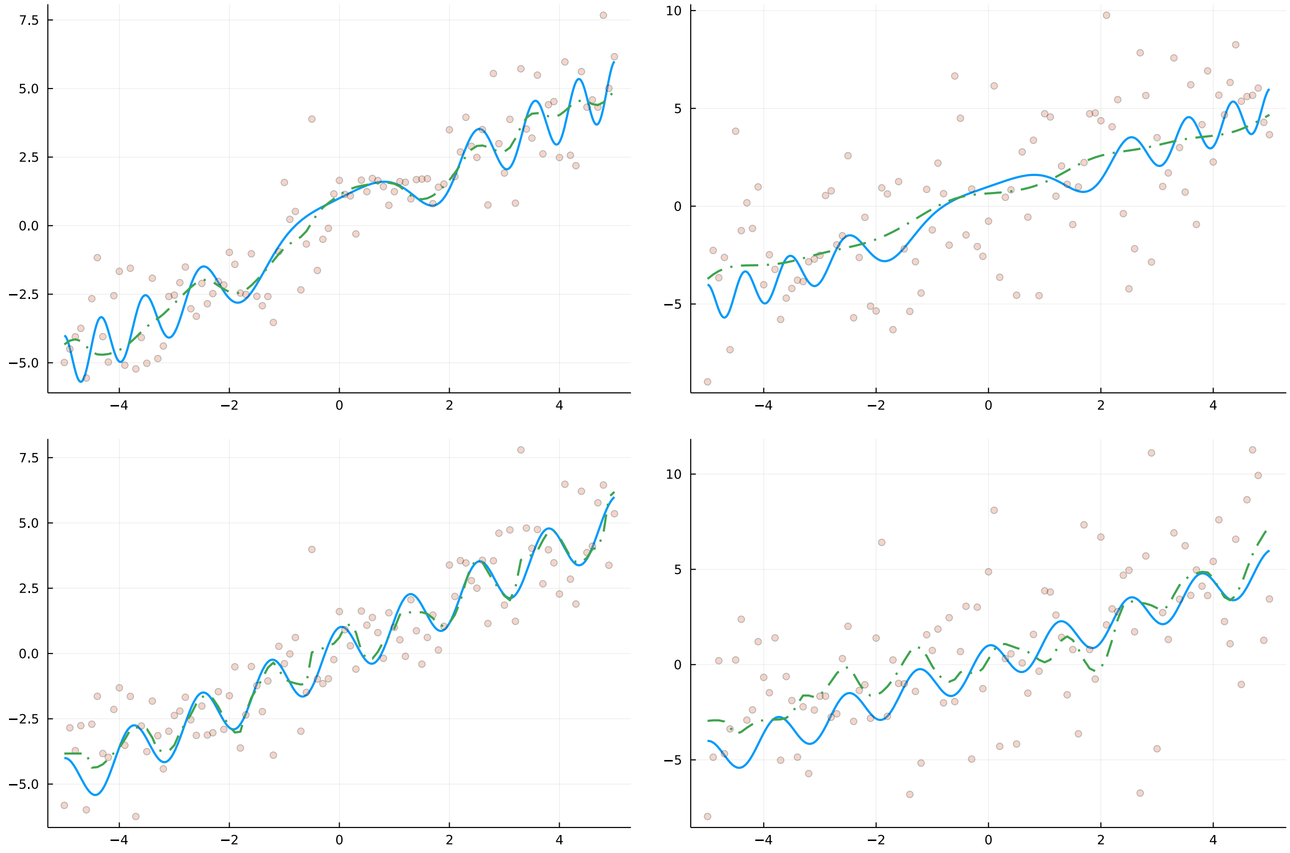


Figure 13

3. Use the leave-one-out lemma to revisit the examples you simulated in Part B, using leave-one-out cross validation to select h in each case. Because of the leave-one-out lemma, you won't need to actually refit the model N times!

CV	0.25	0.65	0.1	0.2
LOO	0.1	0.65	0.1	0.2

Table 3: Optimal hs for both models.

After using the hint we plotted the best models, shown in figure 14. The results are similar, which is reasonable considering the bandwithds, shown in table **??**, were similar.

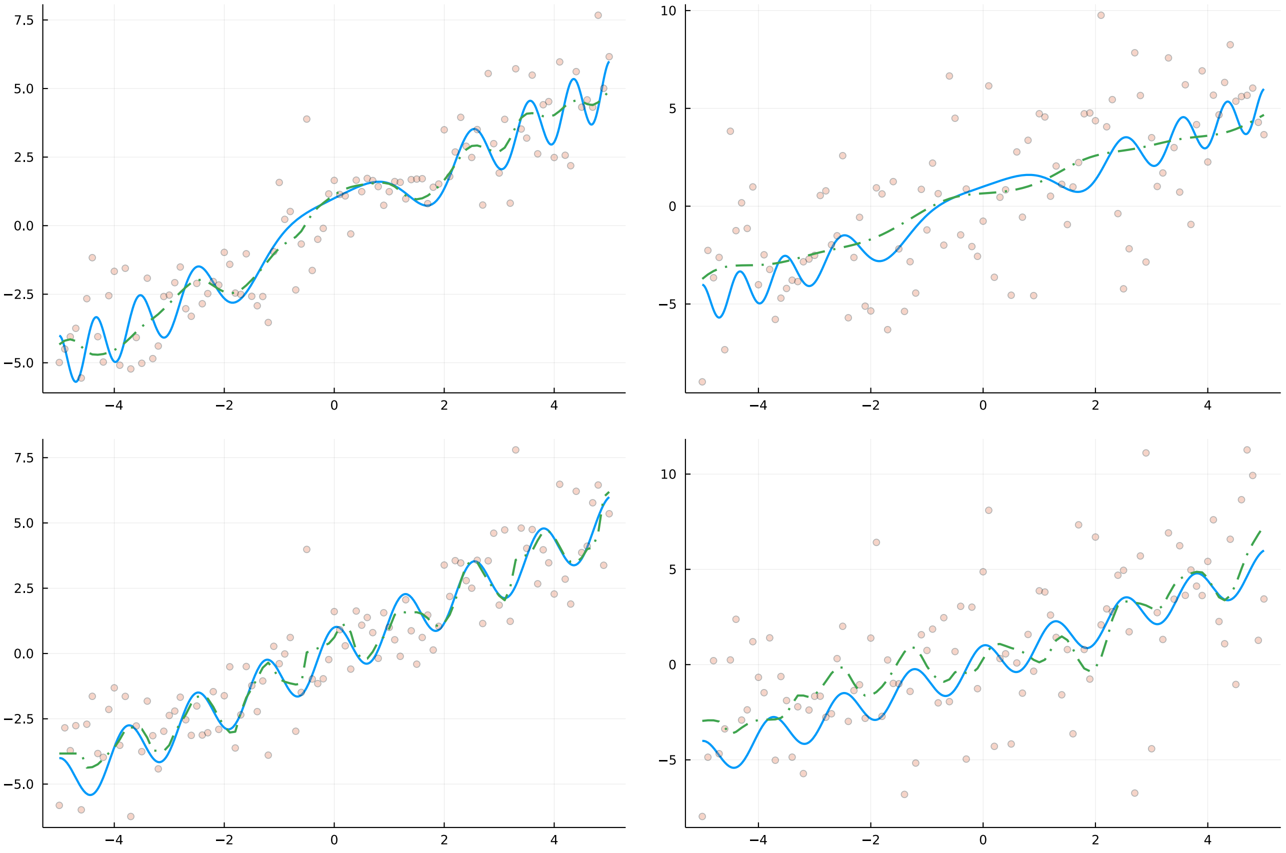


Figure 14: Idem.

9 Local polynomial regression

1. A natural generalization of locally constant regression is local polynomial regression. For points u in a neighborhood of the target point x, define the polynomial

$$g_x(u;a) = a_0 + \sum_{k=1}^{D} a_j (u-x)^k$$

for some vector of coefficients $a = (a_0, \dots, a_D)$. As above, we will estimate the coefficients a in $g_x(u;a)$ at some target point x using weighted least squares:

$$\hat{a} = \arg\min_{R^{D+1}} \sum_{i=1}^{n} w_i \left\{ y_i - g_x(x_i;a) \right\}^2 \,,$$

where $w_i \equiv w(x_i, x)$ are the kernel weights defined just above, normalized to sum to one.[10] Derive a concise (matrix) form of the weight vector $\hat{a}$, and by extension, the local function estimate $\hat{f}(x)$ at the target value x.[11] Life will be easier if you define the matrix R_x whose (i,j) entry is $(x_i - x)^{j-1}$, and remember that (weighted) polynomial regression is the same thing as (weighted) linear regression with a polynomial basis.

Note: if you get bogged down doing the general polynomial case, just try the linear case.

Let us express the sum as a quadratic form. Using the R_s matrix we get

$$\begin{aligned} R_x a &= \begin{pmatrix} a_0 + a_1(x_1 - x) + \cdots + a_D(x_1 - x)^D \\ \vdots \\ a_0 + a_1(x_n - x) + \cdots + a_D(x_n - x)^D \end{pmatrix} \\ &= \begin{pmatrix} g_x(x_1|a) \\ \vdots \\ g_x(x_n|a) \end{pmatrix}, \end{aligned}$$

[10] We are fitting a different polynomial function for every possible choice of x. Thus $\hat{a}$ depends on the target point x, but we have suppressed this dependence for notational ease.

[11] Observe that at the target point x, $g_x(u = x; a) = a_0$. That is, only the constant term appears. But this is not the same thing as fitting only a constant term!

so

$$\sum_{i=1}^n w_i\{y_i - g_x(x_i;a)\}^2 = (y - R_x a)'\mathrm{diag}(w)(y - R_x a)$$

To get $\hat{a}$ we differentiate f, getting

$$\begin{aligned}\frac{\partial}{\partial a}(y - R_x a)'\mathrm{diag}(w)(y - R_x a) &= -(y'\mathrm{diag}(w)R_x)' - R_x'\mathrm{diag}(w)y + 2R_x'\mathrm{diag}(w)R_x a \\ &= -2R_x'\mathrm{diag}(w)y + 2R_x'\mathrm{diag}(w)R_x a.\end{aligned}$$

Setting it to zero

$$R_x'\mathrm{diag}(w)R_x a = R_x'\mathrm{diag}(w)y,$$

so

$$\hat{a} = \left(R_x'\mathrm{diag}(w)R_x\right)^{-1} R_x'\mathrm{diag}(w)y,$$

and since the local approximation is evaluated at $x_i = x$ we get

$$\hat{f}(x) = \begin{pmatrix}1 & 0 & 0 & \cdots & 0\end{pmatrix}\hat{a}.$$

2. From this, conclude that for the special case of the local linear estimator ($D = 1$), we can write $\hat{f}(x)$ as a linear smoother of the form

$$\hat{f}(x) = \frac{\sum_{i=1}^n w_i(x)y_i}{\sum_{i=1}^n w_i(x)},$$

where the unnormalized weights are

$$\begin{aligned}w_i(x) &= K\left(\frac{x - x_i}{h}\right)\{s_2(x) - (x_i - x)s_1(x)\} \\ s_j(x) &= \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)(x_i - x)^j.\end{aligned}$$

Note that in this case

$$\hat{f}(x) = \begin{pmatrix}1 & 0\end{pmatrix}\hat{a}.$$

Then we can express

$$\begin{aligned}R_x'\mathrm{diag}(w)R_x &= \begin{pmatrix}1 & \cdots & 1 \\ x_1 - x & \cdots & x_n - x\end{pmatrix}\begin{pmatrix}w_1 & \cdots & 0 \\ \vdots & w_i & \vdots \\ 0 & \cdots & w_n\end{pmatrix}\begin{pmatrix}1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x\end{pmatrix} \\ &= \begin{pmatrix}w_i & \cdots & w_n \\ w_1(x_1 - x) & \cdots & w_n(x_n - x)\end{pmatrix}\begin{pmatrix}1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x\end{pmatrix} \\ &= \begin{pmatrix}\sum w_i & \sum w_i(x_i - x) \\ \sum w_i(x_i - x) & \sum w_i(x_i - x)^2\end{pmatrix}\end{aligned}$$

The inverse is

$$\left(R_x'\mathrm{diag}(w)R_x\right)^{-1} = \frac{1}{\det}\begin{pmatrix}\sum w_i(x_i - x)^2 & -\sum w_i(x_i - x) \\ -\sum w_i(x_i - x) & \sum w_i\end{pmatrix},$$

where

$$\begin{aligned}\det &= \sum_{i=1}^n w_i(x_i - x)^2 - \left(\sum_{i=1}^n w_i(x_i - x)\right)^2 \\ &= \underbrace{\sum_{i=!}^n K(\cdot)(x_i - x)^2}_{s^2} - \underbrace{\left(\sum_{i=1}^n K(\cdot)(x_i - x)\right)^2}_{m^2}\end{aligned}$$

Thus

$$\begin{aligned}
\hat{f}(x) &= \frac{1}{\det}\begin{pmatrix} \sum w_i(x_i-x)^2 & -\sum w_i(x_i-x) \\ -\sum w_i(x_i-x) & \sum w_i \end{pmatrix}\begin{pmatrix} \sum K(\cdot)y_i \\ \sum K(\cdot)(x_i-x)y_i \end{pmatrix} \\
&= \frac{s^2\sum K(\cdot)y_i - m\sum K(\cdot)(x_i-x)y_i}{s^2-m^2} \\
&= \frac{\sum w_i(x)y_i}{\sum w_i(x)}
\end{aligned}$$

3. Suppose that the residuals have constant variance σ^2 (that is, the spread of the residuals does not depend on x). Derive the mean and variance of the sampling distribution for the local polynomial estimate $\hat{f}(x)$ at some arbitrary point x. Note: the random variable $\hat{f}(x)$ is just a scalar quantity at x, not the whole function.

   Write $\hat{f}(x_i) = g_x(x_i|\hat{a})$, so

$$\begin{aligned}
\mathbb{E}\left[\hat{f}(x_i)\right] &= \mathbb{E}\left[R_{x_i}\hat{a}\right] \\
&= R_{x_i}\mathbb{E}\left[\hat{a}\right] \\
&= R_{x_i}\mathbb{E}\left[\left(R'_x\text{diag}(w)R_x\right)^{-1}R'_x\text{diag}(w)y\right] \\
&= R_{x_i}\left(R'_x\text{diag}(w)R_x\right)^{-1}R'_x\text{diag}(w)\mathbb{E}\left[y\right]
\end{aligned}$$

   For the variance we have

$$\begin{aligned}
\text{Var}\left(\hat{f}(x_i)\right) &= \text{Var}\left(R_{x_i}\hat{a}\right) \\
&= R_{x_i}\text{Var}\left(\left(R'_x\text{diag}(w)R_x\right)^{-1}R'_x\text{diag}(w)y\right)R'_{x_i} \\
&= \sigma^2 R_{x_i}\left(R'_x\text{diag}(w)R_x\right)^{-1}R'_x\text{diag}(w^2)R_{x_i}\left(R'_x\text{diag}(w)R_x\right)^{-1}R'_{x_i}.
\end{aligned}$$

4. We don't know the residual variance, but we can estimate it. A basic fact is that if x is a random vector with mean μ and covariance matrix Σ, then for any symmetric matrix Q of appropriate dimension, the quadratic form x^TQx has expectation

$$E(x^TQx) = \text{tr}(Q\Sigma) + \mu^TQ\mu\,.$$

   Write the vector of residuals as $r = y - \hat{y} = y - Hy$, where H is the smoothing matrix. Compute the expected value of the estimator

$$\hat{\sigma}^2 = \frac{\|r\|_2^2}{n - 2\text{tr}(H) + \text{tr}(H^TH)}\,,$$

   and simplify things as much as possible. Roughly under what circumstances will this estimator be nearly unbiased for large n? Note: the quantity $2\text{tr}(H) - \text{tr}(H^TH)$ is often referred to as the "effective degrees of freedom" in such problems.

   Noting that $\mathbb{E}\left[y_i\right] = g_x(x_i|\hat{a})$,

$$\begin{aligned}
\mathbb{E}\left[\|r\|_w^2\right] &= \mathbb{E}\left[r'r\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[r_i^2\right] \\
&= \sum_{i=1}^{n}\mathbb{E}\left[(y_i-\hat{y}_i)^2\right] \\
&= \sum_{i=1}^{n}\left(\mathbb{E}\left[(y_i^2\right] - 2H_i\mathbb{E}\left[y_i)^2\right]\right) \\
&= \sum_{i=1}^{n}\left(\text{Var}\left(y_i\right) + \mathbb{E}\left[(y_i\right]^2 - 2H_i\mathbb{E}\left[y_i)\right]\right) \\
&= n\sigma^2 + \sum_{i=1}^{n}g_x(x_i|\hat{a})\left[g_x(x_i|\hat{a}) - 2H_i\right]
\end{aligned}$$

   Hence

$$\mathbb{E}\left[\hat{\sigma^2}\right])\frac{n\sigma^2 + \sum g_x(x_i|\hat{a})\left[g_x(x_i|\hat{a}) - 2H_i\right]}{n - 2\text{tr}(H) + \text{tr}(H'H)}.$$

5. Write code that fits the local linear estimator using a Gaussian kernel for a specified choice of bandwidth h. Then load the data in "utilities.csv" into R.[12] This data set shows the monthly gas bill (in dollars) for a single-family home in Minnesota, along with the average temperature in that month (in degrees F), and the number of billing days in that month. Let y be the average daily gas bill in a given month (i.e. dollars divided by billing days), and let x be the average temperature. Fit y versus x using local linear regression and some choice of kernel. Choose a bandwidth by leave-one-out cross-validation.

   The chosen bandwith was $h = 6$. The estimated curve is shown in figure 17.

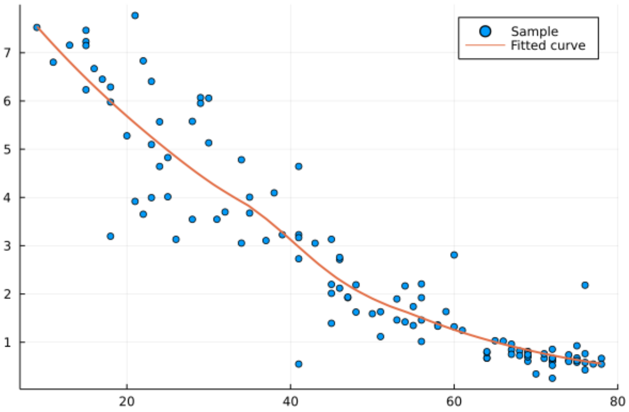


Figure 15: Fitted local linear estimator.

6. Inspect the residuals from the model you just fit. Does the assumption of constant variance (homoskedasticity) look reasonable? If not, do you have any suggestion for fixing it?

   The plot is shown in figure 16. The variance decreases as the temperature increases, so the homoskedasticity assumption is not reasonable.

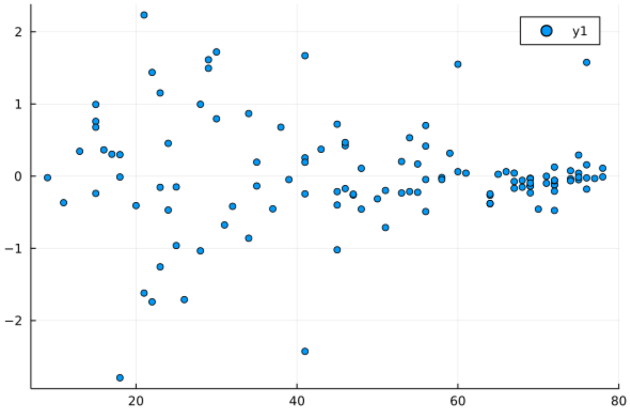


Figure 16: Residuals for the fitted local linear estimator.

7. Put everything together to construct an approximate point-wise 95% confidence interval for the local linear model (using your chosen bandwidth) for the value of the function at each of the observed points x_i for the utilities data. Plot these confidence bands, along with the estimated function, on top of a scatter plot of the data.[13]

   The bands appear mostly constant along the temperature range, which can be explained by a bad approximation using the Gaussian intervals.

[12] On the class GitHub site.

[13] It's fine to use Gaussian critical values for your confidence set.

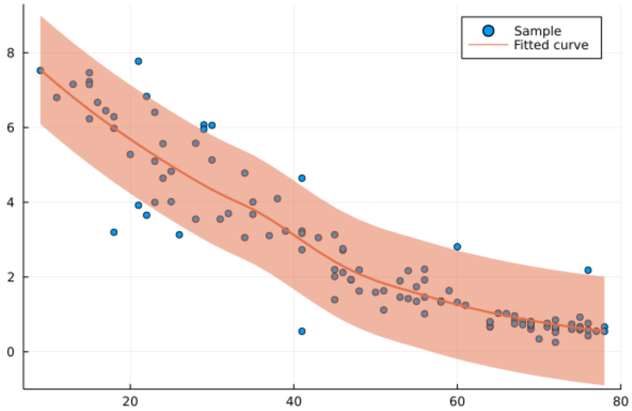


Figure 17: Fitted local linear estimator together with its confidence bands.

10 Gaussian processes

A *Gaussian process* is a collection of random variables $\{f(x) : x \in \mathcal{X}\}$ such that, for any finite collection of indices $x_1, \dots, x_N \in \mathcal{X}$, the random vector $[f(x_1), \dots, f(x_N)]^T$ has a multivariate normal distribution. It is a generalization of the multivariate normal distribution to infinite-dimensional spaces. The set $\mathcal{X}$ is called the index set or the state space of the process, and need not be countable.

A Gaussian process can be thought of as a random function defined over $\mathcal{X}$, often the real line or $\mathbb{R}^p$. We write $f \sim \mathrm{GP}(m, C)$ for some mean function $m : \mathcal{X} \to \mathbb{R}$ and a covariance function $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. These functions define the moments[14] of all finite-dimensional marginals of the process, in the sense that

$$E\{f(x_1)\} = m(x_1) \quad \text{and} \quad \mathrm{cov}\{f(x_1), f(x_2)\} = C(x_1, x_2)$$

for all $x_1, x_2 \in \mathcal{X}$. More generally, the random vector $[f(x_1), \dots, f(x_N)]^T$ has covariance matrix whose (i, j) element is $C(x_i, x_j)$. Typical covariance functions are those that decay as a function of increasing distance between points x_1 and x_2. The notion is that $f(x_1)$ and $f(x_2)$ will have high covariance when x_1 and x_2 are close to each other.

(a) Let's start with the simple case where $\mathcal{X} = [0, 1]$, the unit interval. Write a function that simulates a mean-zero Gaussian process on $[0, 1]$ under the squared exponential covariance function. The function will accept as arguments: (1) finite set of points $x_1, \dots, x_N$ on the unit interval; and (2) a triplet (b, τ_1^2, τ_2^2). It will return the value of the random process at each point: $f(x_1), \dots, f(x_N)$.

Use your function to simulate (and plot) Gaussian processes across a range of values for b, τ_1^2, and τ_2^2. Try starting with a very small value of τ_2^2 (say, 10^{-6}) and playing around with the other two first. On the basis of your experiments, describe the role of these three hyperparameters in controlling the overall behavior of the random functions that result. What happens when you try $\tau_2^2 = 0$? Why? If you can fix this, do—remember our earlier discussion on different ways to simulate the MVN.

Now simulating a few functions with a different covariance function, the Matérn with parameter 5/2:

$$C_{M52}(x_1, x_2) = \tau_1^2 \left\{1 + \frac{\sqrt{5}d}{b} + \frac{5d^2}{3b^2}\right\} \exp\left\{\right\}(\left.\frac{-\sqrt{5}d}{b}\right) + \tau_2^2 \delta(x_1, x_2),$$

where $d = \|x_1 - x_2\|_2$ is the distance between the two points x_1 and x_2. Comment on the differences between the functions generated from the two covariance kernels.

We used a equispaced grid in τ_1 from 0 to 2; b was selected from 0.001 to 1, using six values.

[14] And therefore the entire distribution, because it is normal

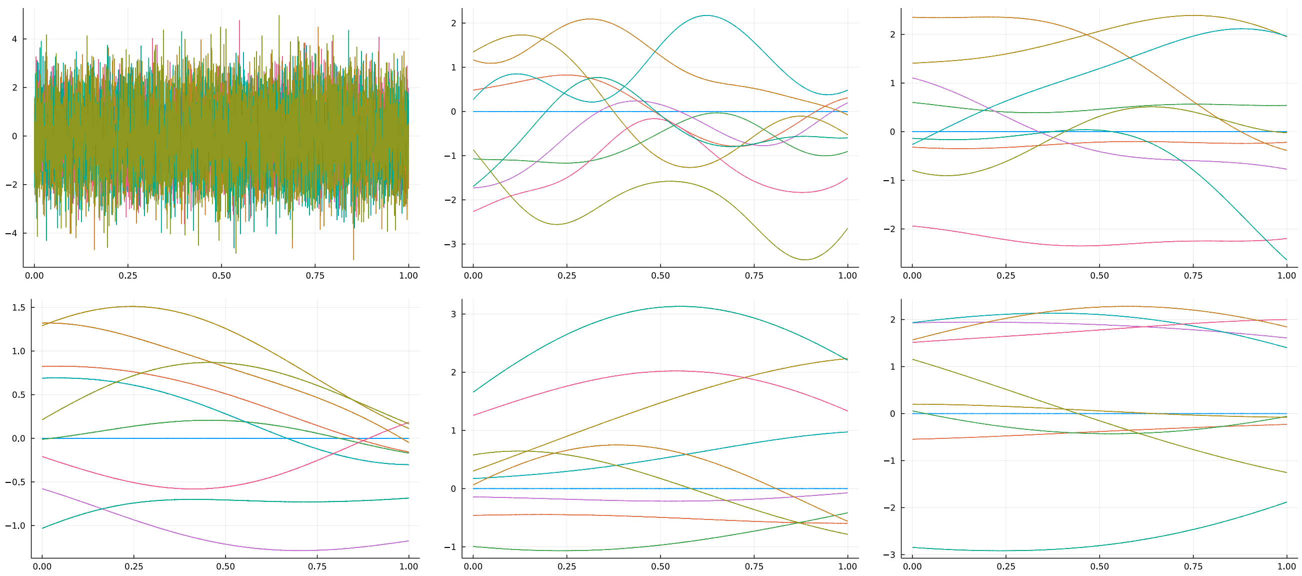


Figure 18: Trajectories for the first Matérn kernel.

The trajectories become smoother as τ_1 increases. The second Matérn kernel also produces more variation on the paths than the first ones.

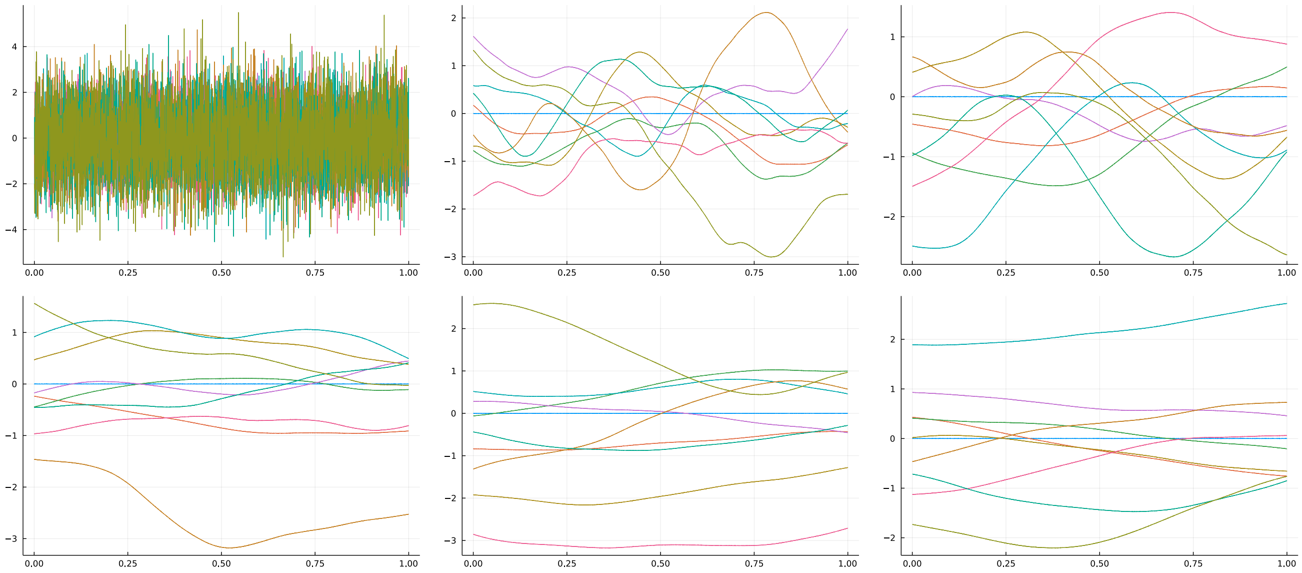


Figure 19: Trajectories for the second Matérn kernel.

(b) Suppose you observe the value of a Gaussian process $f \sim \mathrm{GP}(m, C)$ at points $x_1, \ldots, x_N$. What is the conditional distribution of the value of the process at some new point $x^\star$? For the sake of notational ease simply write the value of the (i, j) element of the covariance matrix as $C_{i,j}$, rather than expanding it in terms of a specific covariance function.

Suppose that we are given a new data point x^*. To predict $f(x^*)$ we start from the joint of $(f(x_1), \ldots, f(x_N))$ and $f(x^*)$, say

$$\begin{pmatrix} f(x^*) \\ f(\mathbf{x}) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} m(x^*) \\ m(\mathbf{x}) \end{pmatrix}, \begin{pmatrix} C^* & \tilde{C}' \\ \tilde{C} & C \end{pmatrix} \right).$$

where.

$$C = C(\mathbf{x}, \mathbf{x})$$
$$\tilde{C} = C(\mathbf{x}, x^*)$$
$$C^* = C(x^*.x^*).$$

Then following the partition lemma we obtain the conditional of $f(x^*)|f, \mathbf{x}$:

$$f(x^*)|f, \mathbf{x} \sim \mathrm{N}\left(m(x^*) + \tilde{C}'C^{-1}(\tilde{y} - m(\tilde{x})), C^* - \tilde{C}'C^{-1}\tilde{C}\right)$$

(c) Prove the following lemma.

Lemma

Suppose that the joint distribution of two vectors y and θ has the following properties: (1) the conditional distribution for y given θ is multivariate normal, $(y \mid \theta) \sim N(R\theta, \Sigma)$; and (2) the marginal distribution of θ is multivariate normal, $\theta \sim N(m, V)$. Assume that R, Σ, m, and V are all constants. Then the joint distribution of y and θ is multivariate normal.

Proof. The joint can be obtained as

$$\begin{aligned} p(\theta, y) &= p(y|\theta)p(\theta) \\ &\propto \exp\left\{-\frac{1}{2}\left(y - R\theta)'\Sigma^{-1}(y - R\theta) + (\theta - m)'V^{-1}(\theta - m)\right)\right\} \end{aligned}$$

We can expand the kernel as (proportional to)

$$y'\Sigma^{-1}y - 2y'\Sigma^{-1}R\theta + \theta'R'\Sigma^{-1}R\theta + \theta'V^{-1}\theta - 2m'V^{-1}\theta.$$

Using the usual trick of completing the square to get a quadratic form we get

$$p(\theta, y) \propto \exp\left\{-\frac{1}{2}\begin{pmatrix}\theta - m \\ y - Rm\end{pmatrix}'\begin{pmatrix}V^{-1} + R'\Sigma^{-1}R & -R'\Sigma^{-1} \\ -\Sigma^{-1}R & \Sigma^{-1}\end{pmatrix}\begin{pmatrix}\theta - m \\ y - Rm\end{pmatrix}\right\},$$

so $p(\theta, m)$ follows a multivariate normal distribution with mean and precision

$$\theta, m \sim \mathrm{N}\left(\begin{pmatrix}\theta - m \\ y - Rm\end{pmatrix}, \begin{pmatrix}V^{-1} + R'\Sigma^{-1}R & -R'\Sigma^{-1} \\ -\Sigma^{-1}R & \Sigma^{-1}\end{pmatrix}\right).$$

□

11 In nonparametric regression and spatial smoothing

(a) Suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for some unknown function f. Suppose that the prior distribution for the unknown function is a mean-zero Gaussian process: $f \sim \mathrm{GP}(0, C)$ for some covariance function C. Let $x_1, \ldots, x_N$ denote the previously observed x points. Derive the posterior distribution for the random vector $[f(x_1), \ldots, f(x_N)]^T$, given the corresponding outcomes $y_1, \ldots, y_N$, assuming that you know σ^2.

(b) As before, suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, N$. Now we wish to predict the value of the function $f(x^\star)$ at some new point $x^\star$ where we haven't seen previous data. Suppose that f has a mean-zero Gaussian process prior, $f \sim GP(0, C)$. Show that the posterior mean $E\{f(x^\star) \mid y_1, \ldots, y_N\}$ is a linear smoother, and derive expressions both for the smoothing weights and the posterior variance of $f(x^\star)$.

(c) Go back to the utilities data, and plot the pointwise posterior mean and 95% posterior confidence interval for the value of the function at each of the observed points x_i (again, superimposed on top of the scatter plot of the data itself). Choose τ_2^2 to be very small, say 10^{-6}, and choose (b, τ_1^2) that give a sensible-looking answer.[15]

(d) Let $y_i = f(x_i) + \epsilon_i$, and suppose that f has a Gaussian-process prior under the Matern(5/2) covariance function C with scale τ_2^1, range b, and nugget τ_2^2. Derive an expression for the marginal distribution of $y = (y_1 \ldots, y_N)$ in terms of (τ_1^2, b, τ_2^2), integrating out the random function f. This is called a marginal likelihood.

(e) Return to the utilities or ethanol data sets. Fix $\tau_2^2 = 0$, and evaluate the log of the marginal likelihood function $p(y \mid \tau_1^2, b)$ over a discrete 2-d grid of points.[16] If you're getting errors in your code with $\tau_2^2 = 0$, use something very small instead. Use this plot to choose a set of values $(\hat{\tau_1^2}, \hat{b})$ for the hyperparameters. Then use these hyperparameters to compute the posterior mean for f, given y. Comment on any lingering concerns you have with your fitted model.

(f) In `weather.csv` you will find data on two variables from 147 weather stations in the American Pacific northwest.

pressure : the difference between the forecasted pressure and the actual pressure reading at that station (in Pascals)

[15] If you're bored with the utilities data, instead try the data in ethanol.csv, in which the NOx emissions of an ethanol engine are measured as the engine's fuel-air equivalence ratio (E in the data set) is varied. Your goal would be to model NOx as a function of E using a Gaussian process.

[16] Don't just use a black-box optimizer; we want to make sure we get the best solution if there are multiple modes.

temperature : the difference between the forecasted temperature and the actual temperature reading at that station (in Celsius)

There are also latitude and longitude coordinates of each station. Fit a Gaussian process model for each of the temperature and pressure variables. Choose hyperparameters appropriately. Visualize your fitted functions (both the posterior mean and posterior standard deviation) on a regular grid using something like a contour plot or color image. Read up on the `image`, `filled.contour`, or `contourplot`[17] functions in R. An important consideration: is Euclidean distance the appropriate measure to go into the covariance function? Or do we need separate length scales for the two dimensions, i.e.

$$d^2(x,z) = \frac{(x_1 - z_1)^2}{b_1^2} + \frac{(x_2 - z_2)^2}{b_2^2} \, .$$

Justify your reasoning for using Euclidean distance or this “nonisotropic” distance.

[17] in the lattice library