Chapter 9

More models

The more tools you have available, the better statistician you will be. In this chapter, we will consider some model building blocks (compounding and mixtures), issues with fitting models (identifiability specifically) and some examples of Bayesian models in fairly common contexts.

9.1 Constructing other distributions

We should not feel bound by the standard set of distributions for our modelling. We can make any density we wish. If we have some function $f(x) : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$, all we need to be able turn it into a density is

$$\int_{-\infty}^{\infty} f(x)\ dx < \infty.$$

That said, there are natural ways to create distributions that have features like extra spread or multiple modes using compounding and mixtures.

A *compound distribution* is formed by having a distribution assigned to parameters in another distribution and integrating out the uncertain parameters (this is similar to some of the hierarchical constructions that we have seen in the previous chapter).

Example 9.1.1

A t-distribution can be thought of as a compound distribution.
Consider

$$\begin{aligned} X|\tau &\sim \mathsf{N}(0, \tau^{-1}), \\ \tau &\sim \mathsf{Gamma}(\alpha, \beta). \end{aligned}$$

We want

$$\pi(x) = \int_0^\infty \pi(x|\tau)\pi(\tau)\ d\tau.$$

We know that

$$\pi(x|\tau)\pi(\tau) = \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)}\tau^{\alpha-1/2}\exp\left[-\left(\beta + x^2/2\right)\tau\right].$$

Spotting a Gamma density, we have that

$$\pi(x) = \frac{\beta^\alpha}{\sqrt{2\pi}\Gamma(\alpha)}\frac{\Gamma(\alpha+1/2)}{(\beta+x^2/2)^{\alpha+1/2}}.$$

Ignoring constants, we see that

$$\begin{aligned}\pi(x) &\propto \left(\beta + \frac{x^2}{2}\right)^{-\alpha-1/2} \\ &\propto \left(1 + \frac{x^2}{2\beta}\right)^{-\frac{2\alpha+1}{2}},\end{aligned}$$

which is a t-distribution with 2α degrees of freedom, location parameter 0 and scale parameter $\sqrt{\beta/\alpha}$.

Example 9.1.2

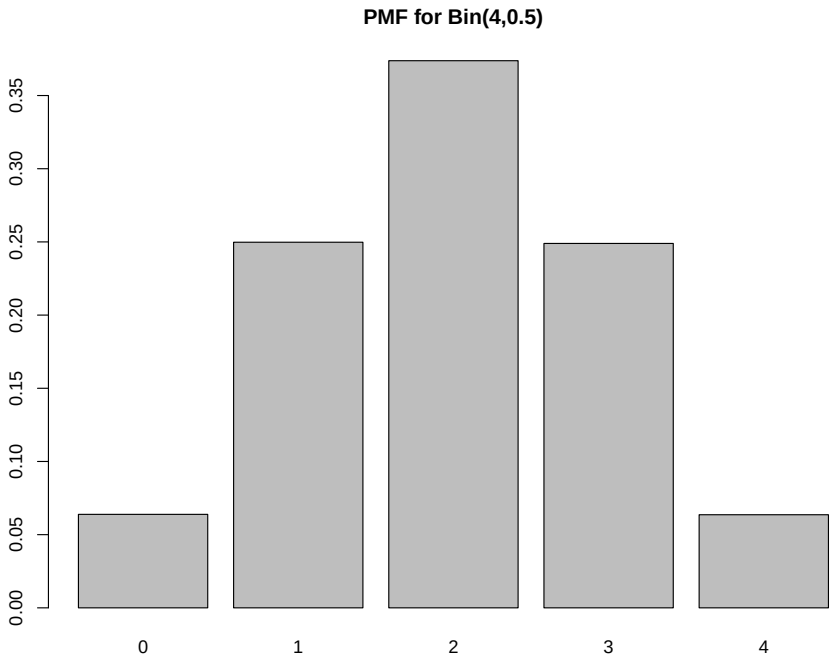
A discrete example:

$$\begin{aligned}Y|n &\sim \text{Bin}(n, 0.5), \\ n-1 &\sim \text{Poisson}(\lambda).\end{aligned}$$

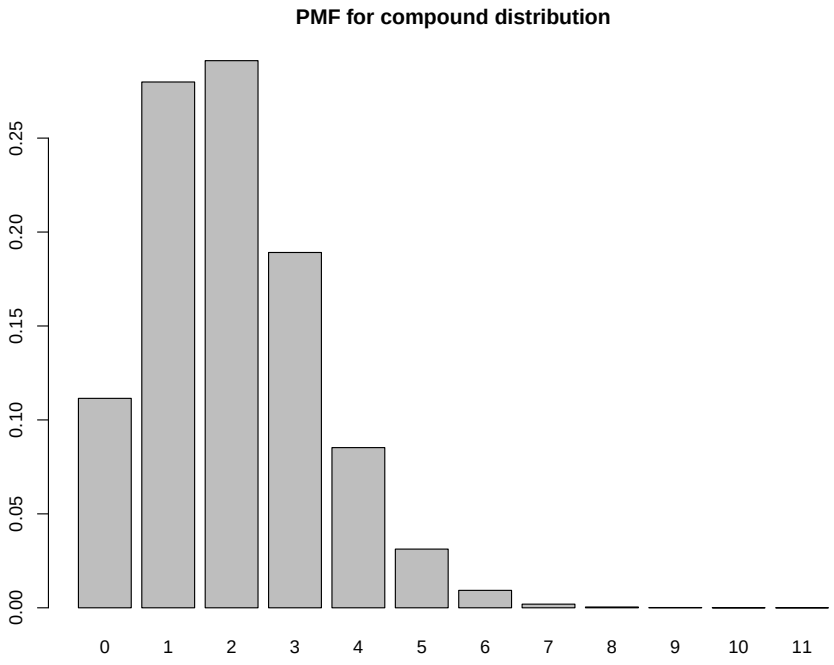
We have that

$$\pi(y) = \frac{e^{-\lambda}}{2y!}\left[\sum_{n=1}^{\infty}\frac{n}{(n-y)!}\left(\frac{\lambda}{2}\right)^{n-1}\right].$$

Compare the probability mass function for $Y \sim \text{Bin}(4, 0.5)$, which has variance 1,



with the probability mass function for our compound distribution with $\lambda = 3$, which has variance of approximately 1.75.



Another neat way of creating useful distributions is through mixtures. There are two primary types of mixture: sums and products. Let $\pi_1(\theta)$ and $\pi_2(\theta)$ be pdfs for θ, then we have the following valid densities

$$\pi(\theta) \propto w_1\pi_1(\theta) + w_2\pi_2(\theta)$$

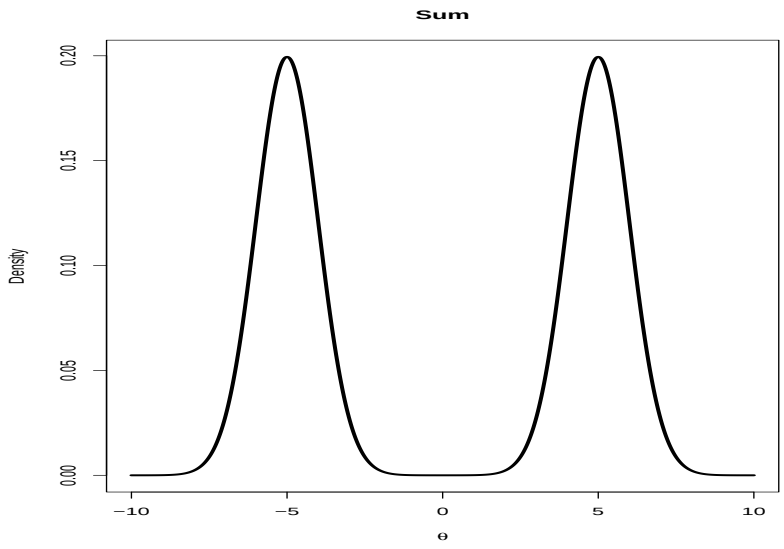
and

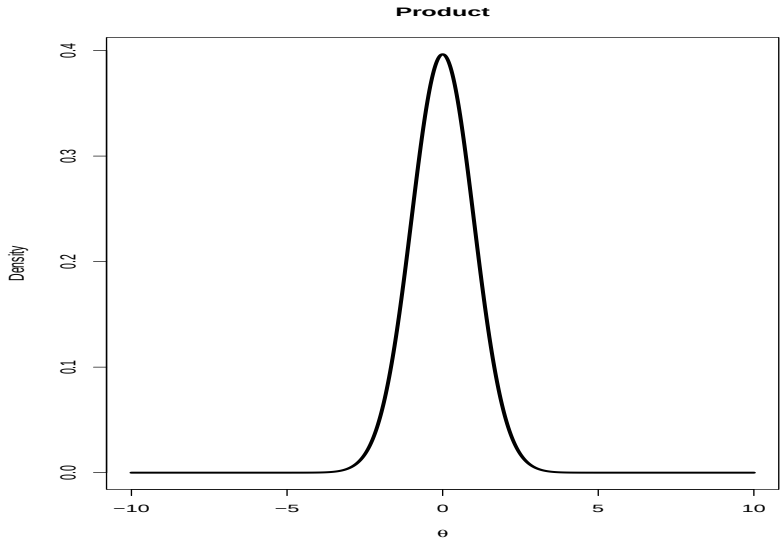
$$\pi(\theta) \propto \pi_1(\theta)^{w_1}\pi_2(\theta)^{w_2}$$

This extends to any number of mixture components.

Example 9.1.3

Let's have two densities, $\pi_1(\theta)$ and $\pi_2(\theta)$, based on N(-5,1) and N(5,1) respectively, and $w_1 = w_2 = 0.5$.





There are three other types of distribution manipulation that are commonly used to create custom distributions: *mirroring*, *folding* and *truncating*.

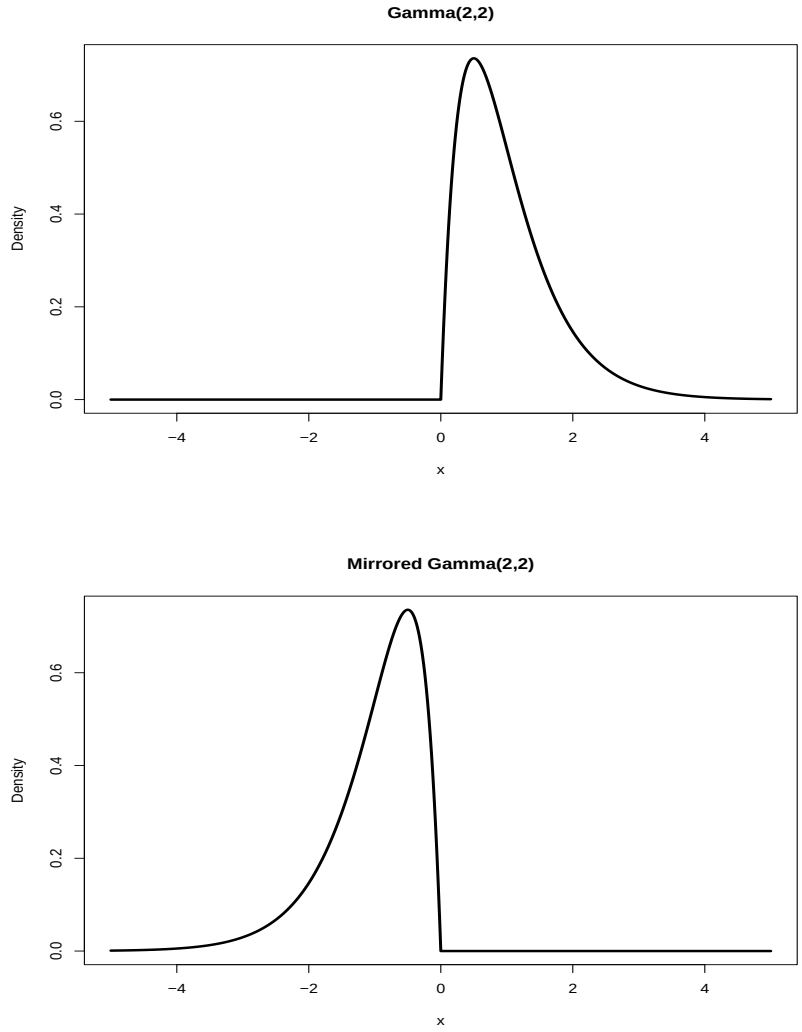
Mirroring a distribution essentially is reflecting the density about a lower or upper bound. This is usually accompanied by a shift in the distribution. Let $f(x)$ be a density for x that has a lower bound at a. A mirrored version of $f(x)$, $g(x)$ say, would be

$$g(x) = \begin{cases} f(2a - x) & x < a, \\ 0 & \text{otherwise.} \end{cases}$$

This is clearly still a valid density:

$$\int_{-\infty}^{\infty} g(x)dx \;=\; \int_{-\infty}^{a} f(2a - x)dx \;=\; -\int_{\infty}^{a} f(u)du \;=\; 1.$$

Example 9.1.4



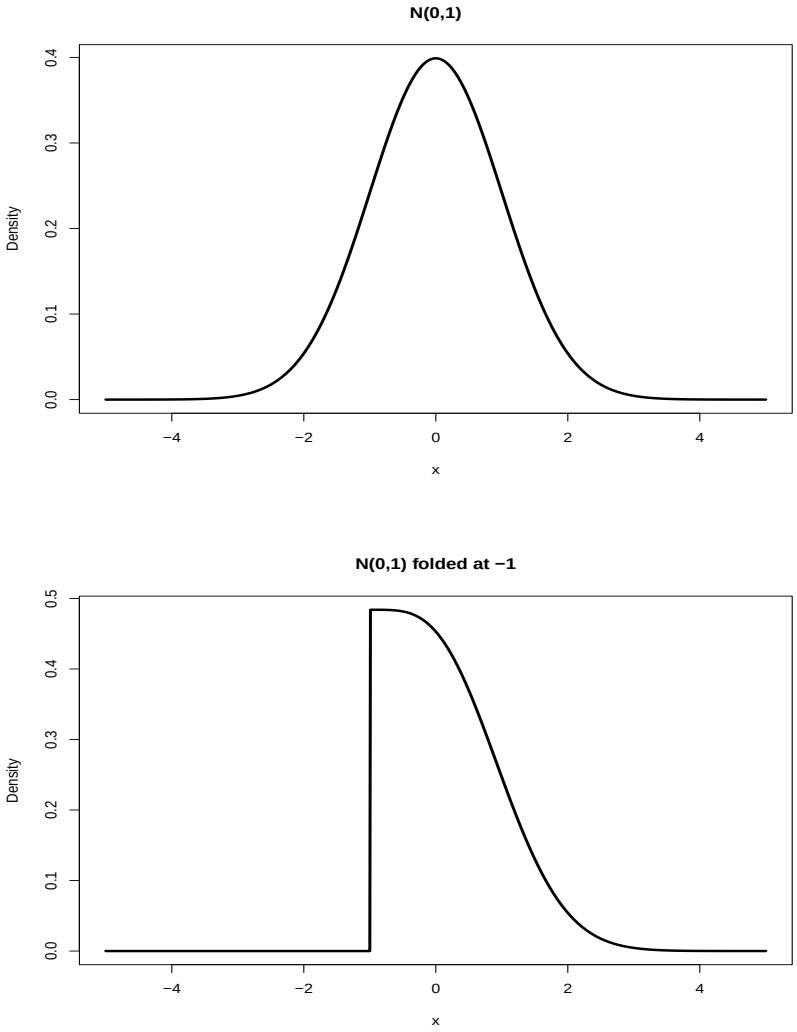
Folding a distribution means reflecting part of density in a boundary that you have imposed. It is useful for turning distributions defined on the real line into bounded distributions. Let $f(x)$ be a density that is positive for all $x \in \mathbb{R}$. A folded version of $f(x)$ at boundary a, $g(x)$ say, could be

$$g(x) = \begin{cases} f(x) + f(2a - x) & x > a, \\ 0 & \text{otherwise.} \end{cases}$$

This is a valid density:

$$\begin{aligned} \int_{-\infty}^{\infty} g(x)dx &= \int_{a}^{\infty} f(x) + f(2a - x)dx = \int_{a}^{\infty} f(x)dx + \int_{a}^{\infty} f(2a - x)dx \\ &= \int_{a}^{\infty} f(x)dx + \int_{-\infty}^{a} f(u)du = 1. \end{aligned}$$

Example 9.1.5



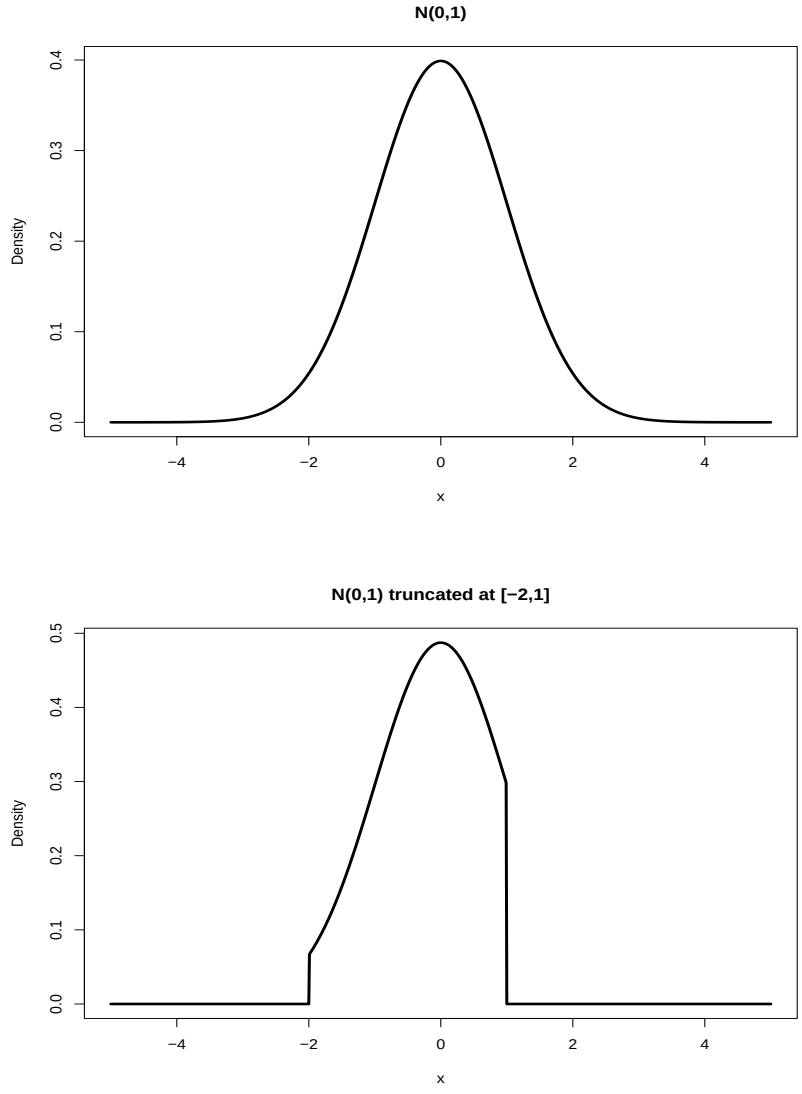
Truncation is much less subtle. We impose a bound or bounds on a distribution. Let $f(x)$ be a density that is positive for all $x \in \mathbb{R}$. A truncated version of $f(x)$ over $[a, b]$, $g(x)$ say, would be

$$g(x) = \begin{cases} \frac{f(x)}{\int_a^b f(y)dy} & x \in [a, b], \\ 0 & \text{otherwise.} \end{cases}$$

This too is a valid density:

$$\int_{-\infty}^{\infty} g(x)dx \;=\; \frac{\int_a^b f(x)dx}{\int_a^b f(y)dy} \;=\; 1.$$

Example 9.1.6



Transformations should also not be forgotten. Transformations are used throughout statistics to move data onto scales that are more amenable to modelling. In the following table, we have some unknown θ that we are transforming into ϕ.

Name	Function	Inverse	Domain	Range
Standardise	$\frac{\theta-\mu_\theta}{\sigma_\theta}$	$\phi\sigma_\theta+\mu_\theta$	$\mathbb{R}$	$\mathbb{R}$
Normalise	$\frac{\theta-\min(\theta)}{\max(\theta)-\min(\theta)}$	$[\max(\theta)-\min(\theta)]\,\phi+\min(\theta)$	$\mathbb{R}$	[0,1]
Power	θ^α	$\phi^{1/\alpha}$	$\mathbb{R}$ or $\mathbb{R}^+$	$\mathbb{R}$ or $\mathbb{R}^+$
Logarithm	$\log(\theta)$	$\exp(\theta)$	$\mathbb{R}^+$	$\mathbb{R}$
Logit	$\log\left(\frac{\theta}{1-\theta}\right)$	$\frac{1}{1+\exp(-\phi)}$	(0,1)	$\mathbb{R}$
Probit	$\Phi(\theta)$	$\Phi^{-1}(\phi)$	(0,1)	$\mathbb{R}$

9.2 Identifiability

An aspect of a statistical model is *identifiable* when it cannot be changed without there being a change in the distribution of the observed variables. More precisely, a statistical model is identifiable if it is possible to find the single true values of this model's parameters after observing an infinite number of observations.

If we can alter part of a model with no consequences for the data distributions, then that part of the model is *unidentifiable*. In this case, two or more sets of parameters give rise to equivalent data distributions.

Example 9.2.1

$X_i \sim N(\mu_1+\mu_2, \sigma^2)$, where $i = 1, \ldots, n$.
σ^2 is identifiable.
μ_1 and μ_2 are unidentifiable.
This is clear from the log-likelihood

$$L(\mu_1, \mu_2, \sigma^2 | x_\bullet) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu_1-\mu_2)^2$$

where we can always change μ_1 to compensate for any change to μ_2, but we cannot counteract changes to σ^2 so easily. However, $\mu = \mu_1+\mu_2$ is identifiable.

In a Bayesian context, suppose that $\pi(x|\boldsymbol{\theta})$ depends on some function of $\boldsymbol{\theta}$, $\boldsymbol{g}(\boldsymbol{\theta})$ say, but not

on the rest, $\boldsymbol{h}(\boldsymbol{\theta})$ say. When we derive the posterior, we get

$$\begin{aligned}\pi(\boldsymbol{\theta}|x) &\propto \pi(x|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ &= \pi(x|\boldsymbol{g}(\boldsymbol{\theta}))\pi(\boldsymbol{\theta}) \\ &= \pi(x|\boldsymbol{g}(\boldsymbol{\theta}))\pi(\boldsymbol{h}(\boldsymbol{\theta})|\boldsymbol{g}(\boldsymbol{\theta}))\pi(\boldsymbol{g}(\boldsymbol{\theta})) \\ &\propto \pi(\boldsymbol{g}(\boldsymbol{\theta})|x)\pi(\boldsymbol{h}(\boldsymbol{\theta})|\boldsymbol{g}(\boldsymbol{\theta})).\end{aligned}$$

Therefore, the conditional posterior distribution of $\boldsymbol{h}(\boldsymbol{\theta})$ given $\boldsymbol{g}(\boldsymbol{\theta})$ is the same as the conditional prior distribution of $\boldsymbol{h}(\boldsymbol{\theta})$ given $\boldsymbol{g}(\boldsymbol{\theta})$, and, no matter how many we observe, we can never learn $\boldsymbol{h}(\boldsymbol{\theta})$ precisely.

Example 9.2.1 continued.
Here, we have $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma^2)^T$, $\boldsymbol{g}(\boldsymbol{\theta}) = (\mu_1 + \mu_2, \sigma^2)^T$ and $\boldsymbol{h}(\boldsymbol{\theta}) = \mu_1 - \mu_2$.
From the likelihood, we know that $\pi(\underline{x}|\boldsymbol{\theta}) = \pi(\underline{x}|\boldsymbol{g}(\boldsymbol{\theta}))$,
and we will find that

$$\pi(\boldsymbol{\theta}|\underline{x}) \propto \pi(\underline{x}|\boldsymbol{g}(\boldsymbol{\theta}))\pi(\boldsymbol{h}(\boldsymbol{\theta})|\boldsymbol{g}(\boldsymbol{\theta})).$$

So we will never find the difference and will never be able to determine the values of μ_1 and μ_2 beyond what we have specified in our prior.

When using maximum likelihood estimation, a lack of identifiability leads to ill-posed optimisation. In these cases, penalty or regularisation terms are added to the log-likelihood to remove the problems. In Bayesian inference, the priors automatically do the job for us:

$$\log[\pi(\theta|x)] = L(\theta|x) + \log[\pi(\theta)].$$

This does not mean that Bayesians should ignore this though. Identifiability issues still appear when sampling from posteriors (especially when the priors are weak), and we should always investigate dependencies between parameters in the posterior.

9.3 Linear regression

Linear regression is one of the most used and simplest statistical models. The textbook Bayesian formulation of a linear regression model is straightforward:

$$\begin{aligned}y_i|\alpha, \beta, \sigma^2, x_i &\sim \mathsf{N}(\alpha + \beta x_i, \sigma^2), \\ \alpha|\sigma_\alpha^2 &\sim \mathsf{N}(0, \sigma_\alpha^2), \\ \beta|\sigma_\beta^2 &\sim \mathsf{N}(0, \sigma_\beta^2), \\ \sigma^2 &\sim \mathsf{InvGamma}(a, b),\end{aligned}$$

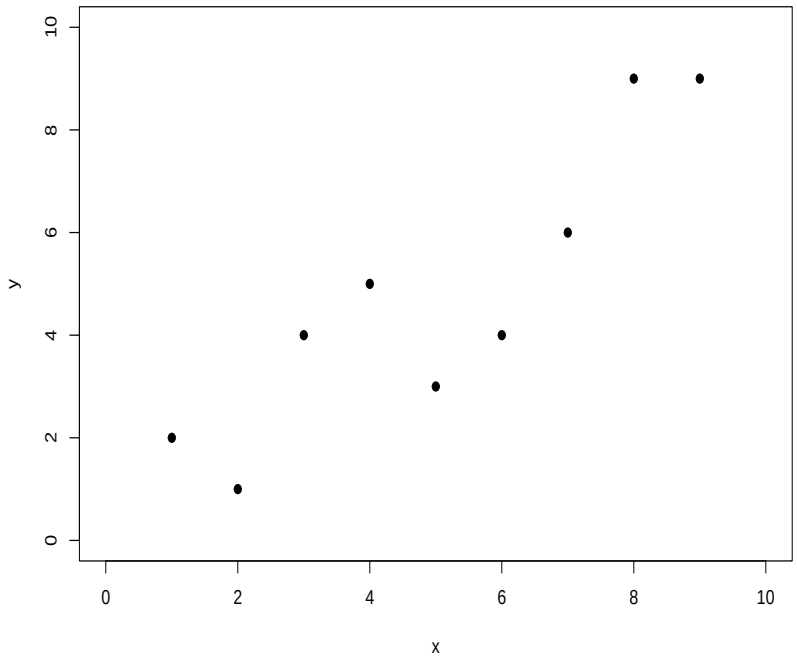
where we have used the standard conjugate form for the priors and the hyperparameters a, b, σ_α^2 and σ_α^2 need to be specified.

This model can be coded in the following way in Stan:

```
data {
  int<lower=0> N;          // num observations
  real x[N];               // observed explanatory variable
  real y[N];               // observed response variable
}
parameters {
  real alpha;             // intercept
  real beta;              // gradient
  real<lower=0> sigma_2; // error variance
}
model {
  // Prior
  alpha ~ normal(0,10000);
  beta ~ normal(0,10000);
  sigma_2 ~ inv_gamma(0.01,0.01);

  // Likelihood
  for (n in 1:N)
      y[n] ~ normal(alpha + beta*x[n], sqrt(sigma_2));
}
```

Example 9.3.1

Let's have the following data that we believe follows a straight line relationship:

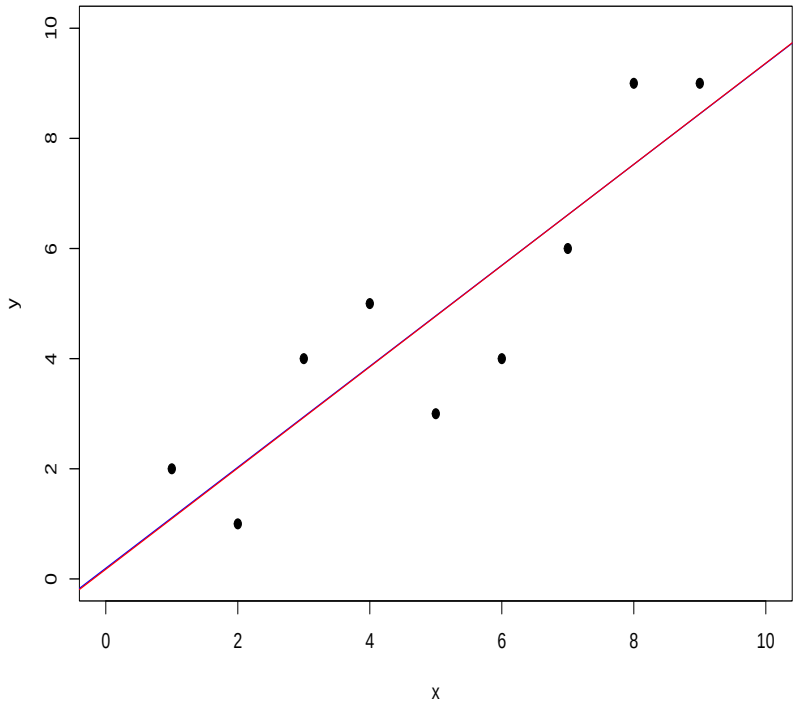


Estimating parameters using least squares fitting in R is trivial:

```
Call:
lm(formula = y ~ x, data = reg_data)

Coefficients:
(Intercept)            x
     0.1944       0.9167
```

Compare this with summaries from Stan:

```
$summary
              mean     se_mean        sd        50%      n_eff     Rhat
alpha    0.1754029 0.012460406 1.1587019  0.1859671   8647.270 1.000327
beta     0.9191603 0.002216499 0.2068693  0.9176843   8710.781 1.000408
sigma_2  2.5749959 0.028843551 1.9171726  2.0599911   4417.993 1.000472
lp__    -8.7725645 0.026130836 1.8224132 -8.3245564   4863.928 1.000254
```



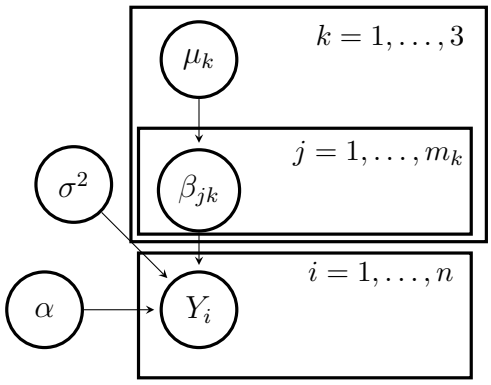
Extension to multiple linear regression (with m explanatory variables) is straightforward:

$$\begin{aligned} y_i|\alpha, \boldsymbol{\beta}, \sigma^2, \boldsymbol{x}_i &\sim \mathsf{N}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i, \sigma^2), \qquad i = 1, \ldots, n, \\ \alpha|\sigma_\alpha^2 &\sim \mathsf{N}(0, \sigma_\alpha^2), \\ \beta_j|\sigma_\beta^2 &\sim \mathsf{N}(0, \sigma_\beta^2), \qquad j = 1, \ldots, m, \\ \sigma^2 &\sim \mathsf{InvGamma}(a, b), \end{aligned}$$

In Stan, we have to extend the data format to handle arrays:

```
data {
  int<lower=0> N;          // num observations
  int<lower=1> M;          // num explanatory variables
  real x[N,M];             // observed explanatory variables
  real y[N];               // observed response variable
}
```

We can also imagine building hierarchical priors in this setting if we believed that certain explanatory variables had similar effects on the response. If we had six explanatory variables with three different expected behaviours (specifically, the first four are similar and the final two are not similar to any other), we could have:



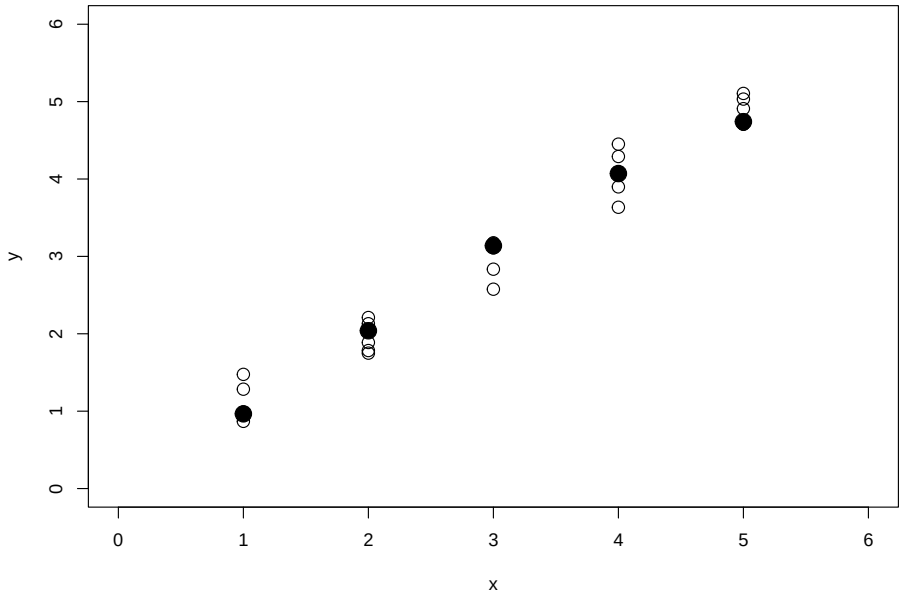
9.3.1 Errors-in-variables

Errors-in-variables regression is a very natural extension of the linear regression models of the previous section. A strong assumption in linear regression is that there is no (or relatively little) error in the recorded x values. Relaxing this assumption gives rise to the errors-in-variables model, which is simple enough to write down:

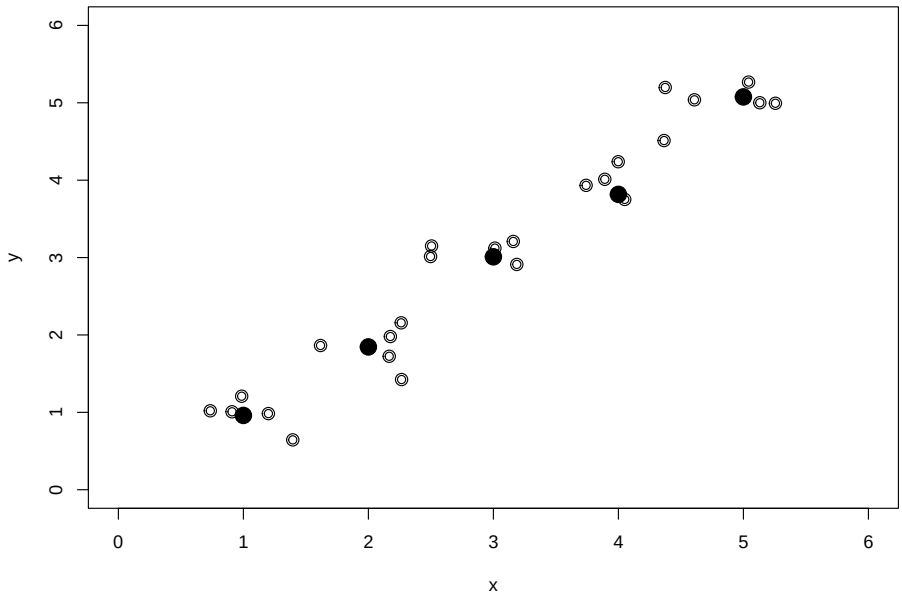
$$\begin{aligned} \widetilde{y}_i &= \alpha + \beta \widetilde{x}_i, \\ y_i|\sigma_y^2 &\sim \mathsf{N}(\widetilde{y}_i, \sigma_y^2), \\ x_i|\sigma_x^2 &\sim \mathsf{N}(\widetilde{x}_i, \sigma_x^2), \end{aligned}$$

for $i = 1, \ldots, n$.

In simple linear regression, our errors act in a straightforward manner. In the following plot, we have filled dots as the "true" values and unfilled dots as possible observations of that truth.



In errors-in-variables regression, our errors come in two directions.



Thinking about the errors-in-variables model, we can see that we have an identifiability issue (dropping the i index for clarity):

$$\begin{aligned} y &= \widetilde{y} + \epsilon_y \\ &= \alpha + \beta\widetilde{x} + \epsilon_y, \\ &= \alpha + \beta\left(x - \epsilon_x\right) + \epsilon_y, \\ &= \alpha + \beta x + (\epsilon_y - \beta\epsilon_x). \end{aligned}$$

Now, if we have $\epsilon_x \sim \mathsf{N}(0, \sigma_x^2)$ and $\epsilon_y \sim \mathsf{N}(0, \sigma_y^2)$, then we will have

$$(\epsilon_y - \beta\epsilon_x) \sim \mathsf{N}(0, \sigma_y^2 + \beta^2\sigma_x^2).$$

There are multiple ways to attempt to solve this identifiability issue in a non-Bayesian setting. The most well known is to assume that the ratio σ_y^2/σ_x^2 is known. This would be a very strong piece of information and is enough to constrain the problem because you are effectively going from two parameters to one:

$$\sigma_y^2 = c\sigma_x^2$$

with c known.

As noted in Section 9.2, we automatically circumnavigate this issue in a Bayesian setting through the use of proper prior distributions. The Bayesian model set-up is interesting in that all the $\widetilde{y}_i$ and $\widetilde{x}_i$ become unknown parameters. We have a formula for $\widetilde{y}_i$, but we need to specify a prior distribution for the $\widetilde{x}_i$. In Stan, this can be handled in the following way:

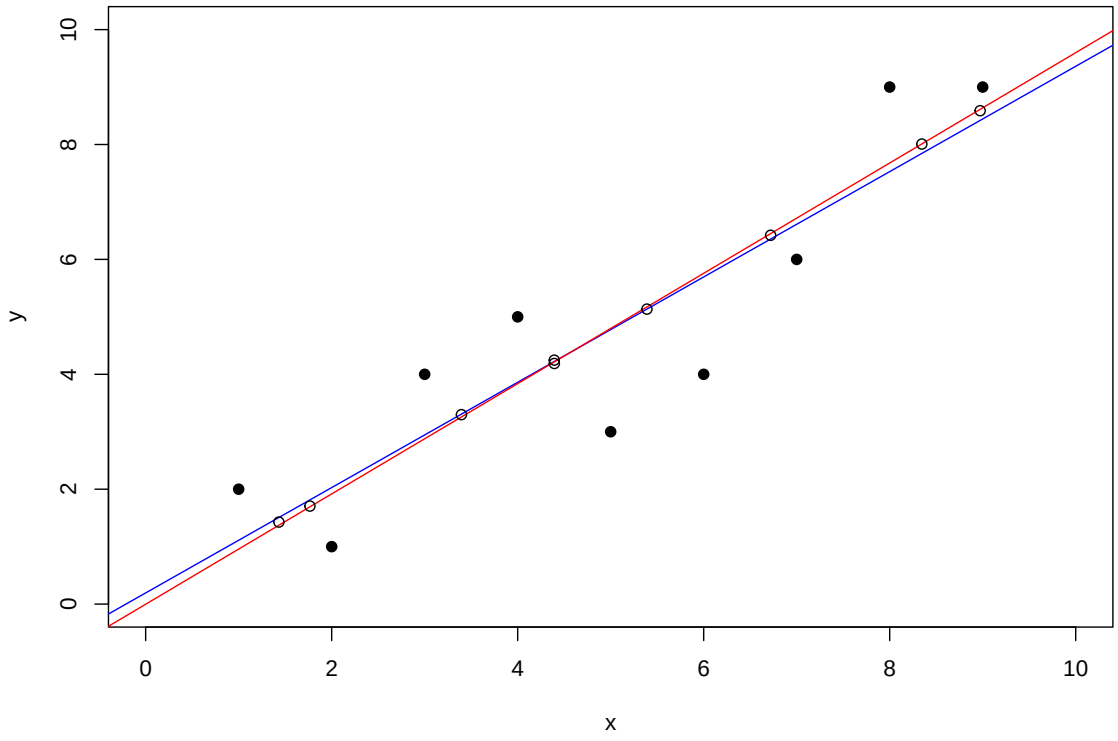
```
parameters {
  real alpha;                          // intercept
  real beta;                           // gradient
  real<lower=0> sigma_2_x;             // error variance for x
  real<lower=0> sigma_2_y;             // error variance for y
  real<lower=0,upper=10> true_x[N];  // the true values of x
}
model {
  // Prior
  alpha ~ normal(0,10000);
  beta ~ normal(0,10000);
  sigma_2_x ~ inv_gamma(0.01,0.01);
  sigma_2_y ~ inv_gamma(0.01,0.01);
  true_x ~ uniform(0,10);   // This is a vectorised form.

  // Likelihood
  for (n in 1:N){
      x[n] ~ normal(true_x[n], sqrt(sigma_2_x));
      y[n] ~ normal(alpha + beta*true_x[n], sqrt(sigma_2_y));
      }
}
```

If we also want to get out the $\widetilde{y}_i$, they can be computed in a `transformed parameters` block:

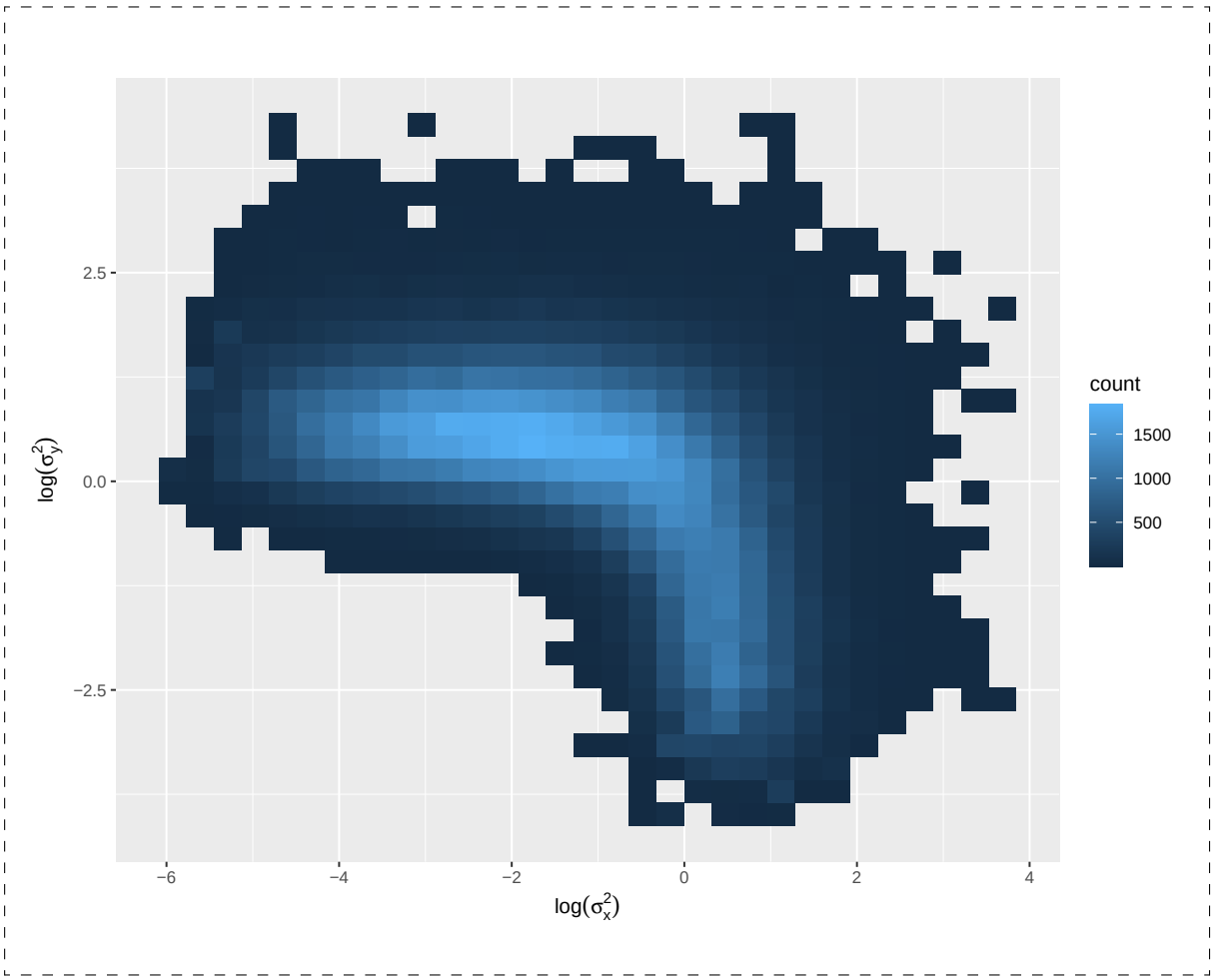
```
transformed parameters {
  real true_y[N];   // the true values of y

  for (n in 1:N)
      true_y[n] = alpha + beta * true_x[n];
}
```

Example 9.3.2

Going back to our example for simple linear regression. We may ask what would the relationship be if we applied an errors-in-variables model.



There is not a great deal of difference in the fits (probably because the errors are similar in both directions and the prior information is weak). We will see an example in the practical where it does make a real difference. The unfilled circles here are the posterior mean estimates of the true set of x and y pairs.
There is still a hangover in the posterior samples from the identifiability issue though. If we plot the 2d-density estimate for $\log(\sigma_x^2)$ and $\log(\sigma_y^2)$, there is clear dependence of the type that can make sample from the posterior problematic.



9.3.2 Logistic regression

Now, we further extend into generalised linear models. Logistic regression marries linear regression with classification through the use of a Bernoulli likelihood rather than a normal one. Let y_i be the ith observation of whether something is true (encoded as 1) or not (encoded as 0). A simple logistic regression model is as follows:

$$\begin{aligned} y_i|\alpha,\beta,x_i &\sim \text{Bernoulli}\left[g(x_i,\alpha,\beta)\right], \\ g(x_i,\alpha,\beta) &= \frac{1}{1+\exp(-\alpha-\beta x_i)}, \\ \alpha &\sim \text{N}(0,\sigma_\alpha^2), \\ \beta &\sim \text{N}(0,\sigma_\beta^2), \end{aligned}$$

where the hyperparameters have been chosen for x_i have been standardised. From earlier,

$$\frac{1}{1+\exp(-\alpha-\beta x_i)}$$

is the inverse of the logit transformation: the logistic function.

Stan has functions to help with this type of model:

```
data {
  int<lower=0> N;
  vector[N] x;
  int<lower=0,upper=1> y[N];
}
parameters {
  real alpha;
  real beta;
}
model {
  // Prior
  alpha ~ normal(0,10000);
  beta ~ normal(0,10000);

  // Likelihood
  y ~ bernoulli_logit(alpha + beta * x);   // This is in vectorised form.
}
```

Example 9.3.3

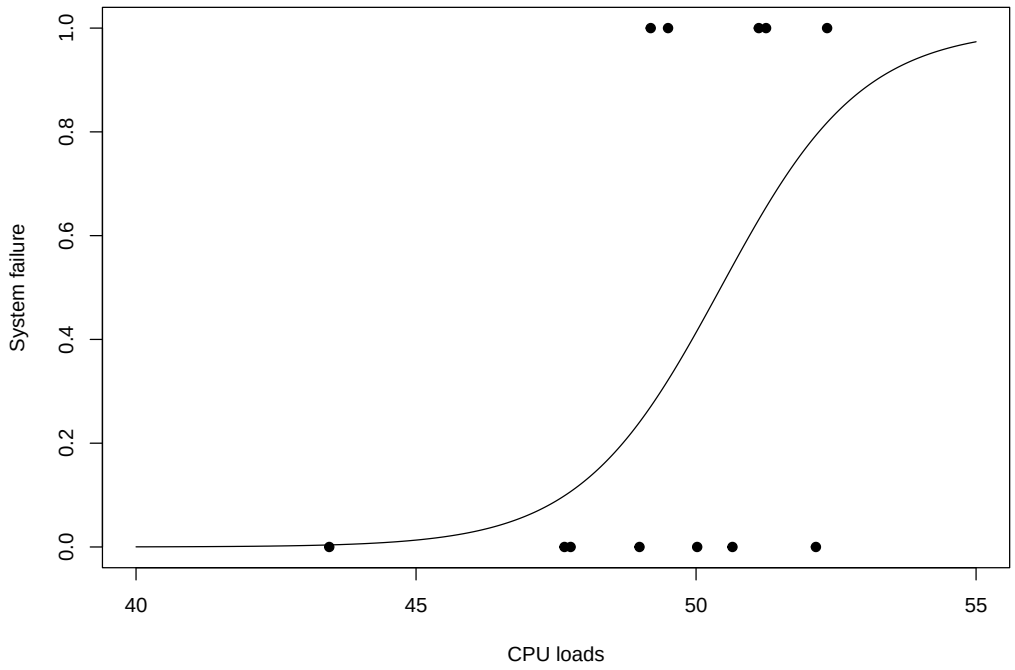
Consider the following data on system failures with respect to average CPU load.

CPU load (%)	System failure
47.65	FALSE
48.99	FALSE
50.65	FALSE
49.19	TRUE
51.25	TRUE
52.34	TRUE
52.14	FALSE
49.50	TRUE
43.45	FALSE
47.76	FALSE
50.02	FALSE
51.12	TRUE

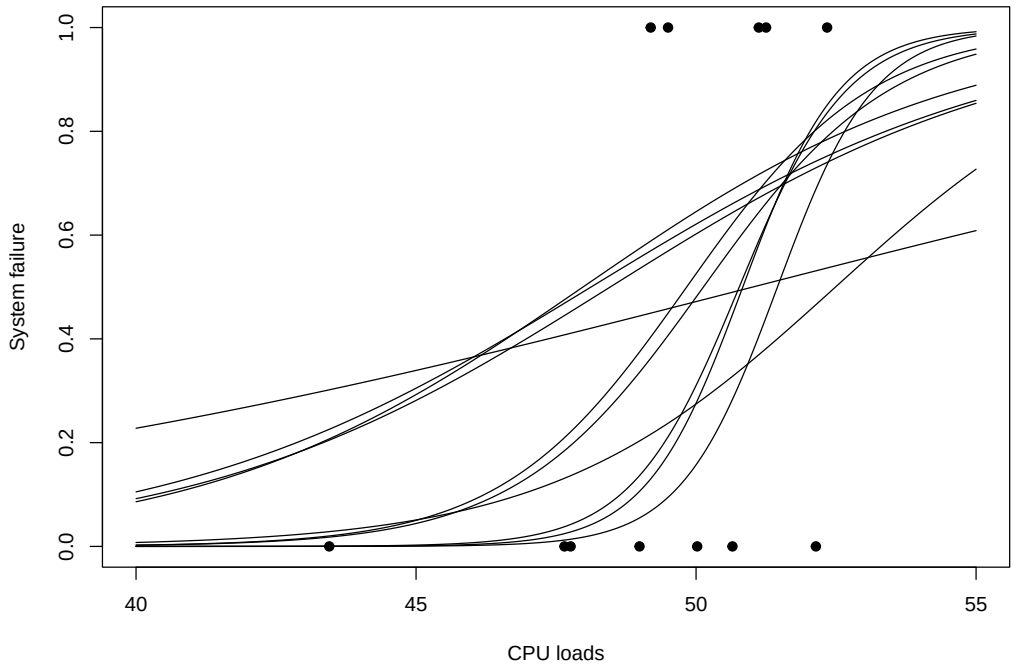
We can use the logistic regression model code to get estimates of the effect CPU load seems to have on system failure.

```
             mean     se_mean         sd
alpha -39.9551183 0.167568775 24.8039672
beta    0.7920968 0.003337940  0.4942338
```

It is more interesting to look at the logistic function with the posterior mean estimates plotted against the data.



It is better to consider the range of plausible relationships from the posterior distribution:



9.4 Compositional data

Compositional data are a type of multivariate data where all the values are restricted to the interval $(0, \kappa)$ and the values for each observation must sum to κ. The value of κ is usually either 1 (when we are considering proportions) or 100 (when we are considering percentages).

Example 9.4.1

Five rocks have been analysed to check their metal composition in terms of iron, nickel and other metals:

$$X = \begin{pmatrix} 0.25 & 0.12 & 0.63 \\ 0.21 & 0.11 & 0.68 \\ 0.29 & 0.12 & 0.59 \\ 0.19 & 0.10 & 0.71 \\ 0.29 & 0.14 & 0.57 \end{pmatrix}$$

We can calculate the correlation matrix:

$$R = \begin{pmatrix} 1.00 & 0.87 & -0.99 \\ & 1.00 & -0.93 \\ & & 1.00 \end{pmatrix}$$

From the raw measurements, we have that all of these rocks had almost exactly the same amount of iron in them in terms of mass

These data cause problems to standard statistical methods because:

1. The data are bounded unlike the most-used multivariate distributions.
2. There is at least one perfect linear relationship in the variables.
3. We are forced to have negative correlation, and this can distort interpretation.

In fact, p-dimensional compositional data are defined on a special subset of $\mathbb{R}^p$: the simplex

$$\mathcal{S}^p = \left\{ \mathbf{x} = [x_1, x_2, \ldots, x_p] \in \mathbb{R}^p \,\middle|\, x_i > 0, i = 1, 2, \ldots, p; \sum_{i=1}^{p} x_i = \kappa \right\}.$$

Effectively, the data points are in $p - 1$ dimensions because if we know the value for $p - 1$ of the variables, we can calculate the value of the other using:

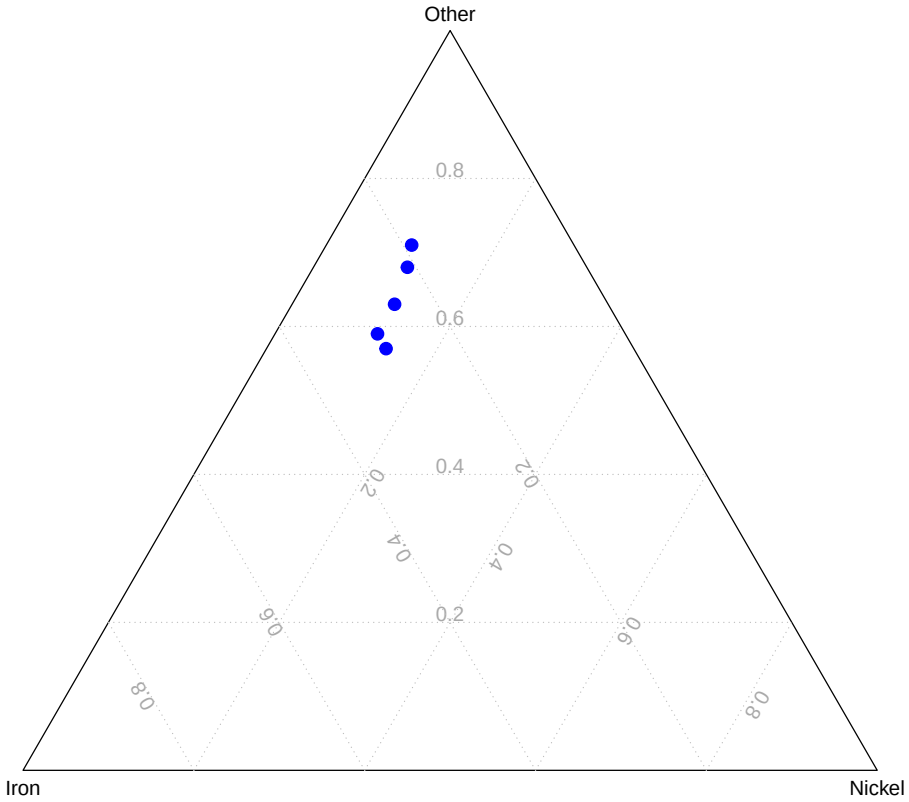
$$x_i = \kappa - \sum_{j \neq i} x_j.$$

A common way of displaying compositional data is through a ternary plot that takes advantage

of the data being constrained on the simplex. For $p = 3$, we just need a triangle.

Example 9.4.2

Here we display the data from the previous example in a ternary plot.



The archetypal distribution defined over the simplex is the *Dirichlet distribution*. If we have compositional data supported over $p-1$-$(0,1)$-simplex, then

$$\pi(\mathbf{x}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{p} x_i^{\alpha_i - 1}$$

where the multivariate Beta function is

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{p} \Gamma(\alpha_i)}{\Gamma\big(\alpha_0\big)}$$

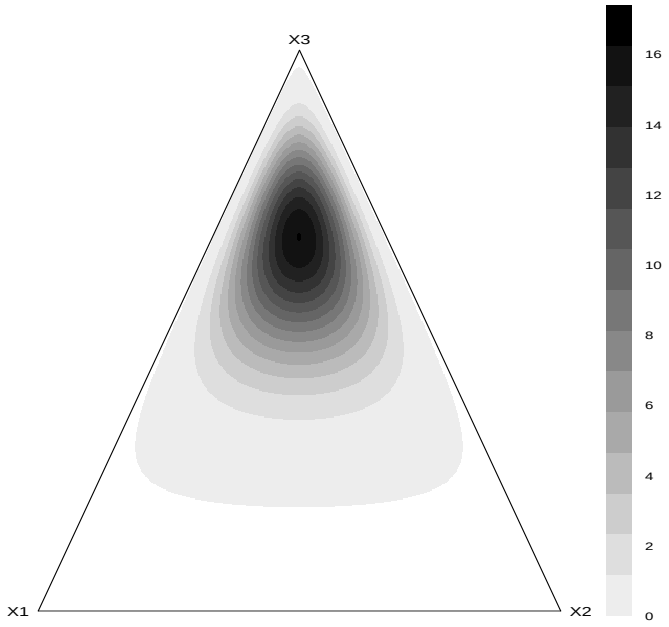
and where

$$\alpha_0 = \sum_{i=1}^{p} \alpha_i.$$

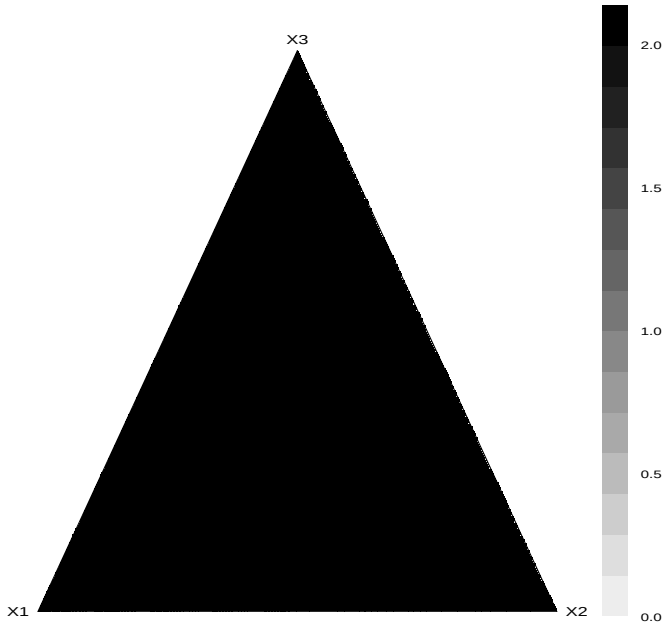
The parameters of the Dirichlet distribution act in a very similar way to the parameters of a Beta

distribution. In fact, there is a strong connection between the two as the marginal distributions of the x_i are Beta ($x_i \sim \text{Be}\left(\alpha_i, \sum_{-i} \alpha_j\right)$ in fact).

Here is a 2d density plot on the simplex for a Dir$(3, 3, 9)$ distribution.



And here is the same for a Dir$(1, 1, 1)$ distribution.



Example 9.4.3

Let's imagine that we believe the data from the previous examples follow a Dirichlet distribution. We might have the following model:

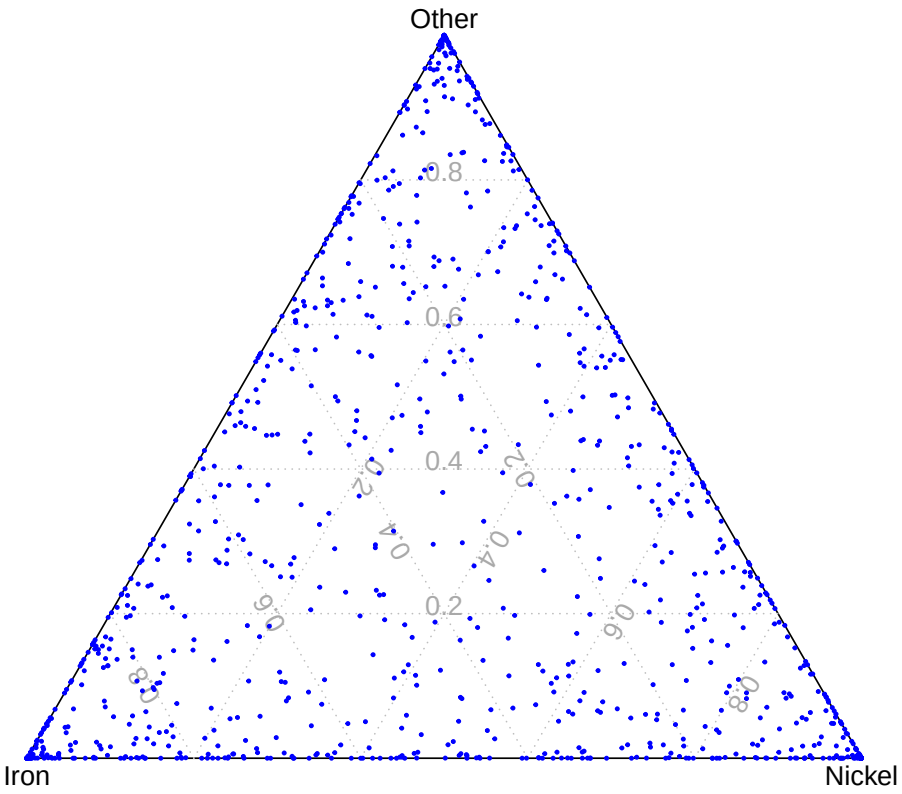
$$\begin{aligned} \mathbf{x}|\boldsymbol{\alpha} &\sim \text{Dir}(\boldsymbol{\alpha}), \\ \alpha_i &\sim \text{Gamma}(1,1) \text{ for } i = 1, \ldots, p. \end{aligned}$$

```
data {
  int<lower=0> N;
  matrix[N,3] x;
}
parameters {
  vector[3] alpha;
}
model {
  vector[3] x_vector;

  // Prior
  alpha ~ gamma(1,1);

  // Likelihood
  for (i in 1:N){
    x_vector = to_vector(x[i]);
    x_vector ~ dirichlet(alpha);
  }
}
```

We can also use Stan to investigate our preposterior for $\mathbf{x}$

```
parameters {
  vector[3] alpha;
}
model {
  // Prior
  alpha ~ gamma(1,1);
}
generated quantities {
  vector[3] x_vector;
  x_vector = dirichlet_rng(alpha);
}
```

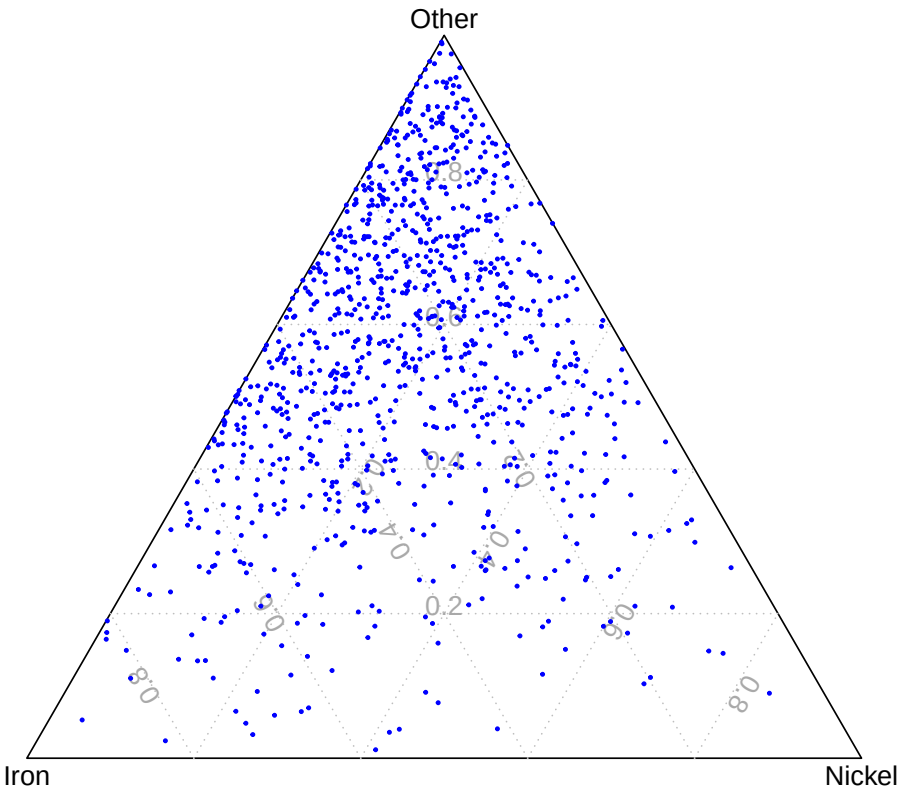
Here is a ternary plot with 1000 samples from the preposterior distribution.



Bringing the data in, we get posterior statistics of $(E)(\alpha_1) = 1.93(0.78)$, $(E)(\alpha_2) = 1.18(0.48)$ and $(E)(\alpha_3) = 4.15(1.70)$, where the posterior standard deviations are given in the brackets.
We can add the following code to the bottom of our posterior sampling code to get a similar sample for our predictive distribution.

```
generated quantities {
  vector[3] x_;
  x_ = dirichlet_rng(alpha);
}
```

Here is a ternary plot with 1000 samples from the predictive distribution.



By using distributions defined over the simplex, we lose the great number of techniques that have been devised for unbounded or multivariate normal datasets. The data are effectively in $p-1$ dimensions so a transformation would be useful. However, a simple linear transformation would not remove the problems with bounds or forced correlation.

There are many different transformation that can be used, but we will focus on the additive log-ratio:

$$\begin{aligned}\mathbf{y} &= (y_1, \ldots, y_{p-1})^T = \text{alr}(\mathbf{x}) \\ &= \left[\log\left(\frac{x_1}{x_p}\right), \ldots, \log\left(\frac{x_{p-1}}{x_p}\right)\right]^T\end{aligned}$$

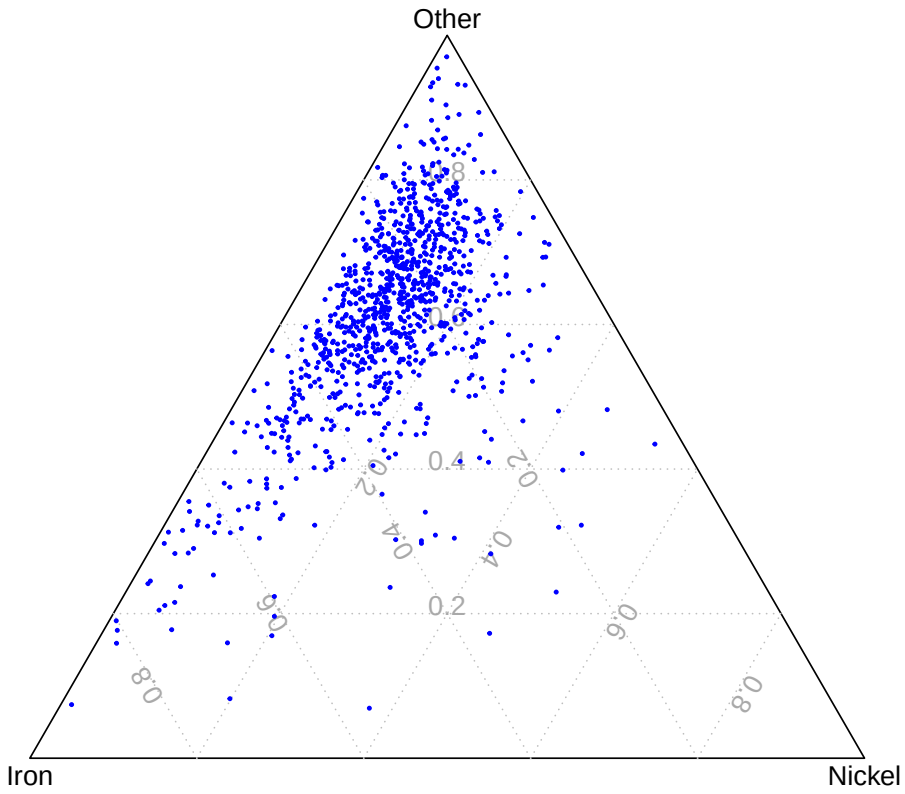
(here, $\mathbf{y} \in \mathbb{R}^{p-1}$ and note that the ordering of the variables is arbitrary). This transformation tends to be useful because it directly utilises the fact that compositional data gives us information about relative size alone. We may then proceed to use methods and distributions that are defined for unbounded real spaces. We need to decide which variable to choose as the divisor and in some cases this is straightforward if there is a clear interpretation. For instance, all types of household expenditure relative to food expenditure.

Example 9.4.4

We can apply the ALR transformation to our data within Stan:

```
data {
  int<lower=0> N;
  matrix[N,3] x;
}
transformed data {
  matrix[N,2] y;
  for (i in 1:N){
    for (j in 1:2){
      y[i,j] = log(x[i,j]/x[i,3]);
    }
  }
}
parameters {
  vector[2] mu;
  cov_matrix[2] Sigma;
}
model {
  vector[2] y_vector;

  // Prior
  mu ~ normal(0,100);
  Sigma ~ inv_wishart(2, diag_matrix(rep_vector(1.0, 2)));

  // Likelihood
  for (i in 1:N){
    y_vector = to_vector(y[i]);
    y_vector ~ multi_normal(mu, Sigma);
  }
}
generated quantities {
  vector[2] y_;
  vector[3] x_;
  real unnormalised_sum;

  y_ = multi_normal_rng(mu, Sigma);
  unnormalised_sum = exp(y_[1]) + exp(y_[2]) + 1;
  x_[1] = exp(y_[1])/unnormalised_sum;
  x_[2] = exp(y_[2])/unnormalised_sum;
  x_[3] = 1/unnormalised_sum;
}
```

Here is a ternary plot with 1000 samples from the predictive distribution.



These kind of constraints come up more often when considering priors for stochastic vectors. A *stochastic vector* is a vector whose elements must sum to one. We may have already seen these as parameters in a multinomial distribution or as rows in a right-stochastic matrix when considering discrete Markov chains.

A multinomial distribution for $x_1, \ldots, x_p$ observed after n trials has the following probability mass function:

$$\pi(\mathbf{x}|\boldsymbol{\theta}) = \frac{n!}{x_1! \cdots x_p!} \prod_{i=1}^{p} \theta_i^{x_i}$$

where the $\theta_i > 0$ and $\sum \theta_i = 1$. The Dirichlet distribution works as a conjugate prior for $\boldsymbol{\theta}$ when we have multinomial data:

$$\begin{aligned}\pi(\boldsymbol{\theta}|\mathbf{x}) &\propto \prod_{i=1}^{p} \theta_i^{x_i} \prod_{i=1}^{p} \theta_i^{\alpha_i - 1} \\ &\propto \prod_{i=1}^{p} \theta_i^{\alpha_i + x_i - 1}.\end{aligned}$$

So a Dir$(\alpha_1, \ldots, \alpha_p)$ prior for $\boldsymbol{\theta}$ becomes a Dir$(\alpha_1 + x_1, \ldots, \alpha_p + x_p)$ posterior.

Example 9.4.5

We are to receive data about eye colour in the form of the number of observations in each of four mutually exclusive categories: [Amber, Brown, Hazel], [Blue], [Green], [Other]. *A priori*, we believe that, in the population we will study, that the first two categories are much more likely so we posit a Dir$(10, 10, 1, 1)$ prior.
The mean for each component of a Dirichlet distribution is

$$\mathsf{E}(\theta_i) = \frac{\alpha_i}{\sum \alpha_j}.$$

So our prior means are $\mathsf{E}(\theta_1) = \mathsf{E}(\theta_2) = 0.45$ and $\mathsf{E}(\theta_3) = \mathsf{E}(\theta_4) = 0.05$ to 2 d.p..
We receive a sample of 150 eye colours: 49, 62, 13, 26 for each category. By conjugacy, we have a Dir$(59, 72, 14, 27)$ posterior, and our posterior means are $\mathsf{E}(\theta_1|\mathbf{x}) = 0.34$, $\mathsf{E}(\theta_2|\mathbf{x}) = 0.42$, $\mathsf{E}(\theta_3|\mathbf{x}) = 0.08$ and $\mathsf{E}(\theta_4|\mathbf{x}) = 0.16$ to 2 d.p..