Solution 1 - Linear regression

The model for N data points is

$$P(\{y_i\}\,|\,\{x_i\},a,c,\sigma^2)=\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(y_i-(ax_i+c))^2}{2\sigma^2}\right]=[2\pi\sigma^2]^{\frac{N}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}[y_i-(ax_i+c)]^2\right].$$

Let's choose a non-informative prior

$$P_0(a,c,\sigma^2)=\frac{\text{const.}}{\sigma^2},$$

and write the posterior as

$$P(a,c,\sigma^2\,|\,\{y_i\},\{x_i\})=\frac{1}{Z}(\sigma^2)^{-\frac{N+2}{2}}\exp\left[-\frac{1}{2\sigma^2}\left[y_i-(ax_i+c)\right]^2\right].$$

The quadratic form under the exponential suggests that the following statistics are going to be useful:

$$N=7,\qquad \langle x\rangle=\frac{\sum_i x_i}{N}=4.0,\qquad \langle y\rangle=\frac{\sum_i y_i}{N}=8.047,$$
$$\langle x^2\rangle=\frac{\sum_i x_i^2}{N}=20.0,\qquad \langle xy\rangle=\frac{\sum_i x_iy_i}{N}=38.119,\qquad \langle y^2\rangle=\frac{\sum_i y_i^2}{N}=73.650.$$

MAP + Laplace

The unnormalized log-posterior is

$$\log\tilde{P}=-\frac{N+2}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}\sum_i\big[y_i-(ax_i+c)\big]^2.$$

Equating the first derivative with 0 results in the following equations

$$\begin{aligned}
0=\partial_a\log\tilde{P} &= \frac{1}{\sigma^2}\sum_i x_i\big[y_i-(ax_i+c)\big]\\
0=\partial_c\log\tilde{P} &= \frac{1}{\sigma^2}\sum_i \big[y_i-(ax_i+c)\big]\\
0=\partial_{\sigma^2}\log\tilde{P} &= -\frac{N+2}{2}\frac{1}{\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_i\big[y_i-(ax_i+c)\big]^2,
\end{aligned}$$

which we can solve to get the MAP estimators

$$\begin{aligned}
a_{\text{MAP}} &= \frac{\langle xy\rangle-\langle x\rangle\langle y\rangle}{\langle x^2\rangle-\langle x\rangle^2}=1.4825\\
c_{\text{MAP}} &= \langle y\rangle-\langle x\rangle a_{\text{MAP}}=2.1172\\
(\sigma^2)_{\text{MAP}} &= \frac{1}{N+2}\sum_i\left[y_i-(a_{\text{MAP}}x_i+c_{\text{MAP}})\right]^2=0.07971
\end{aligned}$$

The second derivative of the log-posterior at the MAP point is

$$\begin{aligned}
\partial_a\partial_a\log\tilde{P}\big|_{\text{MAP}} &= -\frac{1}{\sigma^2_{\text{MAP}}}\sum_i x_i^2=-\frac{N\langle x^2\rangle}{\sigma^2_{\text{MAP}}}\\
\partial_a\partial_c\log\tilde{P}\big|_{\text{MAP}} &= -\frac{1}{\sigma^2_{\text{MAP}}}\sum_i x_i=-\frac{N\langle x\rangle}{\sigma^2_{\text{MAP}}}\\
\partial_c\partial_c\log\tilde{P}\big|_{\text{MAP}} &= -\frac{1}{\sigma^2_{\text{MAP}}}\sum_i 1=-\frac{N}{\sigma^2_{\text{MAP}}}\\
\partial_{\sigma^2}\partial_{\sigma^2}\log\tilde{P}\big|_{\text{MAP}} &= \frac{N+2}{2}\frac{1}{(\sigma^2_{\text{MAP}})^2}-\frac{\sum_i[y_i-(a_{\text{MAP}}x_i+c_{\text{MAP}})]^2}{(\sigma^2_{\text{MAP}})^3}=-\frac{N+2}{2(\sigma^2_{\text{MAP}})^2}
\end{aligned}$$

$$\partial_{\sigma^2}\partial_a \log\tilde{P}\big|_{\rm MAP} = -\frac{1}{(\sigma^2_{\rm MAP})^2}\sum_i x_i\big[y_i-(ax_i+c)\big]=0$$
$$\partial_{\sigma^2}\partial_c \log\tilde{P}\big|_{\rm MAP} = -\frac{1}{(\sigma^2_{\rm MAP})^2}\sum_i \big[y_i-(ax_i+c)\big]=0$$

where the last two mixed derivatives are zero because they are proportional to the first derivatives with respect to a and c, which are zero at the MAP point. Now, we can construct the Hessian, and take the negative inverse to obtain the Laplace covariance matrix

$$\nabla\nabla\log\tilde{P} = -\frac{N}{\sigma^2_{\rm MAP}}\begin{bmatrix}\langle x^2\rangle & \langle x\rangle & 0\\ \langle x\rangle & 1 & 0\\ 0 & 0 & \frac{N+2}{2N\sigma^2_{\rm MAP}}\end{bmatrix}$$
$$\mathrm{Cov_L}\begin{bmatrix}a\\c\\\sigma^2\end{bmatrix} = \big[-\nabla\nabla\log\tilde{P}\big]^{-1} = \frac{\sigma^2_{\rm MAP}}{N}\begin{bmatrix}\frac{1}{\langle x^2\rangle-\langle x\rangle^2} & -\frac{\langle x\rangle}{\langle x^2\rangle-\langle x\rangle^2} & 0\\ -\frac{\langle x\rangle}{\langle x^2\rangle-\langle x\rangle^2} & \frac{\langle x^2\rangle}{\langle x^2\rangle-\langle x\rangle^2} & 0\\ 0 & 0 & \frac{2N\sigma^2_{\rm MAP}}{N+2}\end{bmatrix},$$

which provides the Laplace standard deviations and correlations

$$\mathrm{std_L}(a) = \sqrt{\frac{\sigma^2_{\rm MAP}}{N}\frac{1}{\langle x^2\rangle-\langle x\rangle^2}} = 0.0534$$
$$\mathrm{std_L}(c) = \sqrt{\frac{\sigma^2_{\rm MAP}}{N}\frac{\langle x^2\rangle}{\langle x^2\rangle-\langle x\rangle^2}} = 0.2386$$
$$\mathrm{corr_L}(a,c) = \frac{\mathrm{cov_L}(a,c)}{\sqrt{\mathrm{std_L}(a)\mathrm{std_L}(c)}} = -\frac{\langle x\rangle}{\sqrt{\langle x^2\rangle}} = -0.89443$$
$$\mathrm{std}_L(\sigma^2) = \sqrt{\frac{2}{N+2}}\sigma^2_{\rm MAP} = 0.0376$$

MCMC

Let's construct the STAN code for the model.

```
data {
    int<lower=1> N;
    real x[N];
    real y[N];
}
parameters {
    real a;
    real c;
    real<lower=0> sigma2;
}
model {
    target += -log(sigma2);
    for (i in 1:N){
        target += normal_lpdf(y[i] | a * x[i] + c, sqrt(sigma2));
    }
}
```

Running sampling with STAN's NUTS algorithm over 4 chains, each with 100,000 iterations, the first half of each chain is dropped, and every 10th (thin=10) of the remaining constitutes 20,000 MCMC samples yielding

the following approximations of means, medians, standard deviations and correlation of a, c and σ^2:

$$\begin{aligned}
&\mathbb{E}(a) \approx 1.482, \qquad \mathrm{median}(a) \approx 1.482,\\
&\mathbb{E}(c) \approx 2.119, \qquad \mathrm{median}(c) \approx 2.120\\
&\mathbb{E}(\sigma^2) \approx 0.2377, \qquad \mathrm{median}(\sigma^2) \approx 0.1639\\
&\mathrm{std}(a) \approx 0.0920, \qquad \mathrm{std}(c) \approx 0.4130, \qquad \mathrm{std}(\sigma^2) \approx 0.2828\\
&\mathrm{corr}(a,c) = \frac{\mathrm{cov}(a,c)}{\sqrt{\mathrm{std}(a)\mathrm{std}(c)}} \approx -0.8908, \qquad \mathrm{corr}(a,\sigma^2) \approx 0.03771, \qquad \mathrm{corr}(c,\sigma^2) \approx -0.03118
\end{aligned}$$

Exact

The model can be cast into the general form of linear regression (See Appendix) with the following definitions,

$$K = 2, \qquad b = \begin{bmatrix} a \\ c \end{bmatrix}, \qquad X = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}, \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}.$$

After plugging in the data $\{x_i\}$, $\{y_i\}$, we can calculate the following intermediate results

$$\begin{aligned}
\Lambda &= \left[\frac{X^\top X}{N}\right]^{-1} = \begin{bmatrix} 0.25 & -1.00 \\ -1.00 & 5.00 \end{bmatrix},\\
\bar{b} &= \begin{bmatrix} \bar{a} \\ \bar{c} \end{bmatrix} = \frac{\Lambda X^\top y}{N} = \begin{bmatrix} 1.4825 \\ 2.1171 \end{bmatrix},\\
\bar{v} &= \frac{(y - X\bar{b})^\top (y - X\bar{b})}{N} = 0.10248.
\end{aligned}$$

Expectation values are

$$\begin{aligned}
&\mathbb{E}(a) = \mathrm{mode}(a) = \mathrm{media}(a) = \bar{a} = 1.4825\\
&\mathbb{E}(c) = \mathrm{mode}(c) = \mathrm{media}(c) = \bar{c} = 2.1171\\
&\mathbb{E}(\sigma^2) = \frac{N\bar{v}}{N-K-2} = 0.23912, \qquad \mathrm{mode}(\sigma^2) = \frac{N\bar{v}}{N-K+3} = 0.08967
\end{aligned}$$

Variance of b, standard deviation of σ^2 are

$$\mathrm{Var}(b) = \frac{\bar{v}\Lambda}{N-K-2} = \begin{bmatrix} 0.00855 & -0.03416 \\ -0.03416 & 0.17080 \end{bmatrix}, \qquad \mathrm{std}(\sigma^2) = \frac{\sqrt{2}N\bar{v}}{(N-K-2)\sqrt{N-K-4}} = 0.33817,$$

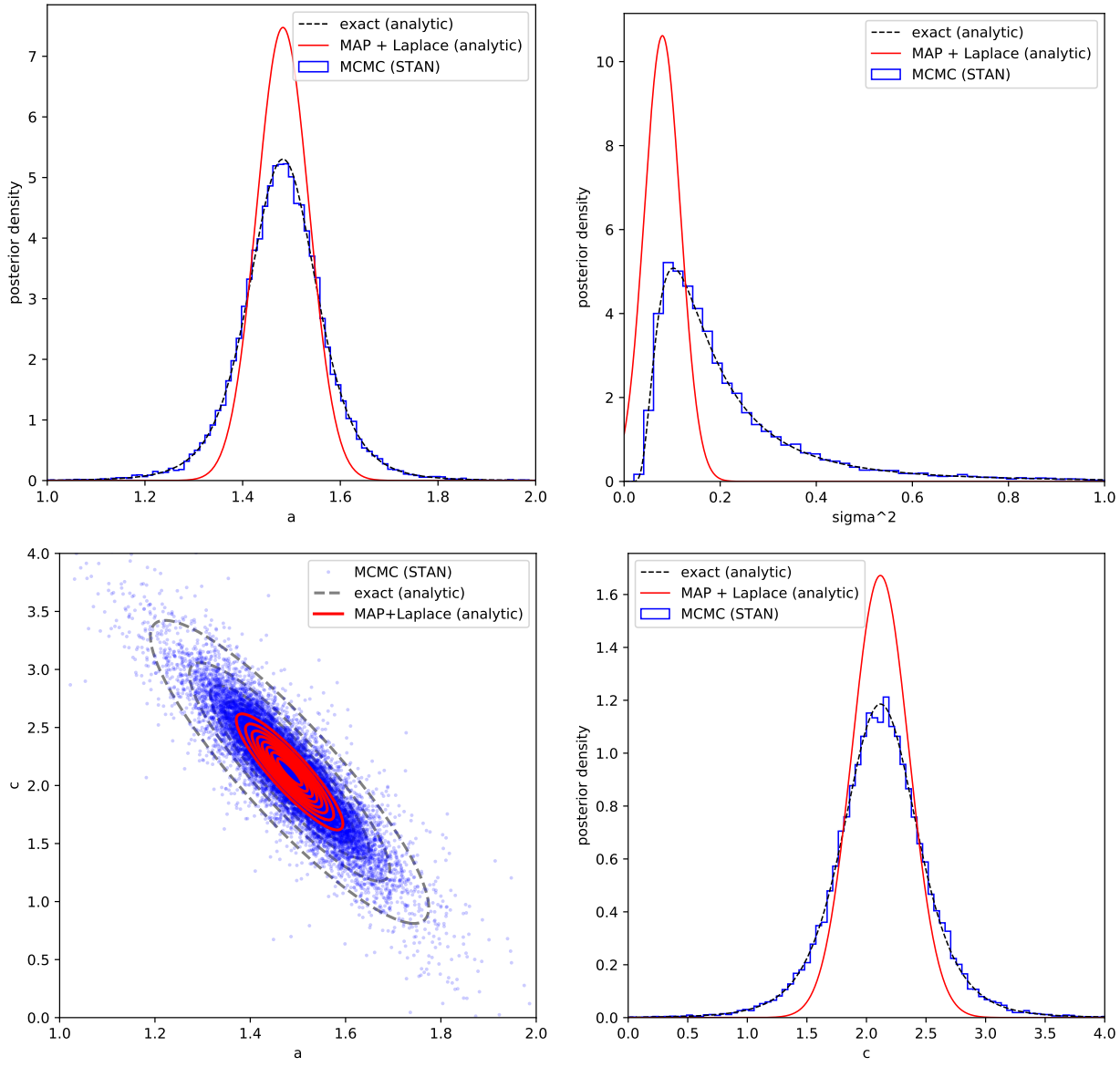
and the standard deviation and correlation of a and c are

$$\mathrm{std}(a) = \sqrt{[\mathrm{Var}(b)]_{1,1}} = 0.09241, \qquad \mathrm{std}(c) = \sqrt{[\mathrm{Var}(b)]_{2,2}} = 0.41328, \qquad \mathrm{corr}(a,c) = \frac{\Lambda_{1,2}}{\sqrt{\Lambda_{1,1}\Lambda_{2,2}}} = -0.89443,$$

and the correlation between σ^2 and a or c is exactly 0, as for all linear regression models with assumed uniform σ^2.

Comparison

We plot the marginals of a, c and σ^2, and the joint marginal of (a, c). MAP + Laplace underestimates σ^2 and the uncertainties of a and c.



Solution 2 - Negative binomial

We assume that the number of failures r is picked before all experiments, it is not generated by a probabilistic distribution. The only data is k, the number of successes, which is generated by a negative binomial distribution,

$$P(k\,|\,p) = \binom{k+r-1}{k} p^k (1-p)^r.$$

Let's use a non-informative prior for p:

$$P_0(p) = \frac{\text{const.}}{p(1-p)},$$

where const. is the appropriate normalization factor in a plausible region $0 < p_{\text{low}} \leq p < p_{\text{high}} < 1$. Now, the posterior of p can be written as

$$P(p\,|\,k) \quad = \quad \text{const.} \times P_0(p) \times P(k\,|\,p) = \frac{1}{Z}\, p^{k-1}\,(1-p)^{r-1}.$$

The unnormalized log-posterior is

$$\log \tilde{P}(p\,|\,k) = (k-1)\log(p) + (r-1)\log(1-p).$$

MAP + Laplace

Equating the first derivative of the log-posterior with zero yields the MAP estimate:

$$0 = \partial_p \log \tilde{P} = \frac{k-1}{p} - \frac{r-1}{1-p} \qquad \Rightarrow \qquad p_{\text{MAP}} = \frac{k-1}{k+r-2} = 0.83333$$

The second derivative of, at the MAP point, can be used to calculate the Laplace standard deviation:

$$\partial_p^2 \log \tilde{P}\big|_{\text{MAP}} = -\frac{k-1}{p_{\text{MAP}}^2} - \frac{r-1}{(1-p_{\text{MAP}})^2} \qquad \Rightarrow \qquad \text{std}_{\text{L}}(p) = \sqrt{-\left[\partial_p^2 \log \tilde{P}\big|_{\text{MAP}}\right]^{-1}} = 0.07607$$

MCMC

Let's construct the STAN code for the model.

```
1  data {
2      int<lower=1> r;  // number of failures
3      int<lower=1> k;  // number of successes
4  }
5  parameters {
6      real<lower=0, upper=1> p;  // success probability
7  }
8  model {
9      target += (k - 1) * log(p) + (r - 1) * log(1 - p);  // log-posterior
10 }
```

Although r is not a probabilistic variable, it still works as an input to STAN, so it needs to be listed under "data" (line 2). We set a minimal allowed value for k to 1 (line 3), because otherwise the posterior is not normalizable over $0 < p < 1$, which throws STAN off. Line 9 defines the unnormalized log-posterior.

After sampling with STAN's NUTS algorithm, 4 chains, 2000 iterations each, dropping the first 1000, and not thinning the remaining (thin = 1), we obtain the following mean, median and standard deviation estimates

$$\begin{aligned} \mathbb{E}(p) &\approx 0.80759 \\ \text{median}(p) &\approx 0.81525 \\ \text{std}(p) &\approx 0.07577 \end{aligned}$$

Exact

We apply the limits $p_{\text{low}} \to 0$, $p_{\text{high}} \to 1$. The analytical formula for the normalization constant Z can be found by integrating over $0 < p < 1$:

$$Z = \int_0^1 dp\, p^{k-1}(1-p)^{r-1} = \mathcal{B}(k,r) = \frac{\Gamma(k)\Gamma(r)}{\Gamma(k+r)} = \frac{(k-1)!(r-1)!}{(k+r-2)!},$$

where $\mathcal{B}$ is the beta function. This allows us to write the posterior as

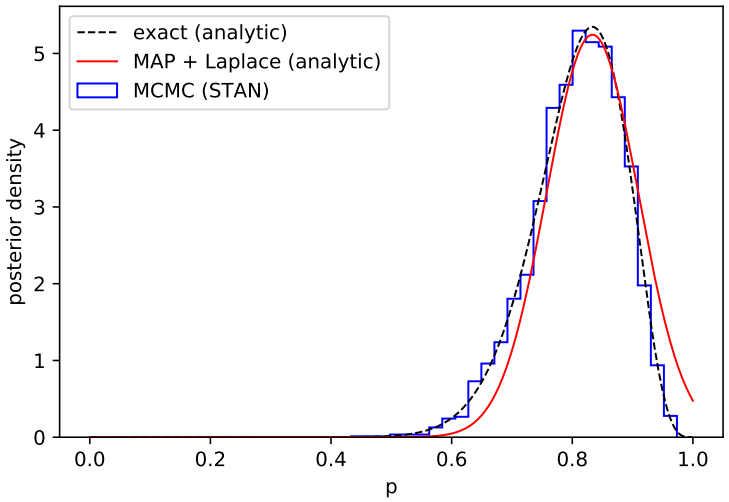
$$P(p\,|\,k) = \frac{p^{k-1}p^{r-1}}{\mathcal{B}(k,r)} =: \frac{p^{\alpha-1}p^{\beta-1}}{\mathcal{B}(\alpha,\beta)} = \text{Beta}(p\,|\,\alpha,\beta),$$

which is the beta distribution for $\alpha = k$ and $\beta = r$. The formulas for mean, median, mode, and standard deviation are known:

$$\begin{aligned}
\mathbb{E}(p) &= \frac{\alpha}{\alpha+\beta} = \frac{k}{k+r} = 0.80769 \\
\text{median}(p) &\approx \frac{\alpha - \frac{1}{3}}{\alpha+\beta-\frac{2}{3}} = 0.81579 \\
\text{mode}(p) &= \frac{\alpha-1}{\alpha+\beta-2} = 0.83333 \\
\text{std}(p) &= \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} = 0.07585
\end{aligned}$$

Comparison

MAP + Laplace method approximates the posterior with a Gaussian centered at the mode. Our MCMC run produced 4000 samples, the histogram of which can be plotted. The exact solution resulted in the beta distribution, which is readily implemented in Python's `scipy.stats.beta`. Let's compare their estimates graphically.



Solution 3 - Normal

From the data $\{x_i\}_{i=1}^N$, let's calculate two important statistics, the empirical mean and empirical variance:

$$N = 7, \qquad \bar{x} = \frac{\sum_i x_i}{N} = 114.71, \qquad \bar{v} = \frac{\sum_i (x_i - \bar{x})^2}{N} = 103.35$$

Each data point is assumed to come from a normal distribution, independent from the other data points. The resulting likelihood is

$$P(\{x_i\} \,|\, \mu, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] = \left[2\pi\sigma^2\right]^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^N (x_i-\mu)^2\right]$$

We use a non-informative prior for μ and σ^2,

$$P_0(\mu, \sigma^2) = P_0(\mu) \times P_0(\sigma^2) = \text{const.} \times \frac{\text{const.}}{\sigma^2} = \frac{\text{const.}}{\sigma^2},$$

where the constants are appropriately chosen to normalize the prior distribution on plausible a interval, $\mu_{\text{low}} \le \mu \le \mu_{\text{high}}$, and $0 < \sigma^2_{\text{low}} \le \sigma^2 \le \sigma^2_{\text{high}}$.

The posterior can be written as

$$P(\mu, \sigma^2 \,|\, \{x_i\}) = \text{const.} \times P_0(\mu, \sigma^2) \times P(\{x_i\} \,|\, \mu, \sigma^2) = \frac{1}{Z}(\sigma^2)^{-\frac{N+2}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^N (x_i-\mu)^2\right],$$

and the unnormalized log-posterior is

$$\log \tilde{P}(\mu, \sigma^2 \,|\, \{x_i\}) = -\frac{N+2}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2.$$

MAP + Laplace

Setting the first derivatives of the log-posterior to zero yields the MAP estimates:

$$0 \;=\; \partial_\mu \log \tilde{P} = \frac{1}{\sigma^2}\sum_i (x_i - \mu) \qquad \Rightarrow \quad \mu_{\text{MAP}} = \frac{\sum_i x_i}{N} = \bar{x} = 114.71$$

$$0 \;=\; \partial_{\sigma^2} \log \tilde{P} = -\frac{N+2}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (x_i-\mu)^2 \qquad \Rightarrow \quad (\sigma^2)_{\text{MAP}} = \frac{\sum_i (x_i - \mu_{\text{MAP}})^2}{N+2} = \frac{N\bar{v}}{N+2} = 80.381$$

The second-oder derivatives at the MAP point are

$$\begin{aligned}
\partial_\mu \partial_\mu \log \tilde{P}\big|_{\text{MAP}} &= -\frac{N}{\sigma^2}\bigg|_{\text{MAP}} = -\frac{N+2}{\bar{v}} \\
\partial_\mu \partial_{\sigma^2} \log \tilde{P}\big|_{\text{MAP}} &= -\frac{1}{\sigma^2}\sum_i (x_i-\mu)\bigg|_{\text{MAP}} = 0 \\
\partial_{\sigma^2} \partial_{\sigma^2} \log \tilde{P}\big|_{\text{MAP}} &= \frac{N+2}{2(\sigma^2)^2} - \frac{\sum_i (x_i-\mu)^2}{(\sigma^2)^3}\bigg|_{\text{MAP}} = \frac{N+2}{2(\sigma^2_{\text{MAP}})^2} - \frac{N\bar{v}}{(\sigma^2_{\text{MAP}})^3} = -\frac{(N+2)^3}{2(N\bar{v})^2}
\end{aligned}$$

The Laplace covariance matrix of (μ, σ^2) is the negative inverse of the Hessian at the MAP point:

$$\text{Cov}_{\text{L}}\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \left[-\nabla\nabla \log \tilde{P}\big|_{\text{MAP}}\right]^{-1} = \begin{bmatrix} \frac{N+2}{\bar{v}} & 0 \\ 0 & \frac{(N+2)^3}{2(N\bar{v})^2} \end{bmatrix}^{-1} = \begin{bmatrix} \frac{\bar{v}}{N+2} & 0 \\ 0 & \frac{2(N\bar{v})^2}{(N+2)^3} \end{bmatrix},$$

from which, we can extract the formulas for the Laplace standard deviations, and correlation

$$\mathrm{std_L}(\mu) = \sqrt{\frac{\bar{v}}{N+2}} = 3.3887, \quad \mathrm{std_L}(\sigma^2) = \sqrt{\frac{2}{N+2}}\frac{N\bar{v}}{N+2} = 37.892, \quad \mathrm{corr_L}(\mu,\sigma^2) = \frac{\mathrm{cov_L}(\mu,\sigma^2)}{\mathrm{std_L}(\mu)\,\mathrm{std_L}(\sigma^2)} = 0.$$

To estimate the probability that $\mu < 100$, we recall that the Laplace approximation assumes a Gaussian shape for the posterior, centered at the MAP point. For μ, this is

$$P(\mu\,|\,\{x_i\}) \approx \mathrm{Norm}(\mu\,|\,\mu_{\mathrm{MAP}}, \mathrm{std_L}(\mu)) = \mathrm{Norm}(\mu\,|\,114.71,\, 3.3887),$$

and the probability of $\mu < 100$ is estimated as

$$P(\mu < 100\,|\,\{x_i\}) = \int_{-\infty}^{100} d\mu\; P(\mu\,|\,\{x_i\}) \;\approx\; \text{cdf-Norm}(100\,|\,114.71,\, 3.3887) = 7\times10^{-6},$$

where cdf-Norm is the cumulative distribution function of a the normal distribution, readily implemented in Python's `scipy.stats.norm` as `cdf`.

MCMC

Let's construct the STAN code for the model.

```
1  data {
2      int<lower=1> N;
3      real x[N];
4  }
5  parameters {
6      real mu;
7      real<lower=0> sigma2;
8  }
9  model {
10     target += -log(sigma2);
11     for (i in 1:N){
12         target += normal_lpdf(x[i] | mu, sqrt(sigma2));
13     }
14 }
```

The "data" section includes N in line 2, because STAN needs to know how many x_i values to expect. In line 11, we loop over the input x_i values and add the log-probability contribution of each, which is readily implemented in STAN's `normal_lpdf(x | mu, sigma)` function (See section 54.1 in STAN manual).

Running sampling with STAN's NUTS algorithm over 4 chains, each with 100,000 iterations, the first half of each chain is dropped, and every 10th (thin=10) of the remaining constitutes 20,000 MCMC samples yielding the following approximates of means, medians, standard deviations and correlation of μ and σ^2:

$$\begin{aligned}
&\mathbb{E}(\mu) \approx 114.69, \qquad \mathrm{median}(\mu) \approx 114.69, \qquad \mathrm{std}(\mu) \approx 5.0706,\\
&\mathbb{E}(\sigma^2) \approx 179.39, \qquad \mathrm{median}(\sigma^2) \approx 135.87, \qquad \mathrm{std}(\sigma^2) \approx 164.42,\\
&\mathrm{corr}(\mu,\sigma^2) = \frac{\mathrm{cov}(\mu,\sigma^2)}{\mathrm{std}(\mu)\mathrm{std}(\sigma^2)} \approx 0.014.
\end{aligned}$$

The probability of $\mu < 100$, can be estimated by counting what fraction of the MCMC samples $\{\mu^{(t)} : t = 1, 2, \ldots T\}$) is below 100, this turned out to be

$$P(\mu < 100) \approx \frac{\#\left[\mu^{(t)} < 100\right]}{T} = \frac{117}{20000} = 5.9\times10^{-3}.$$

Exact

The analytical formula for the normalization constant Z can be calculated with the integral over $\mu \in (-\infty, +\infty)$, $\sigma^2 \in [0, \infty)$:

$$Z = \int_0^\infty d\sigma^2 \; (\sigma^2)^{-\frac{N+2}{2}} \int_{-\infty}^{+\infty} d\mu \; \exp\left[-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\right],$$

where the μ-integral can be written as

$$\int_{-\infty}^{+\infty} d\mu \; \exp\left[-\frac{N\mu^2 - 2\mu\sum_i x_i + \sum_i x_i^2}{2\sigma^2}\right] = \exp\left[-\frac{N\bar{v}}{2\sigma^2}\right] \underbrace{\int_{-\infty}^{+\infty} d\mu \; \exp\left[-\frac{N(\mu - \bar{x})^2}{2\sigma^2}\right]}_{\sqrt{2\pi\sigma^2/N}},$$

and Z can be written as

$$Z = \sqrt{\frac{2\pi}{N}} \int_0^\infty d\sigma^2 \; (\sigma^2)^{-\frac{N+1}{2}} \exp\left[-\frac{N\bar{v}}{2\sigma^2}\right] = \sqrt{\frac{2\pi}{N}} \left(\frac{2}{N\bar{v}}\right)^{\frac{N-1}{2}} \underbrace{\int_0^\infty dz \; z^{\frac{N-3}{2}} \exp(-z)}_{\Gamma\left(\frac{N-1}{2}\right)},$$

where we introduced a new integration variable $z = \frac{N\bar{v}}{2\sigma^2}$, and substituted $\sigma^2 = \frac{N\bar{v}}{2z}$ and $d\sigma^2 = -\frac{N\bar{v}}{2z^2} dz$, and got rid of the minus sign by flipping the integration limits. Now we can write the normalized posterior as

$$P(\mu, \sigma^2 \,|\, \{x_i\}) = \frac{\sqrt{N}}{\sqrt{2\pi\sigma^2}} \frac{1}{\Gamma\left(\frac{N-1}{2}\right)} \left(\frac{N\bar{v}}{2}\right)^{\frac{N-1}{2}} (\sigma^2)^{-\frac{N-1}{2}-1} \exp\left[-\frac{2\left(\frac{N\bar{v}}{2}\right) + N(\mu - \bar{x})^2}{2\sigma^2}\right],$$

which we can identify with the normal-inverse-gamma distribution,

$$\text{N-IG}(\mu, \sigma^2 \,|\, \mu_c, \lambda, \alpha, \beta) = \frac{\sqrt{\lambda}}{\sqrt{2\pi\sigma^2}} \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} \exp\left[-\frac{2\beta + \lambda((\mu - \mu_c)^2)}{2\sigma^2}\right],$$

with parameters $\mu_c = \bar{x}$, $\lambda = N$, $\alpha = \frac{N-1}{2}$ and $\beta = \frac{N\bar{v}}{2}$. The mean, mode, standard deviation and correlation are known

$$\begin{aligned}
&\mathbb{E}(\mu) = \mu_c = \bar{x} = 114.71, \qquad \text{mode}(\mu) = \mu_c = \bar{x} = 114.71, \\
&\text{std}(\mu) = \sqrt{\frac{\beta}{\lambda(\alpha - 1)}} = \sqrt{\frac{\bar{v}}{N - 3}} = 5.0830, \\
&\mathbb{E}(\sigma^2) = \frac{\beta}{\alpha - 1} = \frac{N\bar{v}}{N - 3} = 180.86, \qquad \text{mode}(\sigma^2) = \frac{\beta}{\alpha + \frac{3}{2}} = \frac{N\bar{v}}{N + 2} = 80.381, \\
&\text{std}(\sigma^2) = \frac{\beta}{(\alpha - 1)\sqrt{\alpha - 2}} = \frac{\sqrt{2} N\bar{v}}{(N - 3)\sqrt{N - 5}} = 144.69, \\
&\text{corr}(\mu, \sigma^2) = \frac{\text{cov}(\mu, \sigma^2)}{\text{std}(\mu)\text{std}(\sigma^2)} = 0.
\end{aligned}$$

To calculate the probability that $\mu < 100$, we need to find the marginal posterior distribution of μ:

$$P(\mu \,|\, \{x_i\}) = \int_0^\infty d\sigma^2 \; P(\mu, \sigma^2 \,|\, \{x_i\}) = \sqrt{\frac{N}{2\pi}} \frac{\left(\frac{N\bar{v}}{2}\right)^{\frac{N-1}{2}}}{\Gamma\left(\frac{N-1}{2}\right)} \int_0^\infty d\sigma^2 \; (\sigma^2)^{-\frac{N+2}{2}} \exp\left[-\frac{N\bar{v} + N(\mu - \bar{x})^2}{2\sigma^2}\right]$$

The integral over σ^2 can be evaluated by introducing a new variable $z = \frac{N\bar{v}+N(\mu-\bar{x})^2}{2\sigma^2}$, giving

$$\int_0^\infty d\sigma^2\,[\ldots] = \left[\frac{2}{N(\bar{v}+(\mu-\bar{x})^2)}\right]^{\frac{N}{2}} \underbrace{\int_0^\infty dz\; z^{\frac{N}{2}-1}\exp(-z)}_{\Gamma(N/2)},$$

resulting in

$$P(\mu\,|\,\{x_i\}) = \frac{1}{\sqrt{\pi\bar{v}}}\frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}\left[1+\frac{(\mu-\bar{x})^2}{\bar{v}}\right]^{-\frac{N}{2}} =: \frac{1}{\sqrt{\nu\pi s^2}}\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{(\mu-\mu_c)^2}{\nu s^2}\right]^{-\frac{\nu+1}{2}}$$

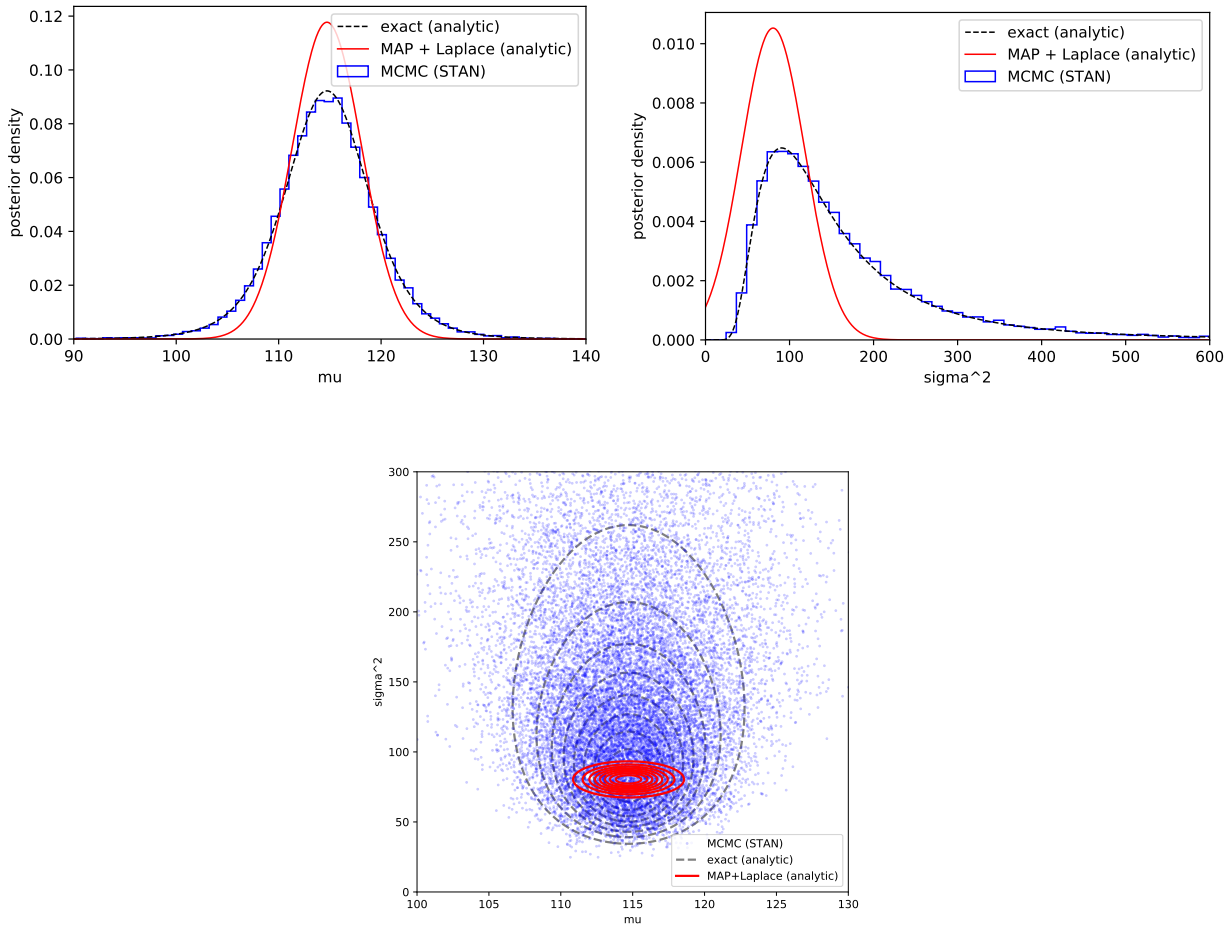
which is a (shifted and scaled) t-distribution with parameters $\nu = N-1$, $\mu_c = \bar{x}$ and $s = \sqrt{\bar{v}/(N-1)}$. The probability in question can now be directly calculated as the cumulative function of this t-distribution:

$$P(\mu < 100\,|\,\{x_i\}) \;=\; \text{cdf-t-dist}(100\,|\,\nu = N-1,\ \text{loc} = \bar{x},\ \text{scale} = \sqrt{\bar{v}/(N-1)}) \;=\; 6.069\times10^{-3}.$$

which is readily implemented in Python's `scipy.stats.t` as `cdf`.

Comparison

Let's plot the estimated posterior marginals of μ and σ^2, and the contour plot of their joint posterior.



Solution 4 - Fixed effect model

Yields $\{y_{g,i}\}$ are indexed by group, $g \in \{A, B, C\}$, and a sequential index within the group $i \in \{1, 2, \dots n_g\}$, where n_g is the number of values in group g. Let's define $G = \sum_g 1$, $N = \sum_g n_g$, i.e. the number of groups and the number of total data points, respectively. Each $y_{g,i}$ is assumed to be normally distributed around the corresponding group mean μ_g,

$$P(y_{g,i} \,|\, \mu_g, \sigma^2) = \mathrm{Norm}(y_{g,i} \,|\, \mu_g, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_{g,i} - \mu_g)^2}{2\sigma^2}\right],$$

the product of these terms give the full likelihood

$$P(\{y_{g,i}\} \,|\, \{\mu_g\}, \sigma^2) = \prod_g \prod_{i=1}^{n_g} \mathrm{Norm}(y_{g,i} \,|\, \mu_g, \sigma^2) = [2\pi\sigma^2]^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_g \sum_{i=1}^{n_g} (y_{g,i} - \mu_g)^2\right].$$

Using a non-informative prior,

$$P_0(\{\mu_g\}, \sigma^2) = P_0(\sigma^2) \prod_g P_0(\mu_g) = \frac{\text{const.}}{\sigma^2} \prod_g \text{const.} = \frac{\text{const.}}{\sigma^2},$$

results in the following posterior

$$P(\{\mu_g\}, \sigma^2 \,|\, \{y_{g,i}\}) = \text{const.} \times P_0(\{\mu_g\}, \sigma^2) \times P(\{y_{g,i}\} \,|\, \{\mu_g\}, \sigma^2) = \frac{1}{Z} (\sigma^2)^{-\frac{N+2}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_g \sum_{i=1}^{n_g} (y_{g,i} - \mu_g)^2\right].$$

Let's calculate the following statistics from the data

$$G = 3, \qquad N = 13, \qquad \{\bar{\mu}_g\} = \left\{\frac{1}{n_g} \sum_i y_{g,i}\right\} = \begin{bmatrix} \bar{\mu}_A \\ \bar{\mu}_B \\ \bar{\mu}_C \end{bmatrix} = \begin{bmatrix} 2507.5 \\ 1569.8 \\ 1337.7 \end{bmatrix}, \qquad \bar{v} = \frac{1}{N} \sum_g \sum_{i=1}^{n_g} (y_{g,i} - \bar{\mu}_g)^2 = 42817.0$$

MAP + Laplace

The log-posterior is

$$\log \tilde{P} = -\frac{N+2}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_g \sum_{i=1}^{n_g} (y_{g,i} - \mu_g)^2.$$

Setting the first derivatives to zero gives the MAP estimates

$$\begin{aligned}
0 = \partial_{\mu_g} \log \tilde{P} &= \frac{1}{\sigma^2} \sum_{i=1}^{n_g} (y_{g,i} - \mu_g) && \Rightarrow \quad (\mu_g)_{\mathrm{MAP}} = \frac{1}{n_g} \sum_{i=1}^{n_g} y_{g,i} = \bar{\mu}_g \qquad \forall g \\
0 = \partial_{\sigma^2} \log \tilde{P} &= -\frac{N+2}{2\sigma^2} + \frac{\sum_g \sum_{i=1}^{n_g} (y_{g,i} - \mu_g)^2}{2(\sigma^2)^2} && \Rightarrow \quad (\sigma^2)_{\mathrm{MAP}} = \frac{\sum_{g,i} \left[y_{g,i} - (\mu_g)_{\mathrm{MAP}}\right]^2}{N+2} = \frac{N\bar{v}}{N+2}
\end{aligned}$$

The second derivatives at the MAP point are

$$\begin{aligned}
\partial_{\mu_g} \partial_{\mu_g} \log \tilde{P}\big|_{\mathrm{MAP}} &= -\frac{n_g}{\sigma^2} \qquad \forall g \\
\partial_{\mu_g} \partial_{\mu_{g'}} \log \tilde{P}\Big|_{\mathrm{MAP}} &= 0 \qquad \text{if } g \neq g' \\
\partial_{\sigma^2} \partial_{\mu_g} \log \tilde{P}\big|_{\mathrm{MAP}} &= -\frac{1}{(\sigma^2)^2} \sum_i \left[y_{g,i} - (\mu_g)_{\mathrm{MAP}}\right] = 0 \\
\partial_{\sigma^2} \partial_{\sigma^2} \log \tilde{P}\big|_{\mathrm{MAP}} &= \frac{N+2}{2(\sigma^2)^2_{\mathrm{MAP}}} - \frac{\sum_{g,i} [y_{g,i} - (\mu_g)_{\mathrm{MAP}}]^2}{(\sigma^2)^3_{\mathrm{MAP}}} = -\frac{N+2}{2((\sigma^2)_{\mathrm{MAP}})^2} = -\frac{(N+2)^3}{2N^2\bar{v}^2},
\end{aligned}$$

Because the mixed derivatives are 0 at the MAP point, we can invert the Hessian ($\nabla\nabla \log \tilde{P}$) term by term, giving the following Laplace standard deviations

$$\mathrm{std_L}(\mu_g) = \sqrt{\frac{\bar{v}}{N+2}}, \qquad \mathrm{std_L}(\sigma^2) = \sqrt{\frac{2}{N+2}}\frac{N\bar{v}}{N+2}.$$

To summarize, the MAP + Laplace estimates are

$$\begin{aligned}
\mu_A: &\quad (\mu_A)_{\mathrm{MAP}} \pm \mathrm{std_L}(\mu_A) = 2507.5 \pm 96.3 \\
\mu_B: &\quad (\mu_B)_{\mathrm{MAP}} \pm \mathrm{std_L}(\mu_B) = 1569.8 \pm 78.6 \\
\mu_C: &\quad (\mu_C)_{\mathrm{MAP}} \pm \mathrm{std_L}(\mu_C) = 1337.7 \pm 111.2 \\
\sigma^2: &\quad (\sigma^2)_{\mathrm{MAP}} \pm \mathrm{std_L}(\sigma^2) = 37108 \pm 13549
\end{aligned}$$

MCMC

STAN can only handle rectangular arrays, and since $\{y_{g,i}\}$ cannot be reliably cast into such an array, we use an additional input: group labels, $l_j \in \{1, 2, 3\}$, signifying which group does the jth data point belong to. With this definition, our data is

$$\begin{aligned}
y = \{y_j : j = 1, 2, \ldots N\} &= [2604, 2665, 2251, 2510, 1559, 1729, 1866, 1414, 1159, 1692, 1528, 1444, 1041] \\
l = \{l_j : j = 1, 2, \ldots N\} &= [1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3].
\end{aligned}$$

This allows us to run the following STAN code.

```
data {
    int<lower=1> G;  // number of groups
    int<lower=1> N;  // number of total data points
    int<lower=1, upper=G> label[N];  // group label
    real y[N];  //data points
}
parameters {
    real mu[G];  // group means
    real<lower=0> sigma2;  // common variance
}
model {
    target += -log(sigma2);
    for (j in 1:N){
        target += normal_lpdf(y[j] | mu[label[j]], sqrt(sigma2));
    }
}
```

Running 4 chains with 2000 iterations each, dropping the first 1000, and keeping all remaining samples (thin=1), we find the following expectation values, medians and standard deviations

	$\mathbb{E}$	median	std
μ_A	2507	2510	131
μ_B	1570	1571	107
μ_C	1340	1339	149
σ^2	69490	59350	40100

Exact

The fixed effect model can be cast into the form of a linear regression with

$$
\begin{aligned}
K &= 3 \\
b &= [\mu_A, \mu_B, \mu_c]^\top \\
y &= [y_{A,1}, y_{A,2} \dots y_{A,n_A},\ y_{B,1}, y_{B,2} \dots y_{B,n_B},\ y_{C,1}, y_{C,2} \dots y_{C,n_C}]^\top \\
X^\top &= \begin{bmatrix} \underbrace{1 \quad 1 \quad \dots \quad 1}_{n_A} & & \\ & \underbrace{1 \quad 1 \quad \dots \quad 1}_{n_B} & \\ & & \underbrace{1 \quad 1 \quad \dots \quad 1}_{n_C} \end{bmatrix}
\end{aligned}
$$

where empty entries in X stand for zeros. Let's compute the relevant statistics

$$
\begin{aligned}
\Lambda &= \left(\frac{1}{N} X^\top X\right)^{-1} = \left(\frac{1}{N} \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}\right)^{-1} = \begin{bmatrix} \frac{N}{n_A} & & \\ & \frac{N}{n_B} & \\ & & \frac{N}{n_C} \end{bmatrix} \\
\bar{b} &= \frac{1}{N} \Lambda X^\top y = \begin{bmatrix} \sum_i y_{A,i}/n_A \\ \sum_i y_{B,i}/n_B \\ \sum_i y_{C,i}/n_C \end{bmatrix} = \begin{bmatrix} \bar{\mu}_A \\ \bar{\mu}_B \\ \bar{\mu}_C \end{bmatrix} \\
\bar{v} &= \frac{1}{N} (y - X\bar{b})^\top (y - X\bar{b}) = \frac{1}{N} \sum_g \sum_{i=1}^{n_g} (y_{g,i} - \bar{\mu}_g)^2
\end{aligned}
$$

Using the formulas in the Appendix, we can write them mean, mode and standard deviation of all μ_g and σ^2 as

$$
\mathbb{E}(b) = \text{median}(b) = \text{mode}(b) = \bar{b} = \begin{bmatrix} \bar{\mu}_A \\ \bar{\mu}_B \\ \bar{\mu}_C \end{bmatrix}, \qquad \text{Var}(b) = \frac{\bar{v}\Lambda}{N-K-2} = \frac{\bar{v}}{N-5} \begin{bmatrix} \frac{13}{4} & & \\ & \frac{13}{6} & \\ & & \frac{13}{3} \end{bmatrix}
$$

$$
\mathbb{E}(\sigma^2) = \frac{N\bar{v}}{N-K-2}, \qquad \text{mode}(\sigma^2) = \frac{N\bar{v}}{N-K+3}, \qquad \text{std}(\sigma^2) == \sqrt{\frac{2}{N-K-4}} \frac{N\bar{v}}{N-K-2}
$$

Numerical values are listed in the table below.

Comparison

Let's compare the summary statistics of the marginals from the three methods. (Note: MAP + Laplace does not estimate expectation value and median, MCMC does not directly give us mode estimates.)

	$\mathbb{E}$		median		mode		std		
	exact	MCMC	exact	MCMC	exact	Laplace	exact	MCMC	Laplace
μ_A	2508	2507	2508	2510	2508	2508	131.8	131.4	96.3
μ_B	1570	1570	1570	1571	1570	1570	107.7	107.0	78.6
μ_C	1338	1340	1338	1339	1338	1338	152.3	148.9	111.2
σ^2	69578	69488		59355	42817	37108	40170	40099	13550

Solution 5 - Random effect model

Similarly to the solution of the fixed effect model, we index the data points $\{y_{g,i}\}$ with group index $g \in \{A, B, C\}$ and individual index $i \in \{1, 2 \dots n_g\}$. Each value is assumed to come from the normal distribution centered at the corresponding group mean $\mu + \delta_g$,

$$P(y_{g,i} \,|\, \mu, \delta_g, \sigma^2) = \mathrm{Norm}(y_{g,i} \,|\, \mu + \delta_g,\, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_{g,i} - (\mu + \delta_g))^2}{2\sigma^2}\right].$$

If we knew the group means, the full likelihood would be written as

$$P(\{y_{g,i}\} \,|\, \mu, \{\delta_g\}, \sigma^2) = \prod_g \prod_{i=1}^{n_g} \mathrm{Norm}(y_{g,i} \,|\, \mu + \delta_g,\, \sigma^2) = [2\pi\sigma^2]^{-\frac{N}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_g \sum_{i=1}^{n_g} (y_{g,i} - (\mu + \delta_g))^2\right].$$

However, the model does not specify the $\{\delta_g\}$ values. Instead, it specifies that they are independently and identically distributed as

$$P(\{\delta_g\} \,|\, \sigma_\delta^2) = \prod_g \mathrm{Norm}(\delta_g \,|\, 0, \sigma_\delta^2) = [2\pi\sigma_\delta^2]^{-\frac{G}{2}} \exp\left[-\sum_g \frac{\delta_g^2}{2\sigma_\delta^2}\right],$$

where G is the number of groups ($G = 3$).

Now, the full likelihood for both the observed y values and the hidden δ values is

$$\begin{aligned} P(\{y_{g,i}\}, \{\delta_g\} \,|\, \mu, \sigma_\delta^2, \sigma^2) &= P(\{y_{g,i}\} \,|\, \mu, \{\delta_g\}, \sigma^2) \times P(\{\delta_g\} \,|\, \sigma_\delta^2) \\ &= [2\pi\sigma^2]^{-\frac{N}{2}} [2\pi\sigma_\delta^2]^{-\frac{G}{2}} \exp\left[-\frac{1}{2\sigma_\delta^2} \sum_g \delta_g^2 - \frac{1}{2\sigma^2} \sum_g \sum_{i=1}^{n_g} (y_{g,i} - (\mu + \delta_g))^2\right], \end{aligned}$$

from which all δ_g need to be marginalized out, to obtain the likelihood for the observed data,

$$\begin{aligned} P(\{y_{g,i}\} \,|\, \mu, \sigma_\delta^2, \sigma^2) &= \int d\{\delta_g\}\, P(\{y_{g,i}\}, \{\delta_g\} \,|\, \mu, \sigma_\delta^2, \sigma^2) \\ &= [2\pi\sigma^2]^{-\frac{N}{2}} [2\pi\sigma_\delta^2]^{-\frac{G}{2}} \prod_g \left\{ \int_{-\infty}^{+\infty} d\delta_g \, \exp\left[-\frac{\delta_g^2}{2\sigma_\delta^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n_g} (y_{g,i} - (\mu + \delta_g))^2\right] \right\} \end{aligned}$$

Let's use a non-informative prior

$$P(\mu, \sigma_\delta^2, \sigma^2) = \text{const.} \times \frac{\sigma_\delta^2 + \sigma^2}{\sigma_\delta^2 \sigma^2},$$

and write the posterior as

$$P(\mu, \sigma_\delta^2, \sigma^2 \,|\, \{y_{g,i}\}) = \frac{1}{Z} (\sigma_\delta^2 + \sigma^2)(\sigma^2)^{-\frac{N+2}{2}} (\sigma_\delta^2)^{-\frac{G+2}{2}} \prod_g \left\{ \int_{-\infty}^{+\infty} d\delta_g \, \exp\left[-\frac{\delta_g^2}{2\sigma_\delta^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n_g} (y_{g,i} - (\mu + \delta_g))^2\right] \right\}.$$

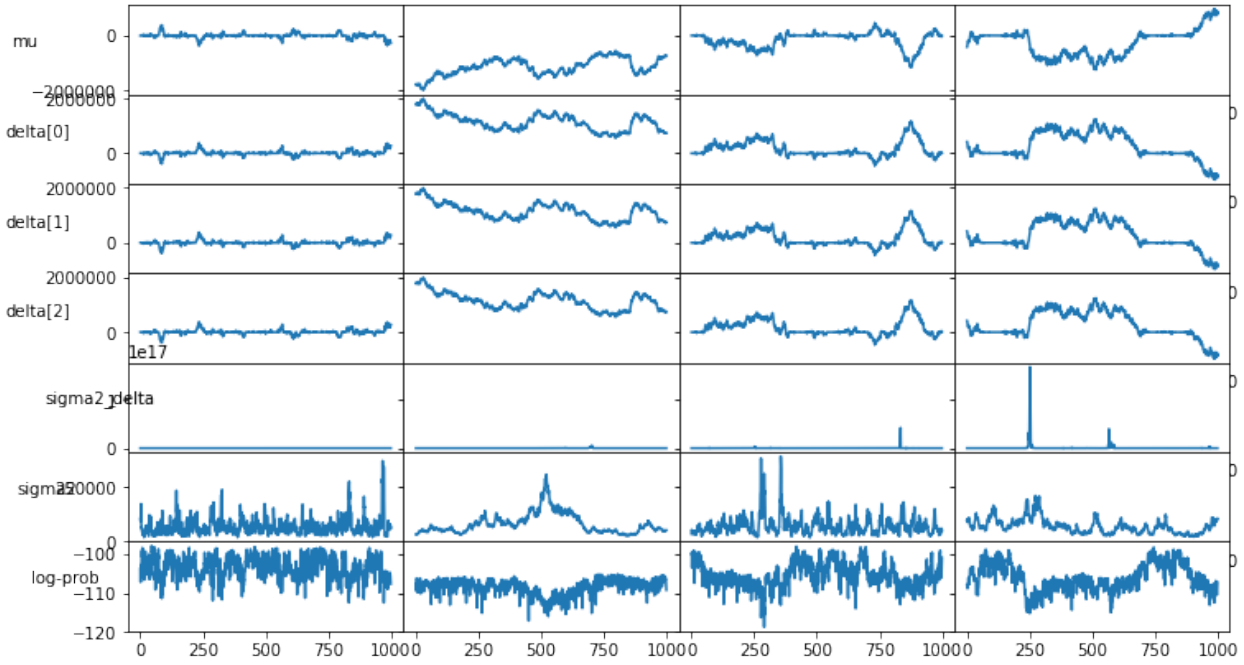
While the integral under the product can be evaluated analytically, to obtain the normalization constant Z, we would need to integrate the result with respect to μ, σ_δ^2 and σ^2. This cannot be done analytically. The MAP + Laplace estimate is possible but overly complicated.

MCMC

Our first attempt (which failed) was the following code

```
data {
    int<lower=1> G;  // number of groups
    int<lower=1> N;  // total number of data points
    int<lower=1, upper=G> label[N];  // group labels
    real y[N];
}
parameters{
    real mu;  // global mean
    real delta[G];  // group deviations
    real<lower=0> sigma2_delta;  // variance of group means
    real<lower=0> sigma2;  // within-group variance
}
model {
    target += log(sigma2_delta + sigma2) - log(sigma2_delta * sigma2); // prior
    for (g in 1:G){
        target += normal_lpdf(delta[g] | 0, sqrt(sigma2_delta));
    }
    for (j in 1:N){
        target += normal_lpdf(y[j] | mu + delta[label[j]], sqrt(sigma2));
    }
}
```

where we explicitly sample the hidden group deviations $\{\delta_g\}$. This is fine, because the means and variances of the MCMC samples will provide marginal values by default anyway. Since we don't have any alternative method to check our results, let's plot the traces after running 4 chains for 20,000 iterations, and keeping every 10th sample from the second half.



This is not every encouraging: 1) Some chains have significantly lower log-probability. 2) μ and δ_g values strongly anti-correlate, which is an indication of non-identifiability. 3) σ_δ^2 has incredibly large range, up to 10^{17} (!). All these are hallmarks of an unstable model.

To make our model more stable, we normalize the raw data, and define priors that are normalizable. This makes sure that STAN draws samples from realistic regions. To normalize the data, we calculate the empirical mean and standard deviation,

$$m := \frac{1}{N}\sum_g \sum_i y_{g,i}^{\text{raw}} = 1804.8, \qquad s := \sqrt{\frac{1}{N}\sum_g \sum_i (y_{g,i}^{\text{raw}} - m)^2} = 520.2,$$

and normalize all data points by subtracting the mean and dividing by the variance

$$y_{g,i} := \frac{y_{g,i}^{\text{raw}} - m}{s},$$

$\sim 95\%$ of which will in the range $[-2.0, 2.0]$. This allows us to define plausible regions for the parameters

$$\begin{aligned} \mu &= 0, \\ \Sigma := \sqrt{\sigma^2 + \sigma_\delta^2} &\sim [0.3, 3.0], \\ \gamma := \sigma_\delta / \Sigma &\sim [0, 1], \end{aligned}$$

where Σ^2 is the total variance, and γ is the fraction of the total variance contributed by σ_δ^2. We can still use a non-informative prior on γ:

$$P_0(\gamma) = \frac{\text{const.}}{\gamma(1-\gamma)},$$

but use well-defined priors for μ and Σ:

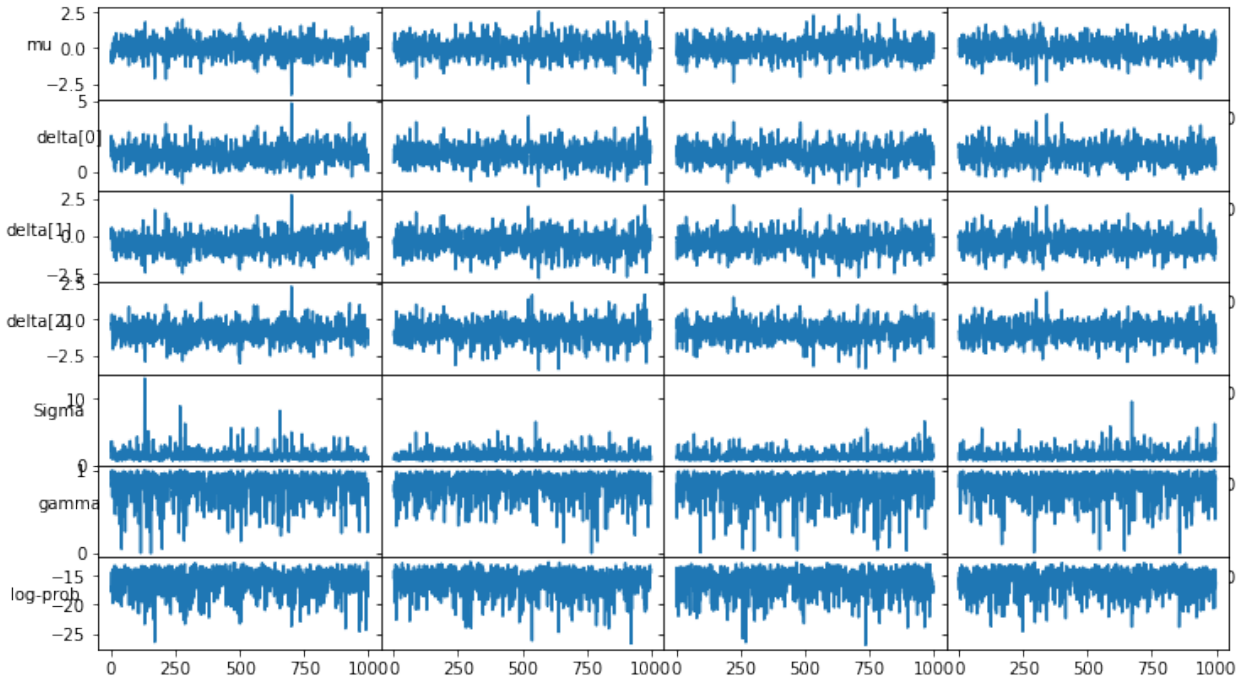
$$\begin{aligned} P_0(\mu) &= \text{Norm}(\mu \,|\, \text{mu}{=}0, \text{sigma}{=}1.0) \\ P_0(\Sigma) &= \text{Log-Norm}(\Sigma \,|\, \text{mu}{=}\ln(1.0), \text{sigma}{=}\frac{\ln(3.0) - \ln(0.3)}{2}) \end{aligned}$$

Using this new parametrization, we write run the following STAN code

```
data {
    int<lower=1> G;  // number of groups
    int<lower=1> N;  // total number of data points
    int<lower=1, upper=G> label[N];  // group labels
    real y[N];
}
parameters{
    real mu;  // global mean
    real delta[G];  // group deviations
    real<lower=0> Sigma;  // total variance
    real<lower=0, upper=1> gamma;  // sigma_delta^2 / Sigma^2
}
model {
    // priors
    target += normal_lpdf(mu | 0, 1.0);
    target += lognormal_lpdf(Sigma | 0, 1.151);
    target += -log(gamma) - log(1 - gamma);

    // likelihoods
    for (g in 1:G){
        target += normal_lpdf(delta[g] | 0, sqrt(gamma)*Sigma);
    }
    for (j in 1:N){
        target += normal_lpdf(y[j] | mu + delta[label[j]], sqrt(1-gamma)*Sigma);
    }
}
```

Running 4 chains, each with 20,000 iterations, out of which we keep only every 10th from the second half produces 4,000 samples. These are shown for the 4 chains below



All four chains show very similar behavior, sampled values are within realistic boundaries. STAN reports the following output

```
Inference for Stan model: anon_model_51faf35cd5be129564173392e7b29923.
4 chains, each with iter=20000; warmup=10000; thin=10;
post-warmup draws per chain=1000, total post-warmup draws=4000.

          mean se_mean   sd   2.5%   25%   50%   75% 97.5% n_eff Rhat
mu         9.5e-3 9.9e-3 0.59  -1.19 -0.36 7.3e-3  0.39  1.19  3575  1.0
delta[0]     1.25    0.01 0.64   0.04  0.83  1.23  1.65  2.54  3520  1.0
delta[1]    -0.43    0.01 0.61  -1.68  -0.8 -0.42 -0.05  0.79  3682  1.0
delta[2]    -0.83    0.01 0.64  -2.13 -1.22 -0.82 -0.43   0.4  3680  1.0
Sigma        1.44    0.01 0.75   0.71  0.98  1.24  1.65  3.34  4000  1.0
gamma        0.81 2.7e-3 0.16   0.38  0.73  0.85  0.92  0.98  3439  1.0
lp__        -16.05    0.03 2.08 -21.16 -17.1 -15.64 -14.54 -13.22  4000  1.0

Samples were drawn using NUTS at Sat Apr 28 17:55:08 2018.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Not forgetting that these estimates are valid for the normalized data, we can obtain the estimates for the unnormalized parameters as:

$$\mathbb{E}(\mu^{\text{raw}}) = m + s \times \mathbb{E}(\mu) = 1809.7, \qquad \text{std}(\mu^{\text{raw}}) = s \times \text{std}(\mu) = 308.5$$
$$\mathbb{E}(\sigma_\delta^{\text{raw}}) = s \times \mathbb{E}(\Sigma\sqrt{\gamma}) = 685.4, \qquad \text{std}(\sigma_\delta^{\text{raw}}) = s \times \text{std}(\Sigma\sqrt{\gamma}) = 409.1$$
$$\mathbb{E}(\sigma^{\text{raw}}) = s \times \mathbb{E}(\Sigma\sqrt{1-\gamma}) = 260.3, \qquad \text{std}(\sigma^{\text{raw}}) = s \times \text{std}(\Sigma\sqrt{1-\gamma}) = 71.2$$
$$\mathbb{E}(\sigma_\delta^2/(\sigma_\delta^2+\sigma^2)) = \mathbb{E}(\gamma) = 0.8056, \qquad \text{std}(\gamma) = 0.1595.$$

Appendix: Bayesian solution of linear regression

For N data points across K features, i.e. $\{([X_{n,k} \,:\, k = 1, 2, \dots K],\ y_n) \,:\, n = 1, 2, \dots N\}$, the linear regression model is defined as

$$y_n = \left(\sum_k X_{n,k} b_k \right) + \varepsilon_n, \qquad X:\ N \times K, \quad y:\ N \times 1, \quad b:\ K \times 1,$$

where $\varepsilon_n \sim \mathrm{Norm}(\cdot \,|\, 0, \sigma^2)$, independent of everything else.

The likelihood can be written as

$$\begin{aligned} P(y \,|\, X, b, \sigma^2) &= \prod_{n=1}^{N} P(y_n \,|\, X_{n,:}, b, \sigma^2) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_n - \sum_k X_{n,k} b_k)^2}{2\sigma^2} \right] \\ &= [2\pi\sigma^2]^{-\frac{N}{2}} \exp\left[-\frac{\sum_n (y_n - \sum_k X_{n,k} b_k)^2}{2\sigma^2} \right] = [2\pi\sigma^2]^{\frac{N}{2}} \exp\left[-\frac{(y - Xb)^\top (y - Xb)}{2\sigma^2} \right], \end{aligned}$$

where matrix-multiplication is assumed between every matrix, X, y, b, and $\top$ denotes transpose.

The numerator inside the exponential can be written as

$$\begin{aligned} (y - Xb)^\top (y - Xb) &= \Big(b^\top - y^\top X (X^\top X)^{-1} \Big) \underbrace{(X^\top X)}_{=: N\Lambda^{-1}} \Big(b - \underbrace{(X^\top X)^{-1} X^\top y}_{=: \bar{b}} \Big) + \underbrace{y^\top y - y^\top X (X^\top X)^{-1} X^\top y}_{=: N\bar{v}} \\ &= N \left[(b - \bar{b})^\top \Lambda^{-1} (b - \bar{b}) + \bar{v} \right], \end{aligned}$$

which shows that it is a good idea to start with computing $\Lambda, \bar{b}$ and $\bar{v}$:

$$\begin{aligned} \Lambda &= \left(\frac{1}{N} X^\top X \right)^{-1}, \\ \bar{b} &= \frac{1}{N} \Lambda X^\top y, \\ \bar{v} &= \frac{1}{N} (y - X\bar{b})^\top (y - X\bar{b}). \end{aligned}$$

Let's assume a non-informative prior for b and σ^2,

$$P_0(b, \sigma^2) = \frac{\text{const.}}{\sigma^2},$$

and write down the posterior

$$P(b, \sigma^2 \,|\, y, X) = \text{const.} \times P_0(b, \sigma^2) \times P(y \,|\, X, b, \sigma^2) = \frac{1}{Z} (\sigma^2)^{-\frac{N+2}{2}} \exp\left(-\frac{N\bar{v}}{2\sigma^2} \right) \exp\left[-\frac{N}{2\sigma^2} (b - \bar{b})^\top \Lambda^{-1} (b - \bar{b}) \right],$$

the normalization constant Z can be calculated by evaluating the integrals over b and σ^2.

$$\begin{aligned} Z &= \int_0^\infty d\sigma^2\ (\sigma^2)^{-\frac{N+2}{2}} \exp\left(-\frac{N\bar{v}}{2\sigma^2} \right) \underbrace{\int db\ \exp\left[-\frac{N}{2\sigma^2} (b - \bar{b})^\top \Lambda^{-1} (b - \bar{b}) \right]}_{\left[\frac{2\pi\sigma^2}{N}\right]^{K/2} \sqrt{\det \Lambda}} \\ &= \left[\frac{2\pi}{N} \right]^{\frac{K}{2}} \sqrt{\det \Lambda} \int_0^\infty d\sigma^2\ (\sigma^2)^{-\frac{N-K+2}{2}} \exp\left(-\frac{N\bar{v}}{2\sigma^2} \right) \\ &= \left[\frac{2\pi}{N} \right]^{\frac{K}{2}} \sqrt{\det \Lambda} \left[\frac{2}{N\bar{v}} \right]^{\frac{N-K}{2}} \underbrace{\int_0^\infty dz\ z^{\frac{N-K}{2} - 1} \exp(-z)}_{\Gamma\left(\frac{N-K}{2}\right)} \end{aligned}$$

Giving the normalized posterior

$$P(b,\sigma^2\,|\,y,X) = \frac{N^{\frac{K}{2}}}{[2\pi\sigma]^{\frac{K}{2}}\sqrt{\det\Lambda}}\frac{\left[\frac{N\bar{v}}{2}\right]^{\frac{N-K}{2}}}{\Gamma\left(\frac{N-K}{2}\right)}(\sigma^2)^{-\frac{N-K}{2}-1}\exp\left[-\frac{2\left(\frac{N\bar{v}}{2}\right)+(b-\bar{b})^\top(N\Lambda^{-1})(b-\bar{b})}{2\sigma^2}\right],$$

which is a multivariate normal-inverse-gamma distribution,

$$P(b\,|\,b_c,\alpha,\beta,V) = \frac{1}{\sqrt{\det V}}[2\pi]^{-\frac{K}{2}}\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-\frac{K}{2}-1}\exp\left[-\frac{2\beta+(b-b_c)^\top V^{-1}(b-b_c)}{2\sigma^2}\right],$$

with parameters $b_c=\bar{b}$, $\alpha=\frac{N-K}{2}$, $\beta=\frac{N\bar{v}}{2}$, $V=\Lambda/N$. Mean, mode, standard deviation and correlation are

$$\begin{aligned}
\mathbb{E}(b) &= \text{mode}(b)=\text{median}(b)=b_c=\bar{b}\\
\mathbb{E}(\sigma^2) &= \frac{\beta}{\alpha-1}=\frac{N\bar{v}}{N-K-2}, \qquad \text{mode}(\sigma^2)=\frac{\beta}{\alpha+1}=\frac{N\bar{v}}{N-K+3}\\
\text{Var}(b) &= \frac{\beta}{\alpha-1}V=\frac{\bar{v}}{N-K-2}\Lambda\\
\text{std}(\sigma^2) &= \frac{\beta}{(\alpha-1)\sqrt{\alpha-2}}=\frac{\sqrt{2}N\bar{v}}{(N-K-2)\sqrt{N-K-4}}\\
\text{cov}(b_k,\sigma^2) &= 0 \quad \forall k,
\end{aligned}$$

where $\text{Var}(b)$ is the full covariance matrix of $b=(b_1,b_2,\dots b_k)$.

The marginals of σ^2 and b are inverse-gamma and multivariate t-distribution, respectively:

$$\begin{aligned}
P(\sigma^2\,|\,y,X) &= \text{IG}(\sigma^2\,|\,\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right)\\
P(b\,|\,y,X) &= \text{t-dist}(b\,|\,\nu,b_c,\Sigma)=\frac{1}{[\pi\nu]^{K/2}\sqrt{\det\Sigma}}\frac{\Gamma\left(\frac{\nu+K}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\left[1+\frac{1}{\nu}(b-b_c)^\top\Sigma^{-1}(b-b_c)\right]^{-\frac{\nu+K}{2}},
\end{aligned}$$

with $\alpha=\frac{N-K}{2}$, $\beta=\frac{N\bar{v}}{2}$, $\nu=N-K$, $b_c=\bar{b}$, $\Sigma=\frac{\bar{v}}{N-K}\Lambda$. Mean, std of b_k and correlation between b_k and $b_{k'}$ are given by

$$\begin{aligned}
\mathbb{E}(b_k) &= \bar{b}_k\\
\text{std}(b_k) &= \sqrt{[\text{Var}(b)]_{k,k}}=\sqrt{\frac{\bar{v}}{N-K-2}\Lambda_{k,k}}\\
\text{corr}(b_k,b_{k'}) &= \frac{\text{cov}(b_k,b_{k'})}{\text{std}(b_k)\text{std}(b_{k'})}=\frac{\Lambda_{k,k'}}{\sqrt{\Lambda_{k,k}\Lambda_{k',k'}}}
\end{aligned}$$

Since $\Lambda=N(X^\top X)^{-1}$, the posterior correlation between the coefficients b_k is entirely determined by the combination of features present in the data, aka the design matrix X.

The marginal of b_k is a t-distribution with the same ν as the joint.

$$P(b_k\,|\,y,X)=\int db_{\neq k}\;P(b\,|\,y,X)=\text{t-dist}(b_k\,|\,\nu=N-K,\,b_c=\bar{b}_k,\,\Sigma=\frac{\bar{v}}{N-K}\Lambda_{k,k})$$