
MULTI-PARAMETER BAYESIAN MODELS
MULTIVARIATE UPDATING PROCESS
POSTERIOR ANALYSIS ($\theta_{1:5}$)
PREDICTIVE DISTRIBUTION MAP
INFORMED NAVIGATION PATH
MULTIVARIATE UPDATING PROCESS
POSTERIOR ANALYSIS ($\theta_{1:3}$)
PRIOR DISTRIBUTIONS (θ_1, θ_3) (Initial Beliefs)
LIKELIHOOD MODEL 1: SHIELD STATUS (New Data)
LIKELIHOOD MODEL 2: ENGINE OUTPUT (New Data)

Learning Objectives

- Review properties of the multivariate normal distribution
- Review conjugate priors for multivariate normal data
- Explore Wishart and Inverse Wishart distributions
- Work with multinomial and Dirichlet distributions
- Apply Bayesian inference to real data examples

Multivariate normal

Definition

A k-dimensional vector $\boldsymbol{y}$ follows a multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ if:

$$p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)$$

- $\boldsymbol{\mu} \in \mathbb{R}^k$ and $\boldsymbol{\Sigma}$ is a symmetric and positive definite matrix. If $\boldsymbol{\Sigma}$ is not of full rank, the distribution is degenerate and does not have a density.
- A positive definite matrix $\boldsymbol{\Sigma}$ satisfies:
  1. $\boldsymbol{x}^T \boldsymbol{\Sigma} x > 0$ for all nonzero vector $\boldsymbol{x}$. $\boldsymbol{x}^T \boldsymbol{\Sigma} \boldsymbol{x} = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$.
  2. Eigenvalues of $\boldsymbol{\Sigma}$ are all positive.
  3. The determinant of $\boldsymbol{\Sigma}$ is positive (since it is the product of all eigenvalues)

Properties

- Any subset of $\boldsymbol{y}$ has a normal distribution.
- Any linear combination of $\boldsymbol{y}$ is also normal.

$$\boldsymbol{A}\boldsymbol{y} \sim N(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$$

- The conditional distributions are also normal. WLOG, let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$,

$$p(\boldsymbol{y}_1|\boldsymbol{y}_2) \sim N(\boldsymbol{m}, \boldsymbol{W})$$

where

$$\boldsymbol{m} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2)$$

$$\boldsymbol{W} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Likelihood

For n observations $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)^T$, notice we can rewrite the likelihood

$$p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp(-\frac{1}{2} \sum_i (\boldsymbol{y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}))$$

into

$$p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-n/2} \exp(-\frac{1}{2} tr(\boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{S}}))$$

with $\tilde{\boldsymbol{S}} = \sum_i^n (\boldsymbol{y}_i - \boldsymbol{\mu})(\boldsymbol{y}_i - \boldsymbol{\mu})^T$.

Note that $\tilde{\boldsymbol{S}} = \sum_{i=1}^n (\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T + n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})(\bar{\boldsymbol{y}} - \boldsymbol{\mu})^T$. That is, the sample mean and variance are sufficient statistics for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Multivariate normal with known Σ

- Condition: $\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known $\boldsymbol{\Sigma}$
- Conjugate prior: $\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$
- Posterior: $\boldsymbol{\mu}|\boldsymbol{y}, \boldsymbol{\Sigma} \sim N(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$ where:

$$\boldsymbol{\mu}_n = \boldsymbol{\Lambda}_n(\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0 + n\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}}), \quad \boldsymbol{\Lambda}_n^{-1} = \boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Sigma}^{-1}$$

- Notice for the predictive distribution of a new data point $\tilde{\boldsymbol{y}}$,

$$p(\tilde{\boldsymbol{y}}, \boldsymbol{\mu}|\boldsymbol{y}) = N(\tilde{\boldsymbol{y}}|\boldsymbol{\mu}, \boldsymbol{\Sigma})N(\boldsymbol{\mu}|\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n)$$

  So $(\tilde{\boldsymbol{y}}, \boldsymbol{\mu})$ has a joint normal posterior distribution. That is $\tilde{\boldsymbol{y}}|\boldsymbol{y}$ is also normally distributed.

- Posterior predictive for new data point $\tilde{\boldsymbol{y}}$:

$$\mathbb{E}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \boldsymbol{\mu}_n, \quad \text{var}(\tilde{\boldsymbol{y}}|\boldsymbol{y}) = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}_n$$

Inverse Wishart Distribution

Definition

For a $k \times k$ positive definite matrix $\boldsymbol{\Sigma}$, the Inverse Wishart distribution with $\nu > k - 1$ degrees of freedom and scale matrix $\boldsymbol{\Lambda}$ (positive definite) has density:

$$p(\boldsymbol{\Sigma}|\nu, \boldsymbol{\Lambda}) \propto |\boldsymbol{\Sigma}|^{-\frac{\nu+k+1}{2}} \exp\left(-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}^{-1})\right)$$

An important property is that the prior $\boldsymbol{\Sigma} \sim$ Inv-Wishart$(\nu, \boldsymbol{\Lambda})$ is equivalent to a Wishart prior on the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{\Omega} \sim \text{Wishart}(\nu, \boldsymbol{\Lambda}^{-1})$$

Wishart Distribution and Sampling

- The Wishart distribution is defined through the construction:
  If $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \sim N(0, \boldsymbol{\Lambda}^{-1})$ independently, then $\sum_i \boldsymbol{z}_i \boldsymbol{z}_i^T \sim$ Wishart$(n, \boldsymbol{\Lambda}^{-1})$
- An Inverse Wishart prior on $\boldsymbol{\Sigma}$ with known $\boldsymbol{\mu}$ leads to:

$$\boldsymbol{\Sigma}|\boldsymbol{y} \sim \text{Inv-Wishart}(n + \nu, \boldsymbol{\Lambda} + \tilde{\boldsymbol{S}})$$

  where the scale matrix is updated by the sample sum of squares

Wishart and Inverse Wishart: Properties I

- The prior mean of $\boldsymbol{\Sigma}$ under Inv-Wishart$(\nu, \boldsymbol{\Lambda})$ is $\frac{1}{\nu-k-1}\boldsymbol{\Lambda}$
- To understand the partition structure, let $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

- Define the Schur complement: $\boldsymbol{\Sigma}_{22,1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$
- This satisfies: $|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22,1}|$

Wishart and Inverse Wishart: Properties II

- The Inverse Wishart distribution has a special partition property:

$$\begin{aligned}\boldsymbol{\Sigma}_{11} &\sim \text{Inv-Wishart}(\nu - k_2, \Lambda_{11}) \\ \boldsymbol{\Sigma}_{22,1} &\sim \text{Inv-Wishart}(\nu, \Lambda_{22,1})\end{aligned}$$

  and $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22,1}$ are independent

- For $k_1 = 1$, the Inverse Wishart prior implies Inverse Gamma on marginal variances:

$$\sigma_1^2 \sim \text{Inv-Gamma}\left(\frac{\nu - k + 1}{2}, \frac{\Lambda_{11}}{2}\right)$$

- Note that the inverse Wishart has only a single degrees of freedom parameter governing all dimensions.

MVN with Both μ and Σ Unknown

Normal-Inverse-Wishart Prior

Denote $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \text{NIW}(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Lambda}_0)$ if:

$$\boldsymbol{\Sigma} \sim \text{Inv-Wishart}(\nu_0, \boldsymbol{\Lambda}_0), \quad \boldsymbol{\mu}|\boldsymbol{\Sigma} \sim N\left(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0\right)$$

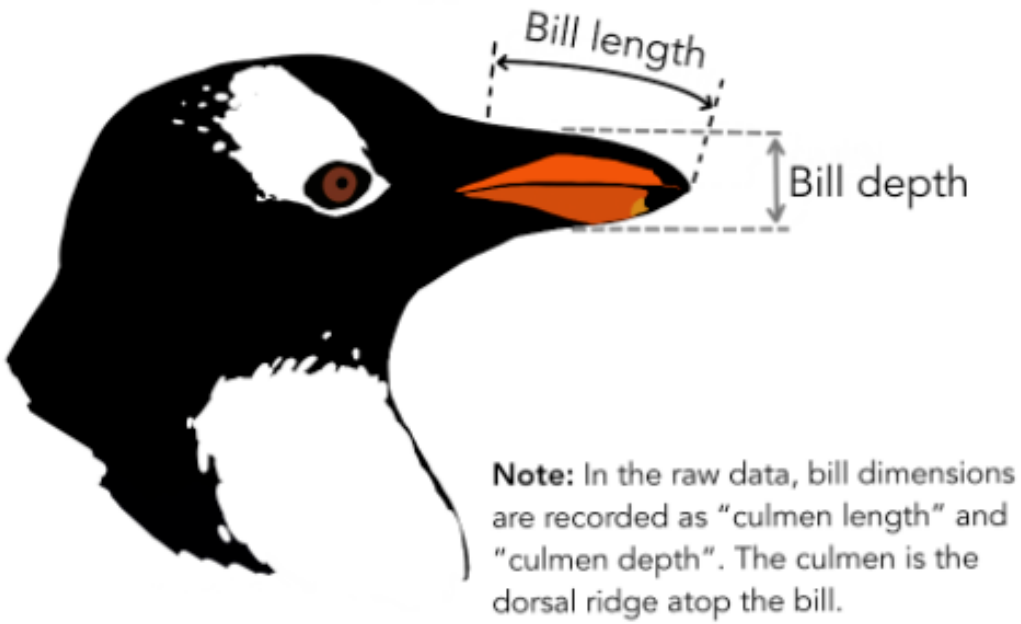
The posterior is also NIW with parameters:

$$\begin{aligned}
\boldsymbol{\mu}_n &= \frac{\kappa_0 \boldsymbol{\mu}_0 + n\bar{\boldsymbol{y}}}{\kappa_0 + n} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\boldsymbol{\Lambda}_n &= \boldsymbol{\Lambda}_0 + \boldsymbol{S} + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_0)(\bar{\boldsymbol{y}} - \boldsymbol{\mu}_0)^T
\end{aligned}$$

The posterior predictive distribution of $\tilde{\boldsymbol{y}}$ is multivariate t.

Example: Palmer penguins

- As an illustration, let us look at the Palmer penguins dataset. It collects various measurements for 344 penguins in Palmer station in Antarctica.
- Let us build a linear model for the bill length.



Example: Palmer penguins

We will use only one of the three penguin species in this example.

```
library(palmerpenguins)
data <- subset(penguins, !is.na(flipper_length_mm) &
                         !is.na(bill_length_mm) &
                         !is.na(species) &
                         !is.na(sex) &
                         species == "Gentoo")
head(penguins)
```

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Gentoo	Biscoe	46.10	13.20	211	4500	female	2007
2	Gentoo	Biscoe	50.00	16.30	230	5700	male	2007
3	Gentoo	Biscoe	48.70	14.10	210	4450	female	2007
4	Gentoo	Biscoe	50.00	15.20	218	5700	male	2007
5	Gentoo	Biscoe	47.60	14.50	215	5400	male	2007
6	Gentoo	Biscoe	46.50	13.50	210	4550	female	2007

Example: Palmer penguins

Consider estimating the joint distribution of flipper length and bill length using a two-dimensional Gaussian model.

```
# 2-d data matrix y
y <- as.matrix(data[, c("flipper_length_mm", "bill_length_mm")])
n <- dim(y)[1]
K <- dim(y)[2]

# Priors
mu0 <- rep(0, K)
nu0 = K + 1
Lambda0 = diag(K)
kappa0 = 1

# Posterior distribution parameters
y_bar <- apply(y, 2, mean)
Lambda_n <- Lambda0 +
            (t(y)-y_bar)%*%t(t(y)-y_bar) +
            kappa0*n/(kappa0+n)*(mu0-y_bar)%*%t(mu0-y_bar)

# to avoid numerical issue with matrix not symmetric
Lambda_n <- (Lambda_n+t(Lambda_n))/2
```

Example: Palmer penguins

Draw samples from the prior predictive distribution

$$\begin{aligned}
\boldsymbol{\Sigma} &\sim \text{Inv-Wishart}(\nu_0, \boldsymbol{\Lambda}_0) \\
\boldsymbol{\mu}|\boldsymbol{\Sigma} &\sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}/\kappa_0) \\
\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} &\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}$$

```
library(LaplacesDemon)

pred_prior <- matrix(0, 1e4, K)

for(i in 1:1e4){
  Sigma_draw <- rinvwishart(nu0, Lambda0)
  mu_draw <- t(rmvn(1, mu0, 1/(kappa0)*Sigma_draw))
  pred_prior[i, ] <- rmvn(1, t(mu_draw), Sigma_draw)
}
```

Example: Palmer penguins

Draw samples from the posterior predictive distribution

$$\begin{aligned}
\boldsymbol{\Sigma}|\boldsymbol{y} &\sim \text{Inv-Wishart}(\nu_n, \boldsymbol{\Lambda}_n) \\
\boldsymbol{\mu}|\boldsymbol{\Sigma}, \boldsymbol{y} &\sim N(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}/\kappa_n) \\
\tilde{\boldsymbol{y}}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{y} &\sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})
\end{aligned}$$

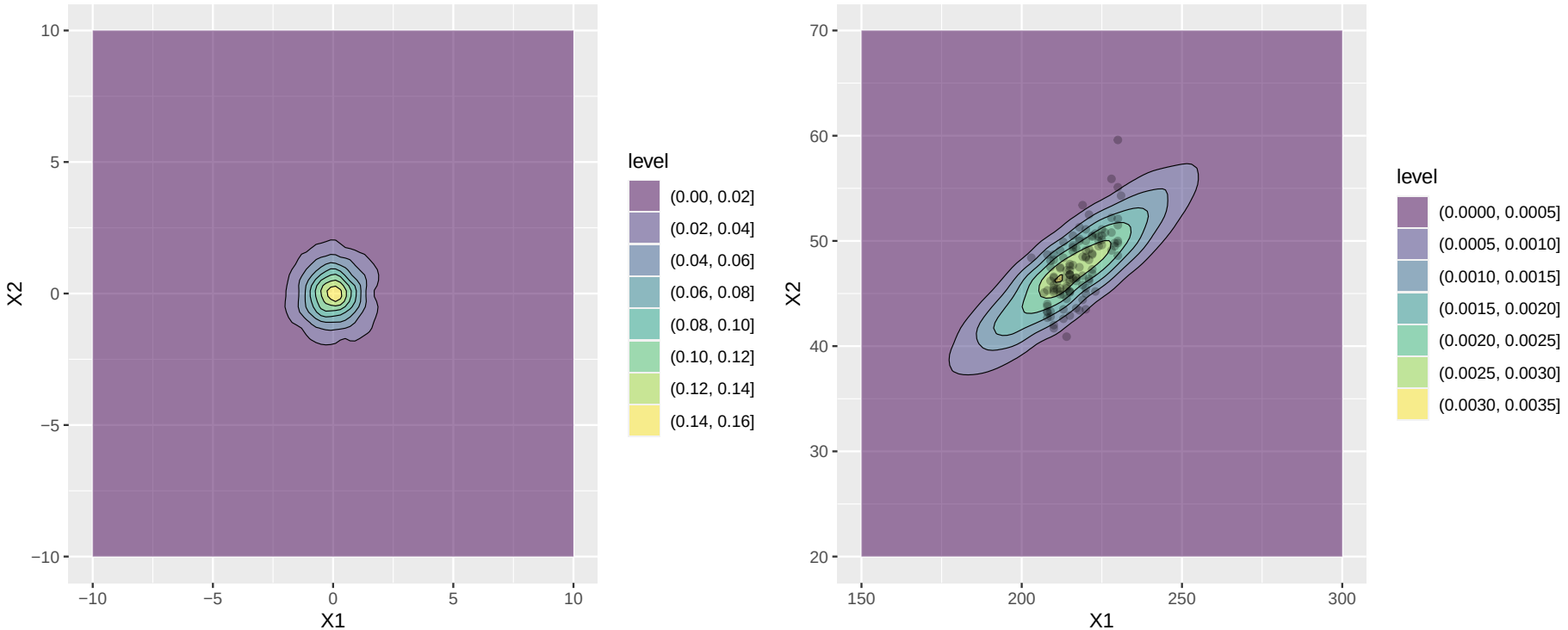
```
pred_post <- matrix(0, 1e4, K)

for(i in 1:1e4){
  Sigma_draw <- rinvwishart(nu0+n, Lambda_n)
  mu_draw <- t(rmvn(1, (kappa0*mu0+n*y_bar)/(kappa0+n),
                       1/(kappa0+n)*Sigma_draw))
  pred_post[i, ] <- rmvn(1, t(mu_draw), Sigma_draw)
}
```

Example: Palmer penguins

Plot the prior and posterior predictive distribution of y.

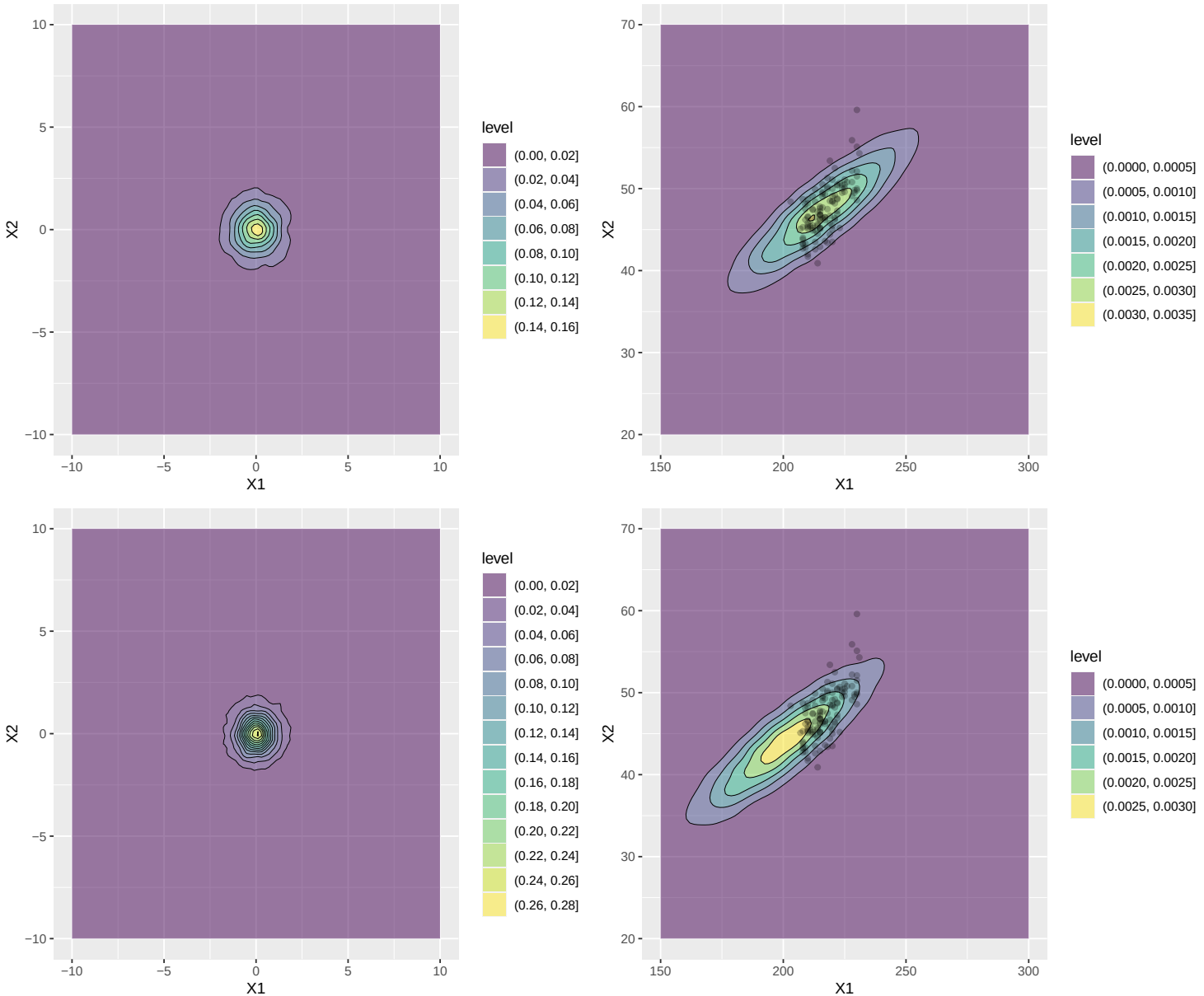
```
# Draw density plot
g1 <- ggplot(data = data.frame(pred_prior), aes(x = X1, y = X2)) +
  geom_density_2d_filled(alpha = 0.5)+
  geom_density_2d(linewidth = 0.25, colour = "black") +
  xlim(c(-10, 10)) + ylim(c(-10, 10))
g2 <- ggplot(data = data.frame(pred_post), aes(x = X1, y = X2)) +
  geom_density_2d_filled(alpha = 0.5)+
  geom_density_2d(linewidth = 0.25, colour = "black") +
  xlim(c(150, 300)) + ylim(c(20, 70)) +
  geom_point(data = data.frame(y),
                    aes(x = flipper_length_mm, y = bill_length_mm),
                    alpha = 0.25)
```

Example: Palmer penguins



Example: Palmer penguins

Comparing with $\kappa_0 = 10$, where the prior for $\boldsymbol{\mu}$ is more concentrated at 0.



Noninformative priors

- Setting $\boldsymbol{\Sigma} \sim$ Inv-Wishart$(k+1, \boldsymbol{I})$ is a commonly used noninformative prior. It has a nice property that the marginal distribution of each off-diagonal elements in the correlation matrix is uniform (see Barnard, McCulloch, Meng, 2000).
- Another noninformative prior is the Jefferys prior,

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(k+1)/2}$$

  which is the limit of NIW as $\kappa_0 \to 0$, $\nu_0 \to -1$ and $|\Lambda_0| \to 0$
- Note again, that in many situations (especially later in hierarchical models), it makes more sense to use a weakly informative prior rather than seeking the noninformativeness.

Alternatives to Inverse Wishart

- A major issue with Inverse Wishart prior for the covariance matrix is that there is only one df parameter for all dimensions. This can be seen from the marginal prior distribution of the variance components.
- Another issue is that the Inverse Wishart imposes a prior dependence between correlation and marginal variances.
- When we start to model covariance matrix in hierarchical models, we could also run into issues where the posterior distribution of $\boldsymbol{\Sigma}$ is heavily concentrated on a degenerate matrix (i.e., the boundary), as the density increases as $|\boldsymbol{\Sigma}| \to 0$.
- Many alternatives have been proposed based on a decomposition

$$\boldsymbol{\Sigma} = diag(\boldsymbol{D})\boldsymbol{R}diag(\boldsymbol{D})$$

  including the scaled inverse-Wishart, LKJ prior for the correlation matrix, etc.

Generating normal and inverse Wishart random variables

- To generate $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we can start with the Cholesky decomposition of $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L}$ is a lower triangular matrix.
- Then we generate $\boldsymbol{z} \sim N(0, \boldsymbol{I})$ and compute $\boldsymbol{x} = \boldsymbol{L}\boldsymbol{z} + \boldsymbol{\mu}$.

- To generate $\boldsymbol{\Sigma} \sim$ Inv-Wishart$(\nu, \boldsymbol{\Lambda})$, first generate $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega} \sim$ Wishart$(\nu, \boldsymbol{\Lambda}^{-1})$ and invert the matrix.
- The Wishart distribution can be sampled by generating ν independent samples $\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_\nu \sim N(0, \boldsymbol{\Lambda}^{-1})$ and let $\boldsymbol{\Omega} = \sum_{i=1}^{\nu} \boldsymbol{\alpha}_i \boldsymbol{\alpha}_i^T$.

- A faster alternative is to use Bartlett's decomposition:
  1. Generate a lower triangular matrix A with $a_{ii} \sim \sqrt{\chi^2_{\nu-i+1}}$ and $a_{ij} \sim N(0,1)$ for $j < i$.
  2. Compute the Cholesky decomposition $\boldsymbol{\Lambda} = \boldsymbol{L}\boldsymbol{L}^T$.
  3. Compute $\boldsymbol{\Omega} = \boldsymbol{L}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{L}^T$.

  This method requires only $k(k+1)/2$ random variable generations and automatically produces the Cholesky decomposition for $\boldsymbol{\Omega}$.

- There are R packages with highly optimized implementations, e.g., mvtnorm, MCMCpack, LaplacesDemon, ...

The Multinomial Distribution

Definition

Consider a categorical random variable with K possible outcomes. Count occurrences of each type for n trials. The count vector $\boldsymbol{y}$ has density:

$$p(\boldsymbol{y}|\boldsymbol{\theta}) \propto \prod_{j=1}^{K} \theta_j^{y_j}$$

with $\sum_{j=1}^{K} \theta_j = 1$.

- Generalization of the binomial distribution to K categories
- Conjugate prior: the Dirichlet distribution

The Dirichlet Distribution

Definition

A distribution on the K-dimensional simplex:

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \propto \prod_{j=1}^{K} \theta_j^{\alpha_j - 1}$$

with $\sum_{j=1}^{K} \theta_j = 1$ and $\alpha_j > 0$ for all j.

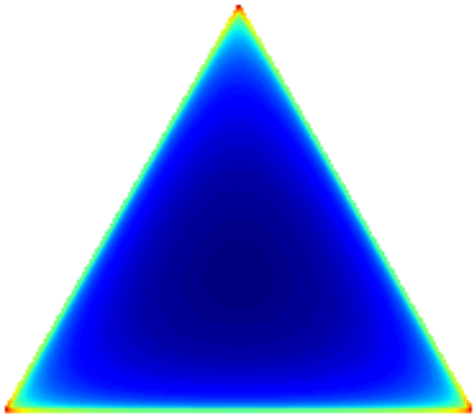
The Dirichlet is the conjugate prior for the multinomial. Under a multinomial likelihood, the posterior is:

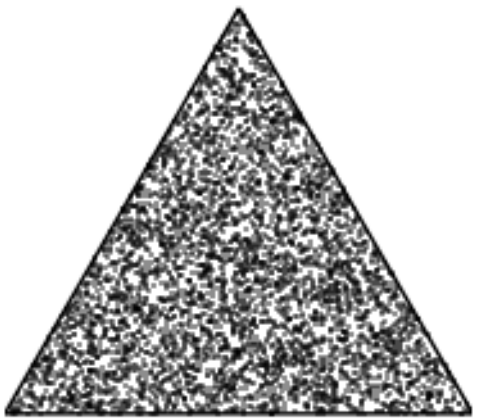
$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \prod_{j=1}^{K} \theta_j^{y_j + \alpha_j - 1}$$

The posterior is also Dirichlet with updated parameters.

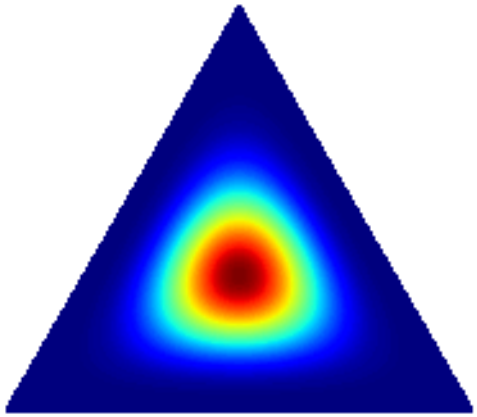
The Dirichlet distribution

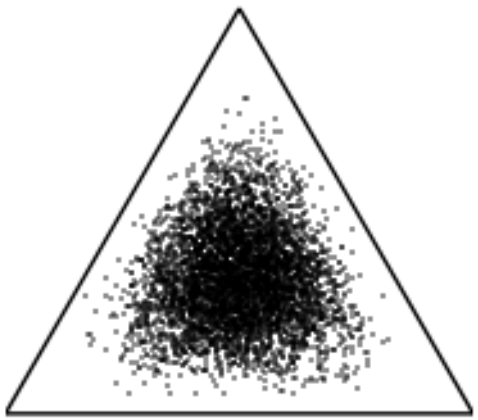
α = (0.999, 0.999, 0.999)



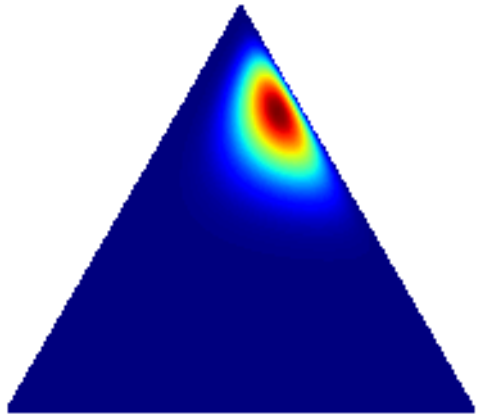


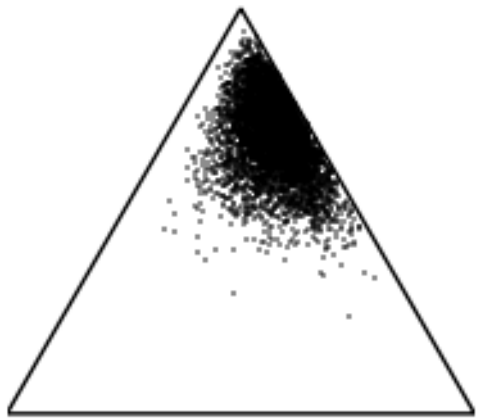
α = (5.000, 5.000, 5.000)





α = (2.000, 5.000, 15.000)





Sampling from the Dirichlet distribution

- There is an important relationship between Dirichlet and gamma distributions. Let

$$Z_j \sim \text{Gamma}(\alpha_j, \beta)$$

for $j = 1, ..., K$ independently, then the vector θ with

$$\theta_j = \frac{Z_j}{\sum_j Z_j}, j = 1, ..., K$$

follow a Dirichlet distribution with parameter α.

- The conditional distribution of a sub-vector of θ given the remaining elements of θ is also a Dirichlet (after rescaling to sum to 1).

Slovenia opinion poll example

- In 1990, a plebiscite was held in Slovenia at which the adult citizens voted on the question of independence.
- A Slovenian public opinion survey had been conducted that included several questions concerning likely plebiscite attendance and voting. In that survey, 2074 Slovenians were asked:
  1. Are you in favor of independence?
  2. Will you attend the plebiscite?
- The rules of the plebiscite were such that only those attending and voting 'yes' would be counted as being in favor of independence.

Table of counts

	Independence		
Attendance	*Yes*	*No*	*DK*
Yes	1,439	78	159
No	16	16	32
DK	144	54	136

Dirichlet model for contingency tables

$n_{11} = 1439$	$n_{10} = 78$	$n_{1\star} = 159$
$n_{01} = 16$	$n_{00} = 16$	$n_{0\star} = 32$
$n_{\star 1} = 144$	$n_{\star 0} = 54$	$n_{\star\star} = 136$

- If we ignore any counts of DK's, a naive estimate of the fraction of people attending voting yes is $n_{11}/(n_{11} + n_{10} + n_{01} + n_{00}) = 0.93$.
- We can try to understand how the DK counts should be distributed into one of the four cells. Let the unobserved true counts $(x_{11}, x_{10}, x_{01}, x_{00}) \sim \text{Mult}(n, \theta)$.
- Given $n_{\star\star}$ people who we know nothing about, it may be reasonable to assume that

$$(n_{\star\star}^{(11)}, n_{\star\star}^{(10)}, n_{\star\star}^{(01)}, n_{\star\star}^{(00)}) \sim \text{Mult}(n_{\star\star}, \theta)$$

  where $n_{\star\star}^{(11)}$ is the number of these people who will attend and vote yes, etc.
- Similarly for the other cells, we can use similar models given the partial information, e.g.,

$$(n_{1\star}^{(11)}, n_{1\star}^{(10)}) \sim \text{Mult}(n_{1\star}, (\frac{\theta_{11}}{\theta_{11} + \theta_{10}}, \frac{\theta_{10}}{\theta_{11} + \theta_{10}}))$$

Dirichlet model for contingency tables

- Thus we can complete a Bayesian model with prior on θ

$$\boldsymbol{\theta} \sim Dir(\alpha)$$
$$(x_{11}, x_{10}, x_{01}, x_{00}) \sim \text{Mult}(n, \theta)$$

where $x_{ij} = n_{ij} + n_{i\star}^{(ij)} + n_{\star j}^{(ij)} + n_{\star\star}^{(ij)}$ and the corresponding latent variables follow

$$(n_{\star\star}^{(11)}, n_{\star\star}^{(10)}, n_{\star\star}^{(01)}, n_{\star\star}^{(00)}) \sim \text{Mult}(n_{\star\star}, \theta)$$
$$(n_{1\star}^{(11)}, n_{1\star}^{(10)}) \sim \text{Mult}(n_{1\star}, (\frac{\theta_{11}}{\theta_{11}+\theta_{10}}, \frac{\theta_{10}}{\theta_{11}+\theta_{10}}))$$
$$(n_{0\star}^{(01)}, n_{0\star}^{(00)}) \sim \text{Mult}(n_{0\star}, (\frac{\theta_{01}}{\theta_{01}+\theta_{00}}, \frac{\theta_{00}}{\theta_{01}+\theta_{00}}))$$
$$(n_{\star 1}^{(11)}, n_{\star 1}^{(01)}) \sim \text{Mult}(n_{\star 1}, (\frac{\theta_{11}}{\theta_{11}+\theta_{01}}, \frac{\theta_{01}}{\theta_{11}+\theta_{01}}))$$
$$(n_{\star 0}^{(10)}, n_{\star 0}^{(00)}) \sim \text{Mult}(n_{\star 0}, (\frac{\theta_{10}}{\theta_{10}+\theta_{00}}, \frac{\theta_{00}}{\theta_{10}+\theta_{00}}))$$

- The model for latent counts here is implicitly assuming data are missing at random. We will define this assumption more explicitly later in the course.

Dirichlet model for contingency tables

- Red line shows the posterior mean. Blue line shows the naive estimates ignoring the DK's.