Multiparameter models

Teacher: Matilde Trevisani

DEAMS

A.A. 2024/2025
(aggiornato: 2025-04-13)

Agenda

Multiparameter models

- marginalization
- Normal model with both mean and varianza unknowm
  - non-informative prior
  - conjugate prior (and semi-conjugate)
- Multinomial model with Dirichlet conjugate prior
- Normal multivariate model with unknown mean and covariance known/unknown

Two-parameters models

A model is specified with two real parameters θ_1, θ_2

$$p(y|\theta_1, \theta_2)$$

the prior is then a bivariate distribution

$$p(\theta_1, \theta_2)$$

and the posterior is then a bivariate distribution as well

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)$$

Suppose that, say, θ_1 is the parameter of interest whereas θ_2 is a *nuisance* parameter.

Thus, we may be particularly interested in the marginal posterior for θ_1.

Marginalization

We find the marginal posterior for θ_1 by marginalizing (*averaging over* θ_2) the joint posterior

$$\pi(\theta_1|y) = \int \pi(\theta_1, \theta_2|y) d\theta_2$$

The joint posterior is factorized

- either as

$$\pi(\theta_1|y) = \int p(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2) d\theta_2$$

- or as a mixture of conditional posteriors

$$\pi(\theta_1|y) = \int \pi(\theta_1|\theta_2, y)\pi(\theta_2|y) d\theta_2$$

Multiparameter models and marginalization

θ_1 and θ_2 can be vectors, and in general we deal with multiparameter models.

The goal is to find the marginal posterior of a quantity of interest

- a model parameter
- a future event, e.g., the predictive posterior distribution which is obtained by marginalizing the posterior distribution

$$
\begin{aligned}
p(\tilde{y} \mid y) &= \int p(\tilde{y}, \theta \mid y) d\theta \\
&= \int p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta
\end{aligned}
$$

(Univariate) normal model with μ and σ^2 unknown

Likelihood

Let

$$y_1, \ldots, y_n | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$$

The likelihood is

$$\begin{aligned}
p(y|\mu, \sigma^2) &\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_j (y_j - \mu)^2 \right\} \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_j (y_j - \bar{y} + \bar{y} - \mu)^2 \right\} \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_j (y_j - \bar{y})^2 + n(\bar{y} - \mu)^2 + 2(\bar{y} - \mu) \sum_j (y_j - \bar{y}) \right) \right\} \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{n}{2\sigma^2} (\hat{\sigma}^2 + (\bar{y} - \mu)^2) \right\} \\
&\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} ((n-1)s^2 + n(\bar{y} - \mu)^2) \right\}
\end{aligned}$$

Likelihood

Let

$$y_1, \ldots, y_n | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$$

The likelihood is

$$\begin{aligned} p(y|\mu, \sigma^2) &\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{n}{2\sigma^2} (\hat{\sigma}^2 + (\bar{y} - \mu)^2) \right\} \\ &\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} ((n-1)s^2 + n(\bar{y} - \mu)^2) \right\} \end{aligned}$$

a function of the sufficient statistics

$$\bar{y} = \frac{1}{n} \sum_j y_j; \quad s^2 = \frac{1}{n-1} \sum_j (y_j - \bar{y})^2 = \frac{n}{n-1} \hat{\sigma}^2.$$

Normal model with noninformative prior

Noninformative prior specification

Consider the improper prior

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

that is, μ and σ^2 are independent and

- $p(\mu) \propto k$
- $p(\sigma^2) \propto (\sigma^2)^{-1}$
  - Equivalently, we could say that (see Gelman p.21, change of variable)
    - $p(\sigma) \propto \sigma^{-1}$
    - $p(\log \sigma^2) \propto k$
  - In the model with known μ such a prior (corresponding to an $\text{Inv-}\chi^2$ with $\nu_0 = 0$) leads to

$$p(\sigma^2|y) \sim \text{Inv-}\chi^2(n, \hat{\sigma}^2)$$

that is, conditional on y, $\sigma^2 =_d \frac{n\hat{\sigma}^2}{\chi^2_n}$

Posterior with noninformative prior

The posterior is

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto (\sigma^2)^{-1} p(y|\mu, \sigma^2) \\ &\propto (\sigma^2)^{-1} (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2) \right\} \\ &\propto \underbrace{(\sigma^2)^{-1/2} \exp\left\{ -\frac{n}{2\sigma^2}(\bar{y}-\mu)^2 \right\}}_{\mu|\sigma^2, y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)} \underbrace{(\sigma^2)^{-(n+1)/2} \exp\left\{ -\frac{1}{2\sigma^2}(n-1)s^2 \right\}}_{\sigma^2|y \sim \text{Inv-}\chi^2(n-1, s^2)} \end{aligned}$$

That is,

$$\begin{aligned} p(\mu, \sigma^2|y) &= p(\mu|\sigma^2, y)\, p(\sigma^2|y) \\ &= N\left(\bar{y}, \sigma^2/n\right) \text{Inv-}\chi^2(n-1, s^2) \end{aligned}$$

The factorization

$$p(\mu, \sigma^2|y) = p(\mu|\sigma^2, y)\, p(\sigma^2|y)$$

derives from

- the posterior for μ conditional on σ^2 is

  $$p(\mu|\sigma^2, y) = N\left(\bar{y}, \sigma^2/n\right)$$

  which follows what was obtained (in the single parameter model) for the a posteriori of μ when σ^2 is known and the a priori for μ is uniform.

- the marginal posterior for σ^2

  $$p(\sigma^2|y) = \text{Inv-}\chi^2(n-1, s^2)$$

  which is like what is obtained in the single parameter model with μ known and with improper prior $p(\sigma^2) \propto 1/\sigma^2$, but here taking into account one less degree of freedom.

Marginal posterior for σ^2

The marginal posterior for σ^2 is

$$\begin{aligned}
p(\sigma^2|y) &= \int p(\mu, \sigma^2|y) d\mu \\
&\propto \int \sigma^{-n-2} \exp\left\{-\frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2)\right\} d\mu \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \int \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right\} d\mu \\
&\qquad \text{Notice that } \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right) d\theta = 1 \\
&\propto \sigma^{-n-2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \sqrt{\frac{2\pi\sigma^2}{n}} \\
&\propto (\sigma^2)^{-(n+1)/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}
\end{aligned}$$

that is

$$\sigma^2|y \sim \text{inv-}\chi^2(n-1, s^2)$$

Marginal posterior for σ^2: notes

Recall that by

$$\sigma^2|y \sim \text{inv-}\chi^2(n-1, s^2)$$

we mean that, conditional on y,

$$\sigma^2 =_d \frac{(n-1)s^2}{X}, \quad X \sim \chi^2_{n-1}$$

and compare this with the usual result on the sampling distribution of s^2,

$$\frac{(n-1)s^2}{\sigma^2}|\sigma^2 \sim \chi^2_{n-1}$$

Note also that it is equivalent to write

$$\sigma^2|y \sim \text{Inv-gamma}\left(n-1/2, (n-1)s^2/2\right)$$

$$(\sigma^2)^{-1}|y \sim \text{Gamma}\left(n-1/2, (n-1)s^2/2\right)$$

Marginal posterior for μ

$$
\begin{aligned}
p(\mu|y) &= \int_0^\infty p(\mu, \sigma^2|y) d\sigma^2 \\
&\propto \int_0^\infty \sigma^{-n-2} \exp\left\{ \underbrace{-\frac{1}{2\sigma^2}((n-1)s^2 + n(\bar{y}-\mu)^2)}_{z=\frac{A}{2\sigma^2}=\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{2\sigma^2}} \right\} d\sigma^2 \\
&\propto \int_0^\infty \left(\frac{A}{2z}\right)^{-(n+2)/2} \exp\{-z\} \frac{A}{2z} dz \\
&\propto A^{-n/2} \int_0^\infty z^{(n-2)/2} \exp\{-z\} dz
\end{aligned}
$$

And recognizing a non-normalized gamma integral $\Gamma(u) = \int_0^\infty x^{u-1} \exp(-x) dx$

$$
\begin{aligned}
&\propto ((n-1)s^2 + n(\bar{y}-\mu)^2)^{-n/2} \\
&\propto \left(1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right)^{-n/2} \Rightarrow p(\mu|y) = t_{n-1}(\bar{y}, s^2/n)
\end{aligned}
$$

Marginal posterior for μ: notes

Hence

$$\mu|y \sim t_{n-1}(\bar{y}, s^2/n)$$

which is equivalent to

$$\left.\frac{\mu - \bar{y}}{s/\sqrt{n}}\right| y \sim t_{n-1}$$

analogous to the usual result for the pivotal quantity

$$\left.\frac{\bar{y} - \mu}{s/\sqrt{n}}\right| \mu, \sigma^2 \sim t_{n-1}$$

1. the posterior distribution of the pivotal quantity does not depend on the data y,
2. the sampling distribution of the pivotal quantity does not depend on the parameters μ and σ^2

In general, a pivotal quantity for the parameter to be estimated is defined as a non-trivial function of the data and the estimand parameter whose sampling distribution is independent of all parameters and data.

Keep in mind that marginal posterior $p(\mu \mid y)$

$$p(\mu \mid y) = \int_0^\infty p(\mu \mid \sigma^2, y) p(\sigma^2 \mid y) d\sigma^2$$

is a mixture of normal distributions where the mixing density is the marginal posterior of σ^2

in order to derive the posterior predictive analytically.

Predictive distribution for $\tilde{y}$

In general

$$p(\tilde{y}|y) = \int\int p(\tilde{y}|\mu, \sigma^2, y)p(\mu, \sigma^2|y)d\mu d\sigma^2$$

and it can be shown that

$$\tilde{y}|y \sim t_{n-1}\left(\bar{y}, \left(1 + \frac{1}{n}\right)s^2\right)$$

Predictive distribution for $\tilde{y}$: proof

First note that we have proven that

$$p(\mu|y) = \int \underbrace{p(\mu|\sigma^2, y)}_{\mathcal{N}(\bar{y}, \sigma^2/n)} p(\sigma^2|y) d\sigma^2 = t_{n-1}(\bar{y}, s^2/n)$$

Then note that

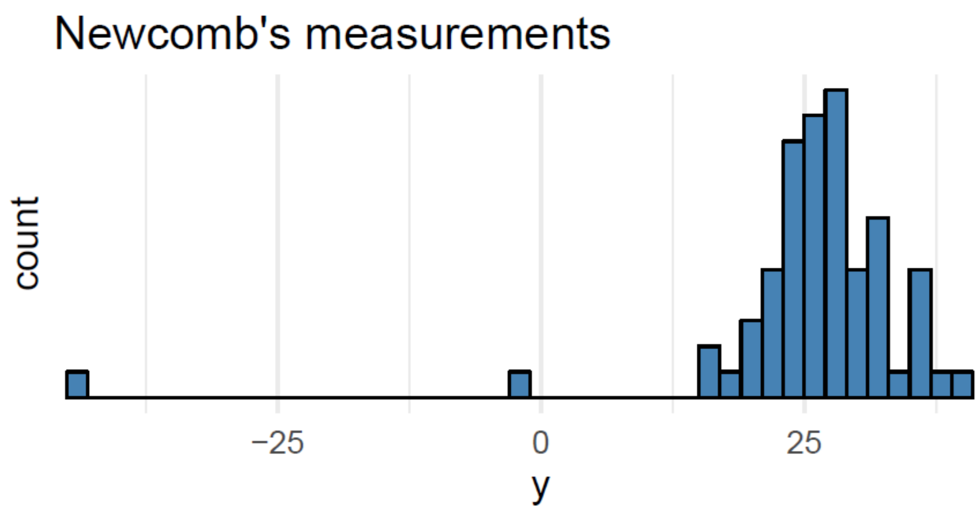
$$\begin{aligned}
p(\tilde{y}|y) &= \int\int p(\tilde{y}|\mu, \sigma^2, y) p(\mu, \sigma^2|y) d\mu d\sigma^2 \\
&= \int\int p(\tilde{y}|\mu, \sigma^2) p(\mu, \sigma^2|y) d\mu d\sigma^2 \\
&= \int \underbrace{\left(\int \underbrace{p(\tilde{y}|\mu, \sigma^2)}_{\mathcal{N}(\mu, \sigma^2)} \underbrace{p(\mu|\sigma^2, y)}_{\mathcal{N}(\bar{y}, \sigma^2/n)} d\mu \right)}_{\mathcal{N}(\bar{y}, \sigma^2(1+1/n))} p(\sigma^2|y) d\sigma^2
\end{aligned}$$

This is the posterior predictive distribution when variance is known, $p(\tilde{y}|\sigma^2, y)$

$$\sim t_{n-1}\left(\bar{y}, s^2\left(1 + \frac{1}{n}\right)\right)$$

Simon Newcomb Experiment (1882): Speed of Light

Newcomb measured ($n = 66$) the time required for light to travel a total distance of 7,422 meters: the distance from his laboratory on the Potomac River to a mirror at the base of the Washington Monument and back.



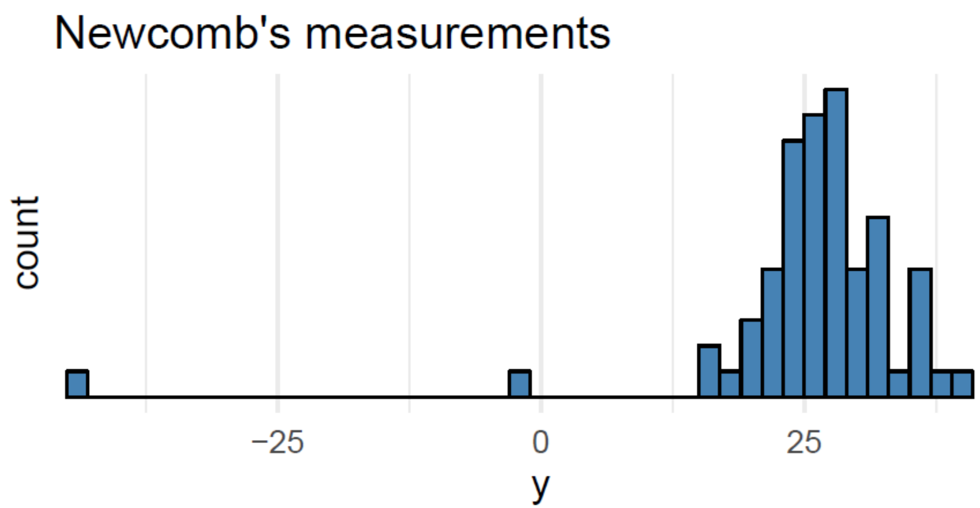
$n = 66$
$\bar{z} = 24826.2$
$s = 10.8$

True value is 24833.02 (in nanoseconds)

Trasformation:
$y = z - 24800$
$\bar{y} = 26.2$

True value: 33.02

Newcomb data: Normal model with noninformative prior

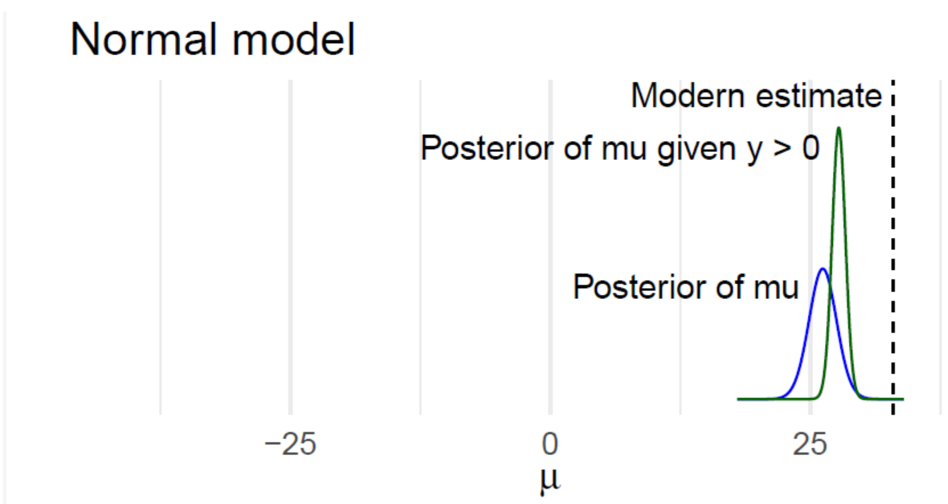


$n = 66$
$\bar{y} = 26.2$
$s = 10.8$

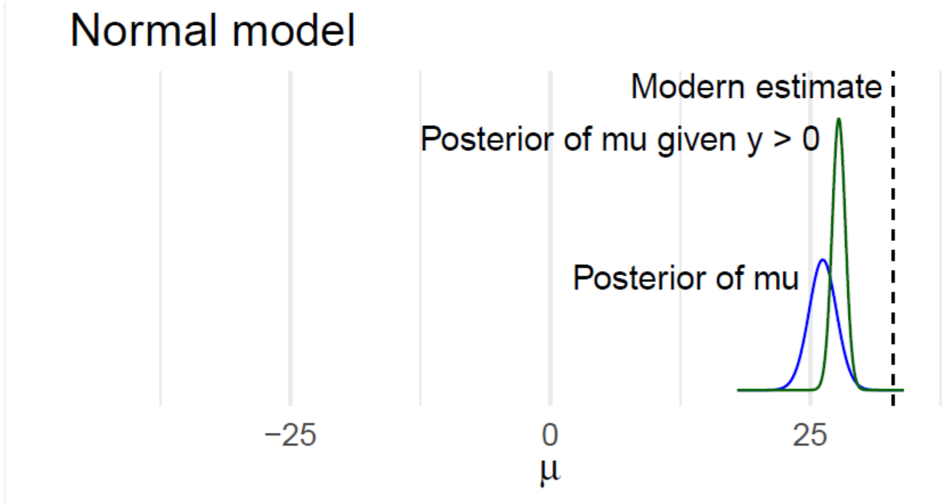
Only positive values:

$\bar{y} = 27.8$
$s = 5.1$

True value: 33.02



Newcomb data: posterior for μ



$$\mu|y \sim t_{65}\left(26.2, \frac{10.8^2}{66}{=}1.329^2\right)$$

posterior interval:
$\bar{y} \pm t_{65,0.975}\frac{s}{\sqrt{66}} =$
$26.2 \pm 1.997 \times 1.329 =$
$[23.6, 28.8]$

only positive data:
$\bar{y} \pm t_{63,0.975}\frac{s}{\sqrt{64}} =$
$27.8 \pm 1.998 \times 0.635 =$
$[26.5, 29.0]$

Normal model with conjugate prior

Conjugate prior specification

The conjugate prior must have the form

$$p(\sigma^2)p(\mu \mid \sigma^2)$$

(see the form of the likelihood in the previous section)

A convenient parametrization is

$$\begin{aligned} \mu \mid \sigma^2 &\sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

which can be written as

$$p(\mu, \sigma^2) = \text{N-Inv-}\chi^2(\mu_0, \sigma_0^2/\kappa_0; \nu_0, \sigma_0^2)$$

- μ and σ^2 are dependent a priori
  $\rightarrow$ if σ^2 is large, then μ has a less precise prior
- Marginally, $p(\mu) = t_{\nu_0}(\mu_0, \sigma_0^2/\kappa_0)$

Joint posterior

Joint posterior is

$$p(\mu, \sigma^2 \mid y) = \text{N-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2)$$

where

$$\begin{aligned}
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\nu_n \sigma_n^2 &= \nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2
\end{aligned}$$

Note that

$$E(\sigma^2 | y) = \frac{\nu_n \sigma_n^2}{\nu_n - 2} = \frac{\nu_0 \sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2}{\nu_0 + n - 2}$$

Marginal for σ^2 and μ

- Conditional $p(\mu \mid \sigma^2, y)$

$$\begin{aligned}\mu \mid \sigma^2, y \sim \mathrm{N}(\mu_n, \sigma^2/\kappa_n) \\ = \mathrm{N}\left(\frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}, \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}\right) \\ = \mathrm{N}\left(\frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n}, \frac{\sigma^2}{\kappa_0 + n}\right)\end{aligned}$$

- Marginal $p(\sigma^2 \mid y)$

$$\sigma^2 \mid y \sim \text{Inv-}\chi^2(\nu_n, \sigma_n^2)$$

- Marginal $p(\mu \mid y)$

$$\mu \mid y \sim t_{\nu_n}(\mu \mid \mu_n, \sigma_n^2/\kappa_n)$$

Multinomial model

Multinomial model for categorical data

See BDA3 p. 69

- Extension of binomial
- y_j number of observations of category j
- $y = (y_1, \ldots, y_k)$ vector of counts of number of observations for each category, with $\sum y_j = n$

- Observational model

$$p(y \mid \theta) \propto \prod_{j=1}^{k} \theta_j^{y_j},$$

  with $\sum \theta_j = 1$
  (Analogously to the binomial, we assume that the distribution is conditional to the number n of observations.)

- Conjugate prior is the Dirichlet

$$p(\theta \mid \alpha) \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \propto \mathcal{D}(\alpha)$$

Conjugate prior

- The Dirichlet, $\mathcal{D}(\alpha)$ - is in the exponential family and - is the multivariate generalization of the $\mathcal{B}eta$ distribution
- The θ_j are nonnegative and sum to 1. More formally, the support of a k-dimensional Dirichlet is the ($k-1$-)simplex of R^k

$$\left\{\theta \in \mathbb{R}^k : \theta_j > 0,\ \sum_{j=1}^{k} \theta_j = 1\right\}$$

- $\mathcal{D}(\alpha)$ is defined for $\alpha_j > 0$ and if $\alpha_0 = \sum_j \alpha_j$, $E(\theta_j) = \alpha_j/\alpha_0$
- The Dirichlet prior contains an equivalent information to $\sum_j(\alpha_j - 1)$ observations with $\alpha_j - 1$ observations of category j
- $\alpha_0 = \sum_j \alpha_j$ determines the *concentration* of a Dirichlet, that is, how much the distribution is dense (high α_0) or sparse (low α_0)

Posterior

- Posterior

$$p(\theta \mid \alpha,\, y) \propto \prod_{j=1}^{k} \theta_j^{y_j+\alpha_j-1} \propto \mathcal{D}(y+\alpha)$$

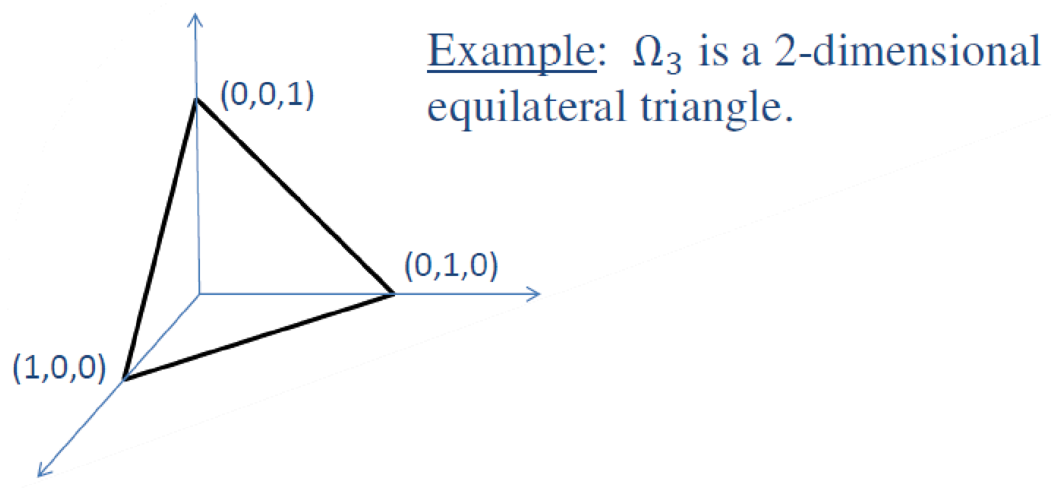
- Notice that $E(\theta_j|y) = \frac{y_j+\alpha_j}{n+\alpha_0} = \frac{y_j}{n} \cdot \frac{n}{n+\alpha_0} + \frac{\alpha_j}{\alpha_0} \cdot \frac{\alpha_j}{n+\alpha_0}$
- There are several plausible noninformative priors:
    - uniform prior: $\alpha_j = 1\,\forall j$
    - improper prior: $\alpha_j = 0\,\forall j$, i.e., uniform on $\log(\theta_j)$
        - posterior is proper if there is at least one observation in each category, so that each component of y is positive.

Example of simplex

The support Ω_k is the space of all the probability vectors of k probabilities that sum to 1.

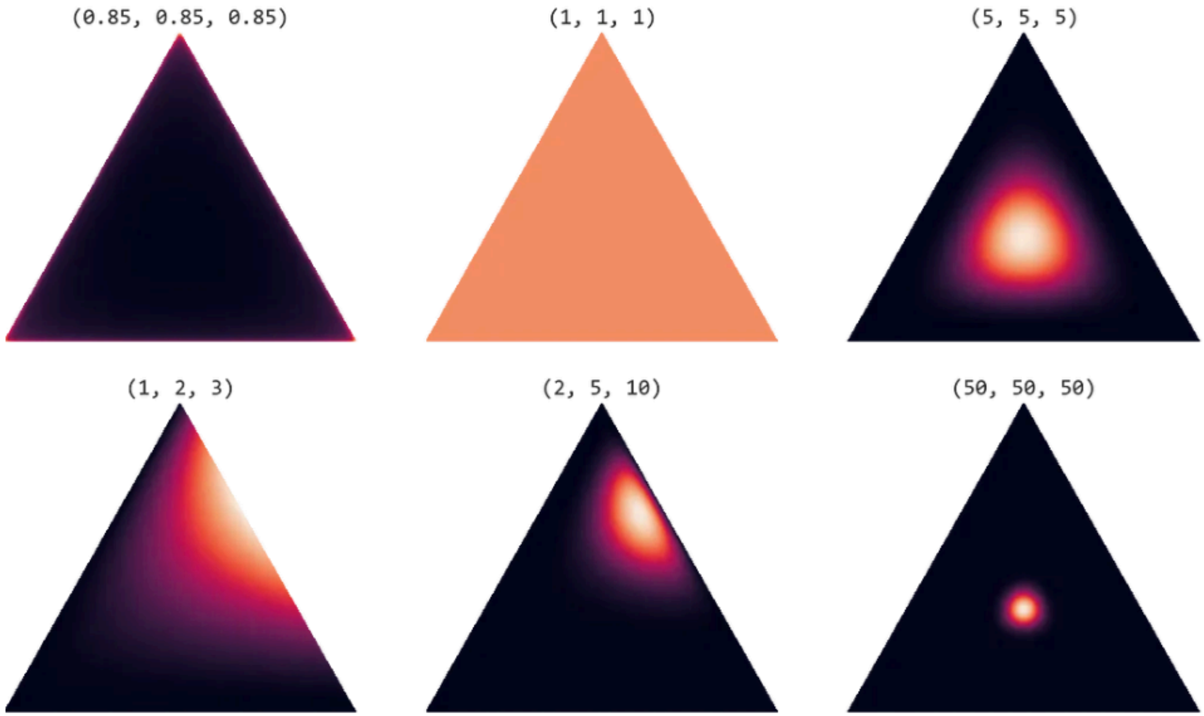
Because of the constraint on the components of a probability vector, Ω_k is $k-1$ dimensional and is the **standard or probability** $K-1$**-simplex**.

For $k=3$ the support is the triangle with vertices $(1,0,0)$, $(0,1,0)$, $(0,0,1)$



$\mathcal{D}(\alpha)$ with different α

3-dimensional Dirichlet distribution on a 2-simplex for different values of α.



Properties:

Symmetric, flat Dirichlet; concentration parameter $\sum_k \alpha_k$.

Multivariate Normal model

Multivariate Normal

BDA3 p. 70-73

- y is a vector of d components
- $y \mid \mu,\, \Sigma \sim \mathrm{N}_d(y \mid \mu,\, \Sigma)$

- Observational model

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

- Multivariate normal with known variance p. 70-71

- Multivariate normal with unknown mean and variance p. 72-73

The two cases follow the development seen in the univariate context but with matrix expressions since the distributions are here multivariate