Multiparameter Model

Anita Wang

2017 - 2018

Introduction

- Virtually every practical problem in statistics involves more than one unknown or unobservable quantity.
- The ultimate aim of a Bayesian analysis is to obtain the **marginal posterior distribution** of the particular parameters of interest.
- We first require the joint posterior distribution of all unknowns, and then we integrate this distribution over the unknowns that to obtain the desired marginal distribution.
- In many problems there is no interest in making inferences about many of the unknown parameters. Parameters of this kind are often called **nuisance parameters**.

Averaging over 'nuisance parameters'

- Suppose θ has two parts, each of which can be a vector, $\theta = (\theta 1, \theta 2)$
- suppose that we are only interested (at least for the moment) in inference for θ1, so θ2 may be considered a **'nuisance' parameter**.
- For instance, in the simple example,

$$y|\mu, \sigma^2 \sim N(\mu, \sigma^2),$$

in which both μ (='θ1') and σ2 (='θ2') are unknown, interest commonly centers on μ.

Joint Posterior Distribution

- We seek the conditional distribution of the parameter of interest given the observed data; in this case, $p(\theta_1|y)$.
- Joint posterior density:

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

- Marginal posterior density:

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

Alternatively,

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

Marginal Posterior Density

- $p(\theta_1|y)$ can be regarded as a mixture of the conditional posterior distributions given the nuisance parameter,θ_2, where $p(\theta_2|y)$.is a weighting function for the different possible values of θ_2.
- The weights depend on the posterior density of θ_2 and thus on a combination of evidence from data and prior model
- $p(\theta_1|y)$ can be computed by marginal and conditional simulation, first drawing θ_2 from its marginal posterior distribution and then θ_1 from its conditional posterior distribution, given the drawn value of θ_2.

Univariate Normal with a Noninformative Prior

- Consider a vector y of n independent observations from a univariate normal distribution, $N(\mu, \sigma^2)$
- Assuming prior independence of location and scale parameters, is uniform on (μ, log σ) or,

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

- The joint posterior distribution

$$p(\mu, \sigma^2 | y) \propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[\sum_{i=1}^{n}(y_i - \bar{y}) + n(\bar{y} - \mu)^2])$$

$$= \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

The conditional posterior distribution

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sample variance. The sufficient statistics are $\bar{y}$ and s^2

- The conditional posterior distribution, $p(\mu|\sigma^2, y)$

we simply use the result derived in lecture 2 for the mean of a normal distribution with known variance and a uniform prior distribution:

$$\mu|\sigma^2, y \sim N(y, \sigma^2 /n).$$

The marginal posterior distribution

- The marginal posterior distribution, $p(\sigma^2|y)$

$$p(\sigma^2|y) \propto \int \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y}-\mu)^2]\right) d\mu$$

$$\propto \sigma^{-n-2} \exp\left(-\frac{1}{2\sigma^2}(n-1)s^2\right)\sqrt{\frac{2\pi\sigma^2}{n}}$$

$$\propto (\sigma^2)^{-\frac{n+1}{2}} \exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)$$

which is a scaled inverse-χ^2 density:

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

The marginal posterior distribution

- Another way to obtain $p(\sigma^2|y)$,

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

$$= \frac{\sigma^{-n-2}\exp(-\frac{1}{2\sigma^2}[(n-1)S^2 + n(\bar{y}-\mu)^2])}{\frac{n}{\sqrt{2\pi}\sigma}\exp\{-\frac{n}{2\sigma^2}(\mu-\bar{y})^2\}}$$

$$\propto \sigma^{-n-1}\exp\{-\frac{1}{2\sigma^2}[(n-1)S^2 + n(\bar{y}-\mu)^2 - n(\bar{y}-\mu)^2]\}$$

$$\propto (\sigma^2)^{-(n+1)/2}\exp(-\frac{(n-1)s^2}{2\sigma^2})$$

which is a scaled inverse-χ^2 density:

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

Joint Posterior Density

- $p(\sigma^2|y)$ is a scaled inverse-χ^2 density:

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

Therefore,

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

Note that this result agrees with the standard frequentist result on the sample variance.

- As we know before,

$$\mu|\sigma^2, y \sim N(\bar{y}, \sigma^2/n)$$

The joint posterior density,

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

Sampling from the joint posterior distribution

- Now that we have $p(\mu|\sigma^2, y)$ and $p(\sigma^2|y)$, inference on μ isn't difficult.
- One method is to use the Monte Carlo approach discussed earlier
  1. Sample σ_i^2 from $p(\sigma^2|y)$
  2. Sample μ_i from $p\left(\mu|\sigma_i^2, y\right)$

Then $\mu_1, \dots, \mu_m$ is a sample from $p(\mu|y)$.

- Note that in this case, it is actually possible to derive the exact density of $p(\mu|y)$.

Marginal Posterior Distribution for μ

- The marginal posterior distribution

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

- Let $z = \frac{A}{2\sigma^2}$, where $A = (n-1)s^2 + n(\bar{y}-\mu)^2$

$$p(\mu|y) \propto A^{-\frac{n}{2}} \int_0^\infty z^{\frac{n-2}{2}} \exp(-z) dz$$

$$\propto [(n-1)s^2 + n(\bar{y}-\mu)^2]^{-\frac{n}{2}} \longleftarrow \text{Gamma integral}$$

$$\propto \left[1 + \frac{n(\bar{y}-\mu)^2}{(n-1)s^2}\right]^{-\frac{n}{2}}$$

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

Marginal Posterior Distribution for μ

- The marginal posterior distribution

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

Therefore,

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

which corresponds to the standard result used for inference on a population mean

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

- The sampling distribution of the pivotal quantity $(\bar{y} - \mu)/(s/\sqrt{n})$ does not depend on the nuisance parameter σ^2, and its posterior distribution does not depend on data.

Posterior Predictive Distribution

- The posterior predictive distribution

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

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

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

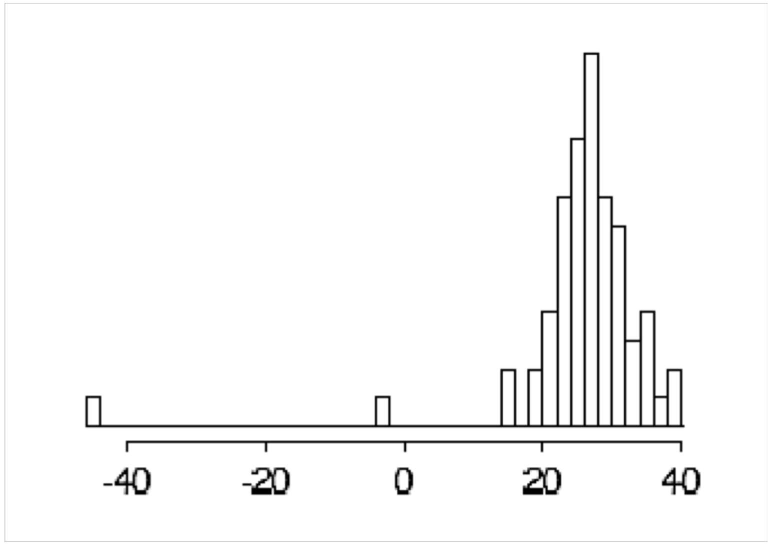
We can derive that

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

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

Example: Estimating the speed of light

- Simon Newcomb set up an experiment in 1882 to measure the speed of light. Newcomb measured the amount of time required for light to travel a distance of 7442 meters. A histogram of Newcomb's 66 measurements is shown in Figure.



The data are recorded as deviations from 24,800 nanoseconds

Example

- We (inappropriately) apply the normal model, assuming that all 66 measurements are independent draws from a normal distribution with mean μ and variance σ^2. The main substantive goal is posterior inference for μ. The mean of the 66 measurements is y = 26.2, and the sample standard deviation is s = 10.8. Assuming the noninformative prior distribution $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$, a 95% central posterior interval for μ is obtained from the t_{65} marginal posterior distribution of μ as $y \pm 1.997s/\sqrt{66} = [23.6, 28.8]$.

A family of conjugate prior distributions

the conjugate prior density must also have the product form $p(\sigma^2)p(\mu|\sigma^2)$:

$$\mu|\sigma^2 \sim N(\mu_0, \frac{\sigma^2}{\kappa_0})$$

$$\sigma^2 \sim \mathrm{Inv} - \chi^2(\nu_0, \sigma_0^2)$$

The joint density is

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma} \frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}} \exp(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu - \mu_0)^2])$$

Conjugate Prior

- This has been labelled as $\mathrm{N}-\mathrm{Inv}-\chi^2(\mu_0, \frac{\sigma_0^2}{\kappa_0}; \nu_0, \sigma_0^2)$ distribution
- its four parameters can be identified as the location and scale of μ and the degrees of freedom and scale of σ^2
- One important thing to note is that with this prior, μ and σ^2 are dependent (i.e. $p(\mu|\sigma^2)$ is a function of σ^2, for example, if σ^2 is large, then a high-variance prior distribution is induced on μ
- This has a different feel from the standard frequentist analysis where $\bar{y}$ and s^2 are independent.

The Posterior Density

- The posterior density satisfies

$$p(\mu,\sigma^2|y) \propto \frac{1}{\sigma}\frac{1}{(\sigma^2)^{\frac{\nu_0}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2+\kappa_0(\mu-\mu_0)^2]\right)$$
$$\times\frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp(-\frac{1}{2\sigma^2}[(n-1)s^2+n(\bar{y}-\mu)^2])$$
$$\propto\frac{1}{\sigma}\frac{1}{(\sigma^2)^{\frac{\nu_n}{2}+1}}\exp\left(-\frac{1}{2\sigma^2}[\nu_n\sigma_n^2+\kappa_n(\mu-\mu_n)^2]\right)$$

The posterior distribution is $\mathrm{N}-\mathrm{Inv}-\chi^2(\mu_n,\frac{\sigma_n^2}{\kappa_n};\nu_n,\sigma_n^2)$

The Posterior Density

where

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

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

The parameters of the posterior distribution combine the prior information and the information contained in the data. For example μ_n is a weighted average of the prior mean and the sample mean, with weights determined by the relative precision of the two pieces of information.

The Conditional Posterior Distribution $p(\mu|\sigma^2, y)$

- By using that $p(\mu|\sigma^2, y) \propto p(\mu, \sigma^2|y)$ with σ as a constant, we get

$$\mu|\sigma^2, y \sim N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

Note that the mean and variance can be written as

$$\mu_n = \frac{\frac{\kappa_0}{\sigma^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}} \qquad \sigma_n^2 = \frac{1}{\frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}}$$

which matches with the fixed variance case discuss earlier.

The Marginal Posterior Distribution $p(\sigma^2|y)$

- $p(\sigma^2|y)$

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

This can be seen by the same way $p(\sigma^2|y)$ was shown in the non-informative prior case or by recognizing the $\mathrm{N}-\mathrm{Inv}-\chi^2$ form of the joint density.

- $p(\mu|y)$

As mentioned before, this can be determined by simulation (see in the next slide). In this case an exact answer can be determined by integrating out σ^2 from the joint density (as in the non-informative case), we get

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

Simulation of $p(\mu|y)$

- we first draw σ^2 from its marginal posterior distribution $p(\sigma^2|y)$,

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

- then draw μ from its normal conditional posterior distribution $p(\mu|\sigma^2, y)$

$$\mu|\sigma^2, y \sim \; N(\mu_n, \frac{\sigma^2}{\kappa_n})$$

using the simulated value of σ^2.

Multinomial Model

- The binomial distribution can be generalized to allow more than two possible outcomes.
- The multinomial sampling distribution is used to describe data for which each observation is one of k possible outcomes.
- If y is the vector of counts of the number of observations of each outcome, then

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

where the sum of the probabilities $\sum_{j=1}^{k} \theta_j$ is 1, and $\sum_{j=1}^{k} y_j = n$ (the number of the observations).

The Prior and Posterior Distribution

- The conjugate prior distribution

Dirichlet: a multivariate generalization of the beta distribution

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

where $\theta_j \in (0,1)$ and $\sum \theta_j = 1$

- The posterior distribution

The resulting posterior distribution for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$.

Prior distribution

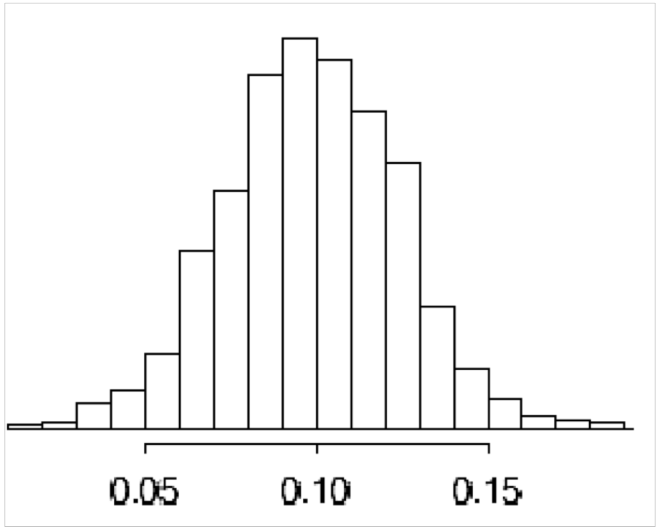
- The prior distribution is mathematically equivalent to a likelihood resulting from $\sum_{j=1}^{k} \alpha_j$ observations with α_j observations of the jth outcome category.
- Noninformative Dirichlet prior distributions:
  - A uniform density is obtained by setting $\alpha_j = 1$ for all j. this distribution assigns equal density to any vector θ satisfying $\sum_{j=1}^{k} \theta_j = 1$.
  - Setting $\alpha_j = 0$ for all j results in an improper prior distribution that is uniform in the $\log(\theta_j)$'s. The resulting posterior distribution is proper if there is at least one observation in each of the k categories,

Example: Pre-election polling

- In late October, 1988, a survey was conducted by CBS News of 1447 adults in the United States to find out their preferences in the upcoming presidential election. Out of 1447 persons, $y_1 = 727$ supported George Bush, $y_2 = 583$ supported Michael Dukakis, and $y_3 = 137$ supported other candidates or expressed no opinion.
- then the data (y_1, y_2, y_3) follow a multinomial distribution, with parameters $(\theta_1, \theta_2, \theta_3)$, the proportions of Bush supporters, Dukakis supporters, and those with no opinion in the survey population.
- An estimand of interest is $\theta_1 - \theta_2$, the population difference in support for the two major candidates

Example

- With a noninformative uniform prior distribution on θ, $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the posterior distribution for $(\theta_1, \theta_2, \theta_3)$ is Dirichlet(728, 584, 138).
- We could compute the posterior distribution of $\theta_1 - \theta_2$ by integration, but it is simpler just to draw 1000 points $(\theta_1, \theta_2, \theta_3)$ from the posterior Dirichlet distribution and then compute θ1 − θ2 for each.



All of the 1000 simulations had $\theta_1 > \theta_2$; thus, the estimated posterior probability that Bush had more support than Dukakis in the survey population is over 99.9%.

Multivariate Normal Model

- y is a vector of length d with mean vector μ (also of length d and $d\times d$ variance matrix Σ), with multivariate normal distribution

$$y|\mu,\Sigma \sim N_d(\mu,\Sigma)$$

- The density of a single observation is

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

- The likelihood of n i.i.d observations is

$$p(y_1,\ldots,y_n|\mu,\Sigma) \propto |\Sigma|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^T\Sigma^{-1}(y_i-\mu)\right)$$

$$= |\Sigma|^{-\frac{n}{2}} \exp(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}S_0))$$

where $\mathrm{tr}(A)$ is the trace of the matrix A (the sum of the diagonal entries) and

$$S_0 = \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T$$

- So the density and likelihood look like what we get in the univariate case, but with matrix and vectors instead.
- Note that most of the inference in this model is a direct analogue to the univariate case. However we need a multivariate analogue to the χ^2 and Inv-χ^2 distributions.

Wishart and Inverse Wishart Distributions

- Wishart distribution ($\text{Wishart}_\nu(\Lambda)$)
- Multivariate analogue of a scaled χ^2 distribution

If $z_1, \ldots, z_\nu \sim N_d(0, \Lambda)$ then

$$\Sigma = \sum_{i=1}^{\nu} z_i z_i^T \sim \text{Wishart}_\nu(\Lambda)$$

like $z_1, \ldots, z_\nu \sim N_d(0, \tau^2)$ then

$$S = \sum_{i=1}^{\nu} z_i^2 \sim \tau^2 \chi_\nu^2$$

Inverse Wishart Distribution

- Inverse Wishart distribution ($\mathrm{Inv}-\mathrm{Wishart}_{\nu}(\Lambda^{-1})$)
- Multivariate analogue of a scaled $\mathrm{Inv}-\chi^2$ distribution

If $\Sigma \sim \mathrm{Wishart}_{\nu}(\Lambda)$ then

$$\Sigma^{-1} \sim \mathrm{Inv}-\mathrm{Wishart}_{\nu}(\Lambda^{-1})$$

Multivariate Normal Models

- Unknown mean μ but known Σ

The conjugate prior distribution for μ is

$$\mu|\Sigma \sim N(\mu_0, \Lambda_0)$$

The posterior density

$$\mu|\Sigma, y \sim N(\mu_n, \Lambda_n)$$

where

$$\mu_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$
$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

Like the univariate case, the posterior mean is a weighted average of the prior mean and the sample average and the posterior precision matrix is the prior precision matrix + data precision matrix.

The Conditional Posterior Distribution

- The marginal posterior distribution of a subset of the parameters, $\mu^{(1)}$ say, is also multivariate normal, with mean vector equal to the appropriate subvector of the posterior mean vector μ_n and variance matrix equal to the appropriate submatrix of Λ_n.
- The conditional posterior distribution of a subset $\mu^{(1)}$ given the values of a second subset $\mu^{(2)}$ is multivariate normal.

$$\mu^{(1)}|\mu^{(2)}, y \sim N\left(\mu_n^{(1)} + \beta^{1|2}\left(\mu^{(2)} - \mu_n^{(2)}\right), \Lambda^{1|2}\right),$$

where the regression coefficients $\beta^{1|2}$ and conditional variance matrix $\Lambda^{1|2}$ are defined by

$$\beta^{1|2} = \Lambda_n^{(12)}\left(\Lambda_n^{(22)}\right)^{-1}$$

$$\Lambda^{1|2} = \Lambda_n^{(11)} - \Lambda_n^{(12)}\left(\Lambda_n^{(22)}\right)^{-1}\Lambda_n^{(21)}$$

Unknown Mean and Variance

- Conjugate Prior

$$\Sigma \sim \mathrm{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1})$$
$$\mu|\Sigma \sim N(\mu_0, \Sigma/\kappa_0)$$

- The posterior distribution satisfies

$$\Sigma|y \sim \mathrm{Inv-Wishart}_{\nu_n}(\Lambda_n^{-1})$$
$$\mu|\Sigma, \mathrm{y} \sim \mathrm{N}(\mu_n, \Sigma/\kappa_n)$$

The Posterior Distribution

where

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y}$$

$$\kappa_n = \kappa_0 + n$$

$$\nu_n = \nu_0 + n$$

$$\Lambda_n = \Lambda_0 + S + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T$$

$$S = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})^T$$

Marginal Distribution

- In addition, it is possible to integrate out the variance matrix showing that

$$\mu|y \sim \mathrm{t}_{\nu_n-\mathrm{d}+1}(\mu_n, \Lambda_n/(\kappa_n(\nu_n - d + 1)))$$

(i.e. multivariate t with $\nu_n - d + 1$ degrees of freedom)

Thanks