Multivariate Normal model

Kwanseok Kim, Gyumin Lee

ESC, YONSEI UNIVERSITY

March 25, 2021

Table of Contents

I. Outline

Outline

- Up until now all of our models have been univariate. However, we are now going to deal with multivariate models.
- This allows us to jointly estimate population means, variances and correlations of a collection of variables.
- We will focus on calculating posterior distributions under semiconjugate prior distributions which works well with Gibbs sampler.

II. Multivariate normal density

Bivariate normal

- $\boldsymbol{Y}_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} = \begin{pmatrix} \text{score on first test} \\ \text{score on second test} \end{pmatrix}$
- Things we might be interested in include the population mean theta,

$$E[\boldsymbol{Y}_i] = \begin{pmatrix} E[Y_{i,1}] \\ E[Y_{i,2}] \end{pmatrix} = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

- and population covariance matrix Σ

$$\Sigma = Cov[\boldsymbol{Y}] = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$$

Bivariate normal

Figure below gives contour plots each one $\boldsymbol{\theta} = (50, 50)^T$, $\sigma_1^2 = 64, \sigma_2^2 = 144$, but value of $\sigma_{1,2}$ varying from -48, 0, 48 from left to right. This implies that each has correlations of -.5, 0, +.5 respectively.

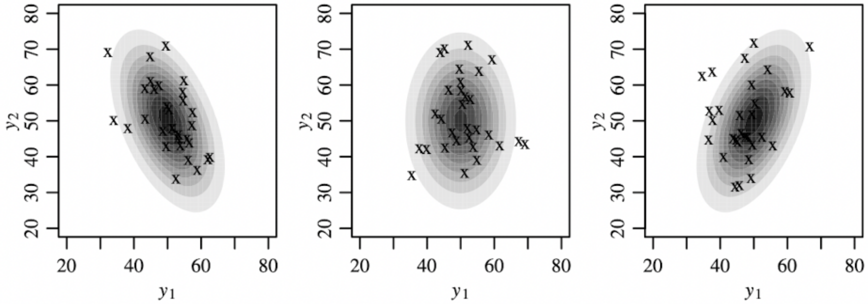


Fig. 7.1. Multivariate normal samples and densities.

Interesting part is that the marginal distribution of each $\boldsymbol{Y_j}$ is a univariate normal w/ mean θ_j and variance σ_j^2

Multivariate normal model

- p-dimensional data vector $\boldsymbol{Y}$

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

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,p} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,p} \\ \vdots & \vdots & & \vdots \\ \sigma_{1,p} & \cdots & \cdots & \sigma_p^2 \end{pmatrix}$$

Some knowledge about matrix

matrix algebra For a matrix $\boldsymbol{A}$

- $|\boldsymbol{A}|$ called determinant, measures how "big" $\boldsymbol{A}$ is
- inverse of $\boldsymbol{A}$ is the matrix $\boldsymbol{A^{-1}}$ s.t. $\boldsymbol{AA^{-1}}$ is $\boldsymbol{I_p}$
- $\boldsymbol{b^T A}$ is 1 x p vector $(\Sigma_{j=1}^{p} b_j a_{j,1}, ..., \Sigma_{j=1}^{p} b_j a_{j,p})$
- $\boldsymbol{b^T A b}$ is a single number $\Sigma_{j=1}^{p} \Sigma_{k=1}^{p} b_j b_k a_{j,k}$.

III. Semiconjugate prior distribution for mean μ

Semiconjugate prior for μ (known Σ)

As in the previous chapters of univariate normal models, we put multivaritae normal model for prior μ

- prior $\boldsymbol{\mu} \sim MVN(\boldsymbol{\mu_0}, \boldsymbol{\Lambda_0})$

$$
\begin{aligned}
p(\boldsymbol{\mu}) &= (2\pi)^{-p/2} \left|\Lambda_0\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^T \Lambda_0^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right\} \\
&= (2\pi)^{-p/2} \left|\Lambda_0\right|^{-1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\Lambda_0^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\Lambda_0^{-1}\boldsymbol{\mu}_0 - \frac{1}{2}\boldsymbol{\mu}_0^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\Lambda_0^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\Lambda_0^{-1}\boldsymbol{\mu}_0\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T\mathbf{A}_0\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{b}_0\right\}
\end{aligned}
$$

where $\boldsymbol{A_0} = \boldsymbol{\Lambda_0^{-1}}$ and $\boldsymbol{b_0} = \boldsymbol{\Lambda_0^{-1}\mu_0}$

Likelihood for multivariate normal

- Our sampling model is $\boldsymbol{Y_1, ..., Y_n} | \boldsymbol{\mu}, \Sigma \overset{\text{i.i.d.}}{\sim} MVN(\boldsymbol{\mu}, \Sigma)$

$$
\begin{aligned}
p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\mu}, \Sigma\right) &= \prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left\{-\left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)/2\right\} \\
&= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)^T \Sigma^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\mu}\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{A}_1 \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{b}_1\right\}
\end{aligned}
$$

where $\boldsymbol{A_1} = n\Sigma^{-1}$ and $\boldsymbol{b_1} = n\Sigma^{-1}\bar{\boldsymbol{y}}$
$\bar{\boldsymbol{y}} = (\frac{1}{n}\Sigma_{i=1}^{n} y_{i,1}, \cdots, \frac{1}{n}\Sigma_{i=1}^{n} y_{i,p})$

Posterior for multivariate model

- By combining the two functions of prior and likelihood, we obtain the following posterior

$$
\begin{aligned}
p\left(\boldsymbol{\mu} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma\right) &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{A}_0 \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{b}_0\right\} \times \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{A}_1 \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{b}_1\right\} \\
&= \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^T \mathbf{A}_n \boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{b}_n\right\}, \text{ where} \\
\mathbf{A}_n &= \mathbf{A}_0 + \mathbf{A}_1 = \Lambda_0^{-1} + n\Sigma^{-1} \text{and} \\
\boldsymbol{b}_n &= \boldsymbol{b}_0 + \boldsymbol{b}_1 = \Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\overline{\boldsymbol{y}}
\end{aligned}
$$

- This implies that the conditional distribution of $\boldsymbol{\mu}$ therefore must be multivariate normal distribution with covariance $\boldsymbol{A_n^{-1}}$ and mean $\boldsymbol{A_n^{-1} b_n}$

Posterior inference

- As we summarize the result from the previous slide,

$$\begin{aligned} \mathrm{Cov}\left[\boldsymbol{\mu} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma\right] &= \Lambda_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1} \\ \mathrm{E}\left[\boldsymbol{\mu} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma\right] &= \boldsymbol{\mu}_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)^{-1} \left(\Lambda_0^{-1}\boldsymbol{\mu}_0 + n\Sigma^{-1}\overline{\boldsymbol{y}}\right) \\ p\left(\boldsymbol{\mu} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \Sigma\right) &= \mathrm{MVN}\left(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n\right) \end{aligned}$$

- Looks a bit complicated, but these are interpretable by the same analogy we learned before.
  1) Posterior precision(inverse variance) is the sum of prior precision and the data precision
  2) Posterior expectation is a weighted average of the prior expectation and the sample mean

IV. Inverse-Wishart distribution

The design of variance-covariance matrix

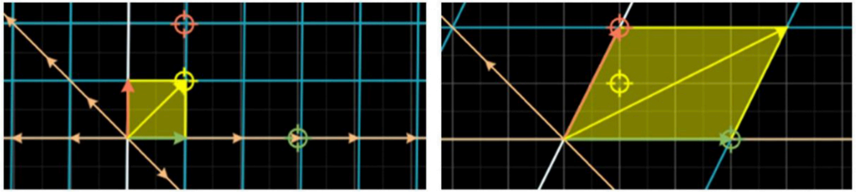
- Just as varaiance σ^2 must be positive, variance-covariance matrix Σ must be positive definite.

$$\boldsymbol{x}'\Sigma\boldsymbol{x} > 0 \text{ for all vectors } \boldsymbol{x}.$$

- Positive definiteness guarantees that all $\sigma_j^2 > 0$ for all j and correlations are between -1 and 1
- Also, covariance matrix should be symmetric so that $\sigma_{j,k} = \sigma_{k,j}$
- Thus, any valid prior distribution should be set of **symmetric, positive definite matrices**. How?

Why positive definite?

- covariance matrix Σ의 positive definite 조건은 그냥 σ^2가 양수여야 한다는 조건의 multivariate 버전인 것!
- Positive definite는 symmetric matrix의 특수한 형태이며, "모든 eigenvalue들이 0보다 크다"와 동치인 조건 $\longrightarrow$ eigenvalue들이 0보다 크다는 것이 왜 중요할까?



Interactive Matrix Visualization (shad.io

$$A = VDV^{-1}$$

$$A(p \times p) = \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} = \begin{bmatrix} v_1 & \cdots & v_p \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_p \end{bmatrix}^{-1}$$

A라는 행렬에 대응하는 선형연산 T 는,
1) 좌표평면을 $v_1, \dots, v_p$를 축으로 하여 $\lambda_1, \dots, \lambda_p$만큼 늘려/줄여주는데,
2) 표준기저 S의 관점에서 보면 e_i 기저벡터가 a_i로 가는 것으로 볼 수 있고,
3) 고유벡터의 기저 V에서 보면, 그냥 기저벡터에 λ_i만큼 곱해주는 것이다.

Empirical covariance matrices

- Sum of squares matrix of a collection of multivariate vectors $z_1, \cdots, z_n$ given by

$$\boldsymbol{\Sigma_{i=1}^n z_i z_i^T = Z^T Z}$$

- $\boldsymbol{Z}$ is nxp matrix where ith row is $\boldsymbol{z}_i^T$, $\boldsymbol{z_i z_i^T}$ is the following pxp matrix

$$\boldsymbol{z}_i \boldsymbol{z}_i^T = \begin{pmatrix} z_{i,1}^2 & z_{i,1}z_{i,2} & \cdots & z_{i,1}z_{i,p} \\ z_{i,2}z_{i,1} & z_{i,2}^2 & \cdots & z_{i,2}z_{i,p} \\ \vdots & & & \vdots \\ z_{i,p}z_{i,1} & z_{i,p}z_{i,2} & \cdots & z_{i,p}^2 \end{pmatrix}$$

- If $\boldsymbol{z}_i$s are samples from a population with zero mean, $\boldsymbol{Z^T Z}/n$ can be thought as sample covariance matrix

$$\frac{1}{n}\left[\mathbf{Z}^T\mathbf{Z}\right]_{j,j} = \frac{1}{n}\sum_{i=1}^n z_{i,j}^2 = s_{j,j} = s_j^2$$
$$\frac{1}{n}\left[\mathbf{Z}^T\mathbf{Z}\right]_{j,k} = \frac{1}{n}\sum_{i=1}^n z_{i,j}z_{i,k} = s_{j,k}$$

Empirical covariance matrices

- If $n > p$ and $\boldsymbol{z_i}$s are linearly independent, then $\boldsymbol{Z^T Z}$ will be positive definite and symmetric.
- Now suggest a given integer ν_0 and a $p * p$ covariance matrix $\boldsymbol{\Phi}_0$
  1. sample $\boldsymbol{z_i}, \cdots, \sim i.i.d.\ MVN(\boldsymbol{0}, \boldsymbol{\Phi}_0)$
  2. calculate $\boldsymbol{Z^T Z} = \Sigma_{i=1}^{\nu_0} \boldsymbol{z_i z_i^T}$
  3. repeat the procedure generating $\boldsymbol{Z_i^T Z_i}$
- $\boldsymbol{Z_i^T Z_i} \sim Wis(\nu_0, \boldsymbol{\Phi_0})$

Properties

- If $\nu_0 > p$, $\boldsymbol{Z^T Z}$ is positive definite with prob. 1
- $\boldsymbol{Z^T Z}$ is symmetric with prob. 1
- $E[\boldsymbol{Z^T Z}] = \nu_0 \boldsymbol{\Phi}_0$

Wishart distribution

- Wishart distribution is a multivariate analogue of the gamma dsitribution

$$\frac{(n-1)S^2}{\sigma^2} \sim X^2(n-1) = \Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right)$$
$$\Leftrightarrow (n-1)s^2 \sim \Gamma\left(\frac{n-1}{2}, \frac{1}{2\sigma^2}\right)$$

- Our prior distribution for the precision $1/\sigma^2$ is a gamma distribution and prior distribution for variance is inverse-gamma distribution
- Similarly, Wishart distribution is a semi-conjugate prior distribution for the precision matirx Σ^{-1} and so, the inverse-Wishart dsitribution is our semi-conjugate prior for covariance matrix Σ

Inverse Wishart distribution as semi-conjugate prior for Σ

- Let's revisit the procedure of creating Wishart distribution
  1. Sample $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{\nu_0} \sim i.i.d.\ MVN(\boldsymbol{0}, \boldsymbol{S_0^{-1}})$
  2. Calculate $\boldsymbol{Z^T Z} = \sum_{i=0}^{\nu_0} z_i z_i^T$
  3. Set $\boldsymbol{\Sigma} = (\boldsymbol{Z^T Z})^{-1}$

$$\Sigma^{-1} \sim \text{Wis}\left(\nu_0, S_0^{-1}\right), E\left[\Sigma^{-1}\right] = \nu_0 S_0^{-1}$$
$$\Sigma \sim \text{Wis}^{-1}\left(\nu_0, S_0^{-1}\right), E[\Sigma] = \frac{1}{\nu_0 - p - 1} S_0$$

- prior 모수 설정 방법!
  1. $\Sigma = \Sigma_0$라는 믿음이 강한 경우: $\nu_0 \uparrow$, $S_0 = (\nu_0 - p - 1)\Sigma_0$로 설정
  2. 믿음이 약하면 $\nu_0 = p + 2$로 설정

  두 경우 모두 Σ가 Σ_0를 중심으로 근처에 설정되도록 함

Full conditional distribution of covariance matrix

- Prior: $\Sigma \sim \text{inv-Wis}(\nu_0, \boldsymbol{S}_0^{-1})$

$$p(\Sigma) = \left[2^{\nu_0 p/2}\pi^{(p)/2}\left|\mathbf{S}_0\right|^{-\nu_0/2}\prod_{j=1}^{p}\Gamma\left(\left[\nu_0+1-j\right]/2\right)\right]^{-1} \times$$
$$|\Sigma|^{-(\nu_0+p+1)/2} \times \exp\left\{-\operatorname{tr}\left(\mathbf{S}_0\Sigma^{-1}\right)/2\right\}.$$

- Likelihood: $\boldsymbol{y}|\boldsymbol{\mu}, \Sigma \sim \text{MVN}(\boldsymbol{\mu}, \Sigma)$

$$p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \mid \boldsymbol{\mu}, \Sigma\right) = (2\pi)^{-np/2}|\Sigma|^{-n/2}\exp\left\{-\sum_{i=1}^{n}\left(\boldsymbol{y}_i-\boldsymbol{\theta}\right)^T\Sigma^{-1}\left(\boldsymbol{y}_i-\boldsymbol{\theta}\right)/2\right\}$$
$$\propto |\Sigma^{-n/2}|\exp\frac{1}{2}tr(S_\mu\Sigma^{-1})$$
$$where\ S_\mu = \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)^T$$

Full conditional distribution of covariance matrix

- Full conditional posterior distribution of Σ: $\Sigma \mid y \sim \text{inv-Wis}(\nu_n + n, [\boldsymbol{S}_0 + \boldsymbol{S}_{\boldsymbol{\mu}}]^{-1})$

$$
\begin{aligned}
& p\left(\Sigma \mid y_1, \cdots y_n, \mu\right) \propto p(\Sigma) p\left(y_1, \cdots y_n \mid \mu, \Sigma\right) \\
& \propto |\Sigma|^{-\left(V_0+p+1\right) / 2} \times \exp \left\{-\operatorname{tr}\left(S_0 \Sigma^{-1}\right) / 2\right\} \times|\Sigma|^{-n / 2} \exp \left(-\frac{1}{2} \operatorname{tr}\left(S_\mu \Sigma^{-1}\right)\right) \\
& =|\Sigma|^{-\left(v_0+p+n+1\right) / 2} \exp \left(-\frac{1}{2} \operatorname{tr}\left(\left[S_0+S_\mu\right] \Sigma^{-1}\right)\right)
\end{aligned}
$$

- Interpretation of the expectation as weighted average

$$
\begin{aligned}
E\left[\Sigma \mid y_1, \cdots y_n, \mu\right] & =\frac{1}{\nu_0+n-p-1}\left(S_0+S_\mu\right) \\
& =\frac{\nu_0-p-1}{\nu_0+n-p-1} \underbrace{\frac{1}{\nu_0-p-1} S_0}+\frac{n}{\nu_0+n-p-1} \underbrace{\frac{1}{n} S_\mu}
\end{aligned}
$$

Summary

- ① Semiconjugate prior for μ (given Σ)

$$\mu \sim \mathrm{MVN}\left(\mu_0, \Lambda_0\right)$$
$$\mu \mid y \sim \mathrm{MVN}\left(\mu_n, \Lambda_n\right)$$
$$\text{where } \begin{cases} \mu_n = \left(\Lambda_0^{-1} + n\Sigma^{-1}\right)\left(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}\right) \\ \Lambda_n = \Lambda_0^{-1} + n\Sigma^{-1} \end{cases}$$

- ② Semiconjugate prior for Σ (given μ)

$$\Sigma \sim \mathrm{Wis}^{-1}\left(\nu_0, S_0^{-1}\right) \quad \left(S_0 = \left(\nu_0 - p - 1\right)\Sigma_0\right)$$
$$\Sigma \mid y \sim \mathrm{Wis}^{-1}\left(\nu_0 + n, \left(S_0 + S_\mu\right)^{-1}\right)$$
$$E\left[\Sigma \mid y\right] = \frac{\nu_0 - p - 1}{\nu_0 + n - p - 1}\Sigma_0 + \frac{n}{\nu_0 + n - p - 1}\left(\frac{1}{n}S_\mu\right)$$

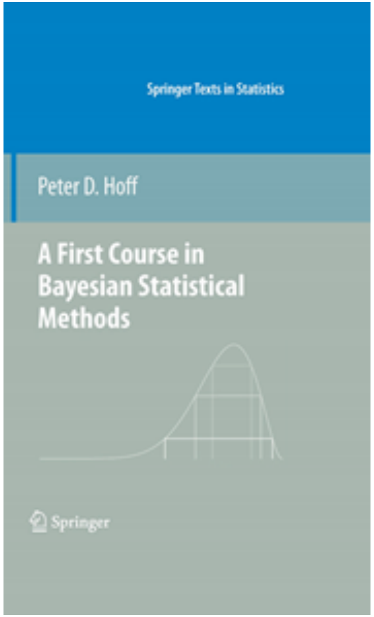
Bayesian Statistics

Ch.7 Multivariate Normal model

이규민

March 25, 2021

FCB



- **A First Course in Bayesian Statistical Methods**
- 목차:
  1. Intro
  2. Belief, probability, and exchangability
  3. One parameter models
  4. Monte Carlo approximation
  5. Normal model
  6. Gibbs sampler
  7. Multivariate normal model
  8. Hierarchical model
  9. Linear regression
  10. Metropolis-Hastings algorithm

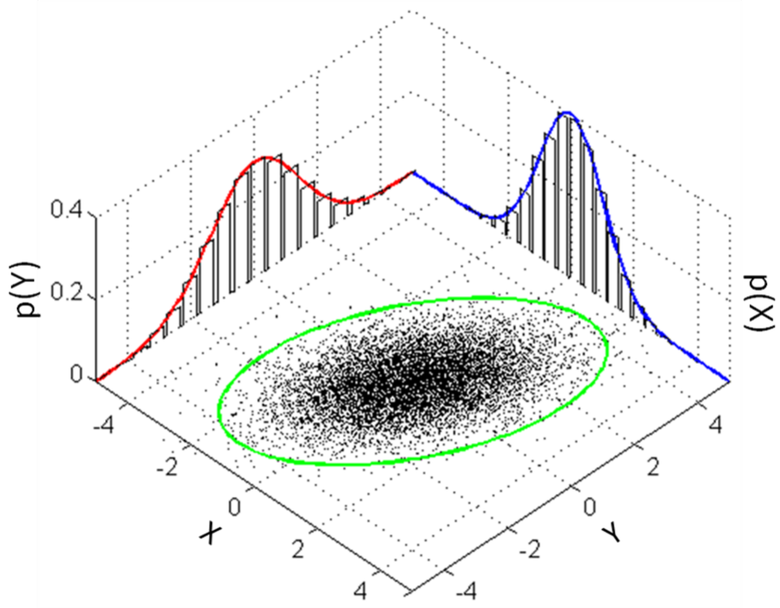
What we did so far...

Full conditional distribution

- $\theta|y_1, ..., y_n, \Sigma \;\sim\; MVN(\mu_n, \Lambda_n)$
- $\Sigma|y_1, ..., y_n, \theta \;\sim\; inv - Wish(\nu_n, S_n^{-1})$

$\rightarrow$ with these, we approximate joint posterior distribution, $\theta, \Sigma|y_1, ..., y_n$

sampling joint distribution by generating dependent sequence of parameters



Gibbs Sampling

sampling joint distribution by generating dependent sequence of parameters

$$\theta_1^{(0)} \rightarrow \theta_2^{(1)} \rightarrow \theta_1^{(1)} \rightarrow ...\theta_2^{(s)} \rightarrow \theta_1^{(s)} \rightarrow ...$$

- 1. Choose a starting value $\theta_1^{(0)}$
- 2. Sample $\theta_2^{(1)}$ from $p(\theta_2|\theta_1^{(0)}, y_1, ..., y_n)$
- 3. Sample $\theta_1^{(1)}$ from $p(\theta_1|\theta_2^{(1)}, y_1, ..., y_n)$
- 4. So on ...

Gibbs Sampling

sampling joint distribution by generating dependent sequence of parameters

$$\Sigma^{(0)} \to \mu^{(1)} \to \Sigma^{(1)} \to ...\mu^{(s)} \to \Sigma^{(s)} \to ...$$

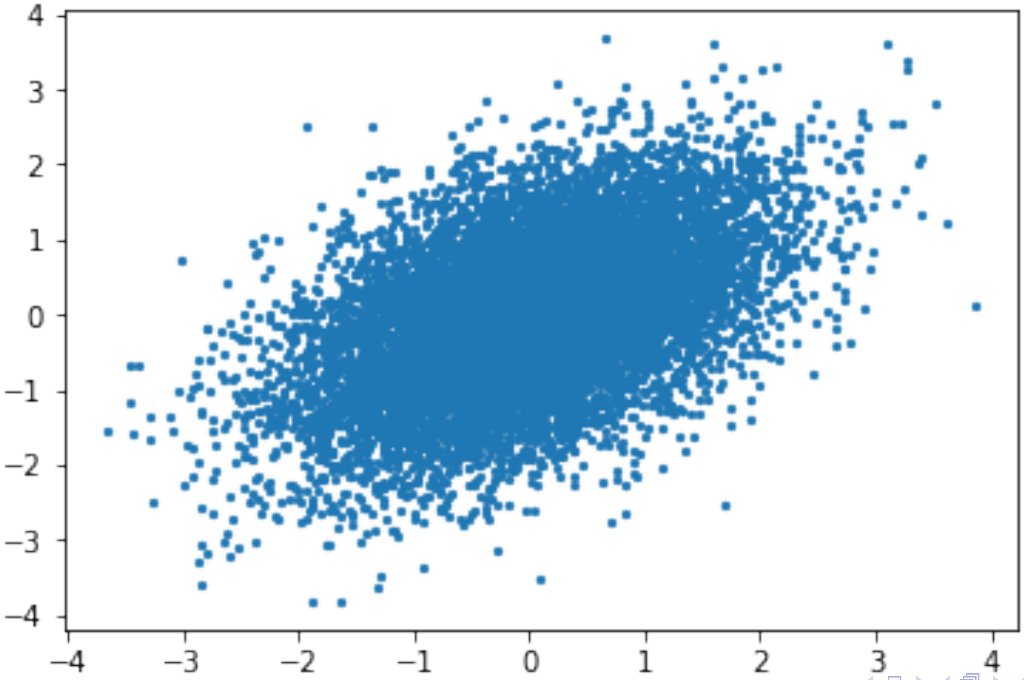
- 1. Choose a starting value $\Sigma^{(0)}$
- 2. Sample $\mu^{(1)}$ from $MVN(\mu_n, \Lambda_n)$ computed from $y_1, ..., y_n$ and $\Sigma^{(0)}$
  - $\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$
  - $\mu_n = \Lambda_n(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$
- 3. Sample $\Sigma^{(1)}$ from $inv - Wish(\nu_n, S_n^{-1})$ computed from $y_1, ..., y_n$ and $\theta^{(1)}$
  - $\nu_n = \nu_0 + n$
  - $S_0 = (\nu_0 - d - 1)\Sigma_0$
  - $S_\mu = \sum_{i=1}^{n} (y_i - \mu)(y_i - \mu)^T$
  - $S_n = S_0 + S_\mu$
- 4. So on ...

Gibbs Sampling

Listing 1: Bivariate normal distribution: true model

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import pearsonr

automatic_samples = np.random.multivariate_normal([0,0], [[1, 0.5], [0.5,1]], 10000)
plt.scatter(automatic_samples[:,0], automatic_samples[:,1], s=5)
```

Gibbs Sampling



Gibbs Sampling

Gibbs sampling by yourself Again bivariate normal model

$$\begin{bmatrix} x_0 \\ x_1 \end{bmatrix} \sim \mathsf{N}\left[\begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{10} & \Sigma_{11} \end{bmatrix} \right]$$

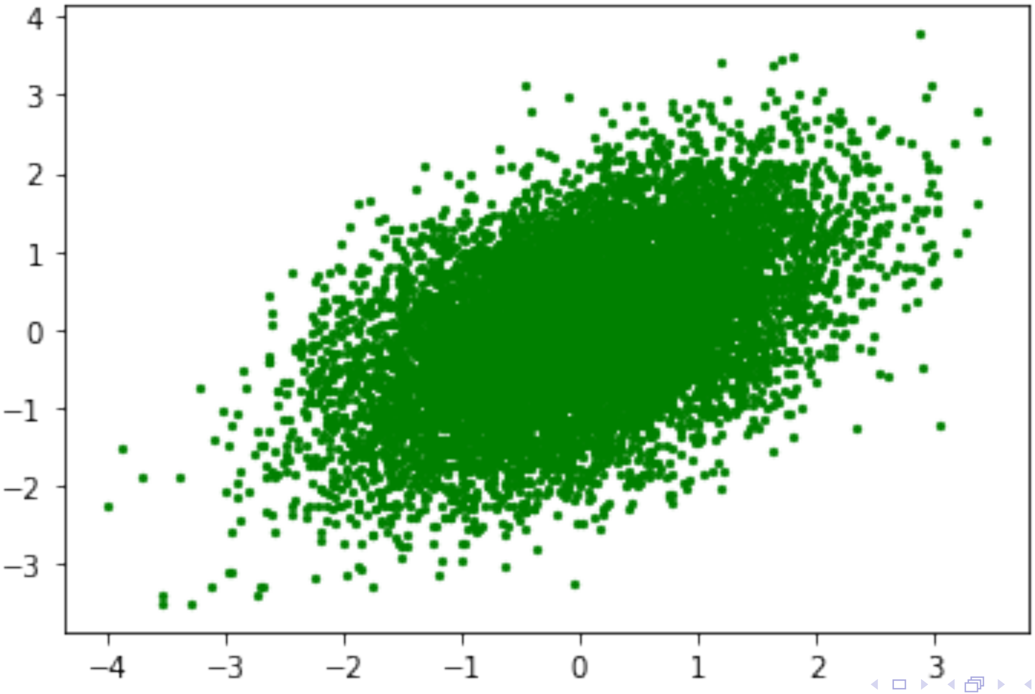
$$p(x_0|x_1) \ \sim \ N(\mu_0 + \Sigma_{01}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{00} - \frac{\Sigma_{01}^2}{\Sigma_{11}})$$

Gibbs Sampling

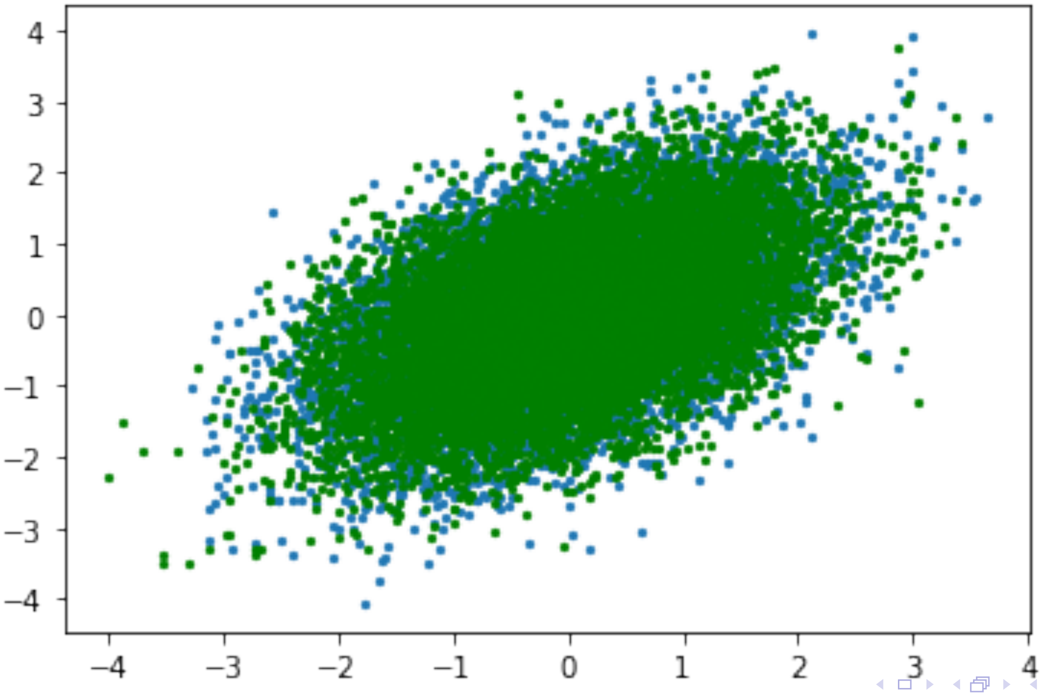
Listing 2: Bivariate normal distribution: generated by Gibbs sampling

```
samples = {'x': [1], 'y': [-1]}

num_samples = 10000

for _ in range(num_samples):
    curr_y = samples['y'][-1]
    new_x = np.random.normal(curr_y/2, np.sqrt(3/4))
    new_y = np.random.normal(new_x/2, np.sqrt(3/4))
    samples['x'].append(new_x)
    samples['y'].append(new_y)

plt.scatter(samples['x'], samples['y'], s=5)
```

Gibbs Sampling



Gibbs Sampling

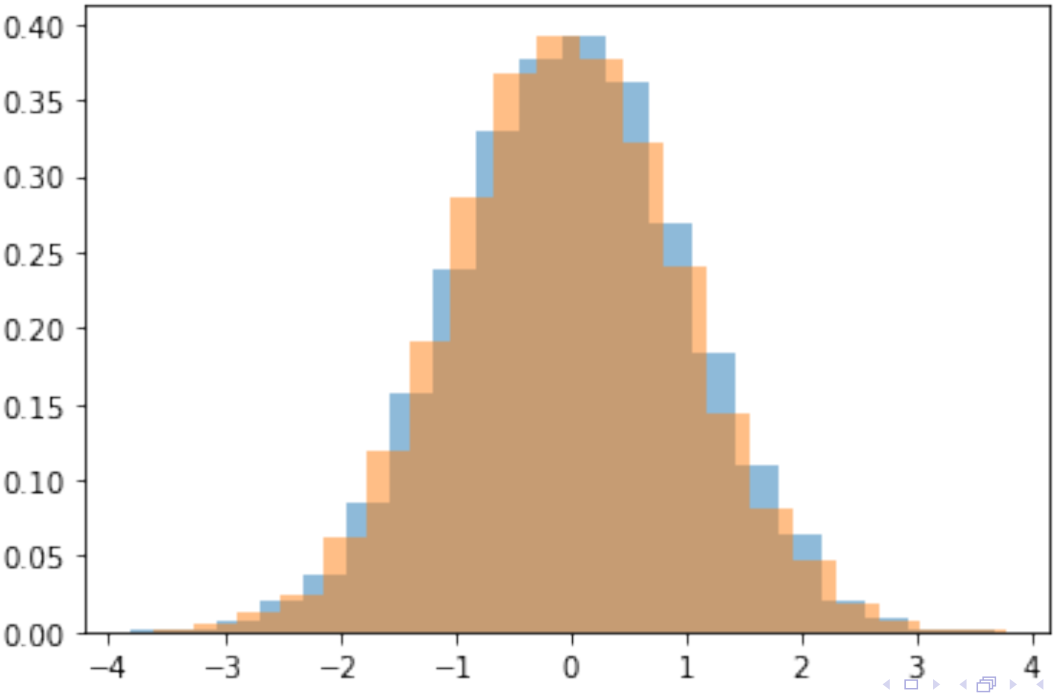


Gibbs Sampling

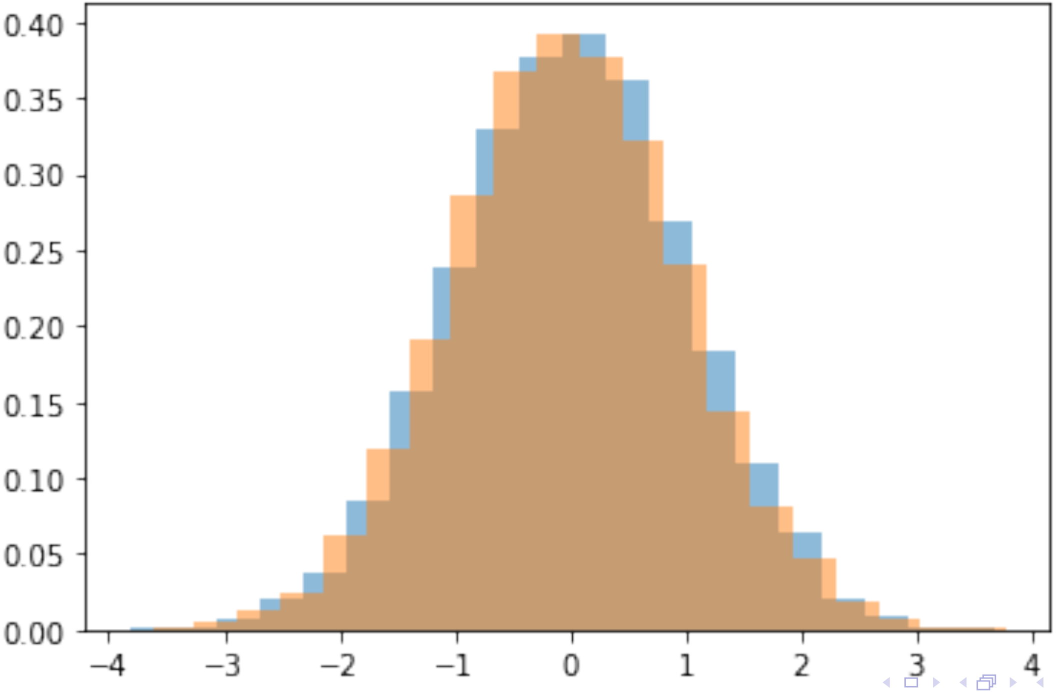
Listing 3: Bivariate normal distribution: comparing histograms

```
plt.hist(automatic_samples[:,0], bins=20, density=True, alpha=0.5)
plt.hist(samples['x'], bins=20, density=True, alpha=0.5)

plt.hist(automatic_samples[:,1], bins=20, density=True, alpha=0.5)
plt.hist(samples['y'], bins=20, density=True, alpha=0.5)
```

Gibbs Sampling



Gibbs Sampling



Gibbs Sampling

Gibbs sampling: data analysis

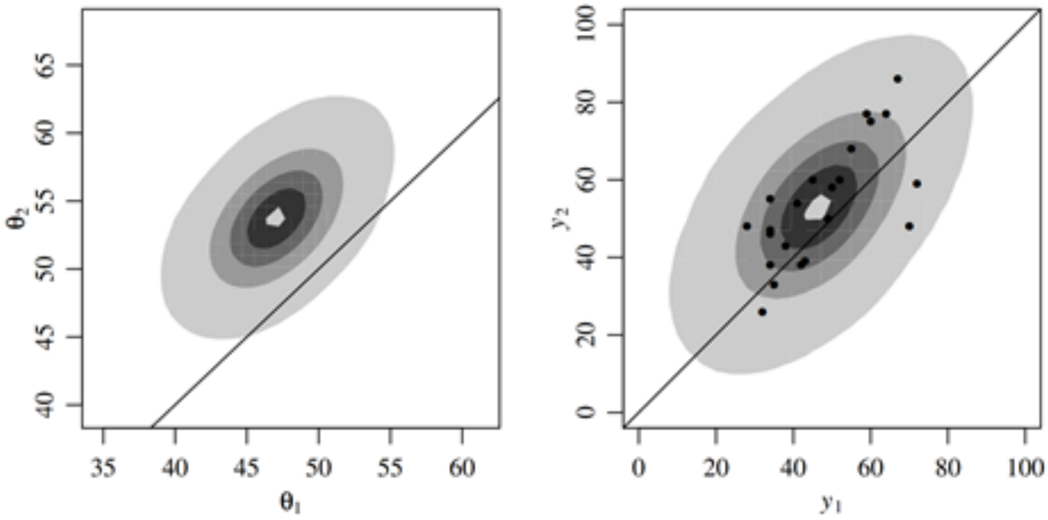


Fig. 7.2. Reading comprehension data and posterior distributions

Gibbs Sampling

Gibbs sampling: look at the whole process!

Gibbs Sampling

Listing 4: Bivariate normal distribution: by yourself!

```
def conditional_sampler(sampling_index, current_x, mean, cov):
    conditioned_index = 1 - sampling_index # 두 r.v. 중고르기
    a = cov[sampling_index, sampling_index] # Sigma00
    b = cov[sampling_index, conditioned_index]  # Sigma01
    c = cov[conditioned_index, conditioned_index]  # Sigma11

    mu = #채워보세요"!"
    sigma = #채워보세요"!"
    new_x = np.copy(current_x)
    new_x[sampling_index] = np.random.randn()*sigma + mu
    # [x_0, x_1] 꼴의 1x2 np.를array return
    return new_x
```

Gibbs Sampling

Listing 5: Bivariate normal distribution: by yourself!

```
def gibbs_sampler(initial_point, num_samples, mean, cov, create_gif=True):
    """
    [input 형태]
    initial_point = [x_0, x_1] = [-9.0, -9.0]
    num_samples = 100
    mean = np.array([0, 0])
    cov = np.array([[10, 3],
                    [3, 5]])
    """
    frames = []  # for GIF
    point = np.array(initial_point)
    samples = np.empty([num_samples+1, 2])  # sampled points
    samples[0] = # 채워보세요!
    tmp_points = np.empty([num_samples, 2]) # inbetween points 중간저장소()
```

Gibbs Sampling

Listing 6: Bivariate normal distribution: by yourself!

```
for i in range(num_samples):
    # 이 for 이loop gibbs sampler 의전부 !
    # point = [x_0, x_1]
    # Sample from p(x_0|x_1)
    point = conditional_sampler(0, point, mean, cov)
    tmp_points[i] = point
    if(create_gif):
        frames.append(plot_samples(samples, i+1, tmp_points, i+1,
        title="Num␣Samples:␣" + str(i)))

    # Sample from p(x_1|x_0)
    point = conditional_sampler(1, point, mean, cov)
    samples[i+1] = point
    if(create_gif):
        frames.append(plot_samples(samples, i+2, tmp_points, i+1,
        title="Num␣Samples:␣" + str(i+1)))
```

NA imputation

	glu	bp	skin	bmi
1	86	68	28	30.2
2	195	70	33	NA
3	77	82	NA	35.8
4	NA	76	43	47.9
5	107	60	NA	NA
6	97	76	27	NA
7	NA	58	31	34.3
8	193	50	16	25.9
9	142	80	15	NA
10	128	78	NA	43.3

NA imputation

Types of missing data

- MCAR: Missing completely at random
- MAR: Missing at random
- MNAR: Missing not at random

NA imputation

NA imputation

- $O_i = (O_1, ..., O_p)^T$ where $O_{i,j} = 1$ for observed data, 0 for o.w.
- Assume "Missing at random" : $O_i$and Y_i are independent
- So O_i doesn't depend on θ or Σ

$$p(o_i, y_{i,j} : o_{i,j} = 1|\theta, \Sigma) = p(o_i) \times p(y_{i,j} : o_{i,j} = 1|\theta, \Sigma) \tag{1}$$

$$= p(o_i) \times \int p(y_{i,j}|\theta, \Sigma) \prod_{y_{i,j}:o_{i,j}=0} dy_{i,j} \tag{2}$$

NA imputation

Again Gibbs sampling

이제는 NA로 몰랐던 자료도 sampling!

$Y_{obs} = y_{i,j} : o_{i,j} = 1$

$Y_{miss} = y_{i,j} : o_{i,j} = 0$

$$\Sigma^{(0)}, Y_{miss}^{(0)} \to \theta^{(1)} \to \Sigma^{(1)} \to Y_{miss}^{(1)} \to \ldots \to \theta^{(s)} \to \Sigma^{(s)} \to Y_{miss}^{(s)} \to \ldots$$

- 1. Choose a starting value $\Sigma^{(0)}, Y_{miss}^{(0)}$
- 2. Sample $\theta^{(1)}$ from $p(\theta|\Sigma^{(0)}, Y_{miss}^{(0)}, Y_{obs})$
- 3. Sample $\Sigma^{(1)}$ from $p(\Sigma|\theta^{(1)}, Y_{miss}^{(0)}, Y_{obs})$
- 4. Sample $Y_{miss}^{(1)}$ from $p(Y_{miss}|\theta^{(1)}, \Sigma^{(1)}, Y_{obs})$
- 5. So on ...