STAT3016/4116/7016:
Introduction to Bayesian Data Analysis

RSFASS, College of Business and Economics, ANU

The Multivariate Normal Model

Introduction

The models we have looked at so far are univariate models.

Rarely does a data set consist of a single measurement. Usually we are interested in multiple outcome measurements.

We need a multivariate model to jointly estimate population means, variance and correlations of a collection of variables.

Example

e.g. algebra test score, calculus test score — multivariate model. this is a

The effectiveness of a new teaching method is evaluated on a sample of 22 primary school students. Let $Y_{i,1}$ denote the pre-instructional score and let $Y_{i,2}$ denote the post-instructional score. For each student, we have a vector of observations $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2})$.

$Y_{i,2}$ & $Y_{i,1}$ are correlated. If independent, then we just use univariate

We might be interested in: pre, post

$$E[\mathbf{Y}] = (E[\mathbf{Y}_1], E[[\mathbf{Y}_2]) = (\theta_1, \theta_2)$$

variance is a matrix instead

$$\Sigma = Cov[\mathbf{Y}] = \begin{pmatrix} Var[\mathbf{Y}_1] & Cov(\mathbf{Y}_1, \mathbf{Y}_2) \\ Cov(\mathbf{Y}_1, \mathbf{Y}_2) & Var[\mathbf{Y}_2] \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{pmatrix}$$

covariance matrix

Note the correlation coefficient: $\rho_{1,2} = \sigma_{1,2} / \sqrt{\sigma_1^2 \times \sigma_2^2}$

$\sigma_{12} = \sigma_{21}$

interest: e.g. $Pr(\theta_1 > \theta_2 \mid data)$

The multivariate normal density

The multivariate normal density is completely described by its first and second order moments. Σ (θ_1, θ_2)

A p-dimensional data vector $\mathbf{Y}$ has a multivariate normal distribution if its sampling density is given by:

recall: $p(y|\theta,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\theta)^2}{2\sigma^2}\right)$

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

where $|\Sigma|$ is the determinant of Σ. for a univariate case the determinant is just σ or σ^2

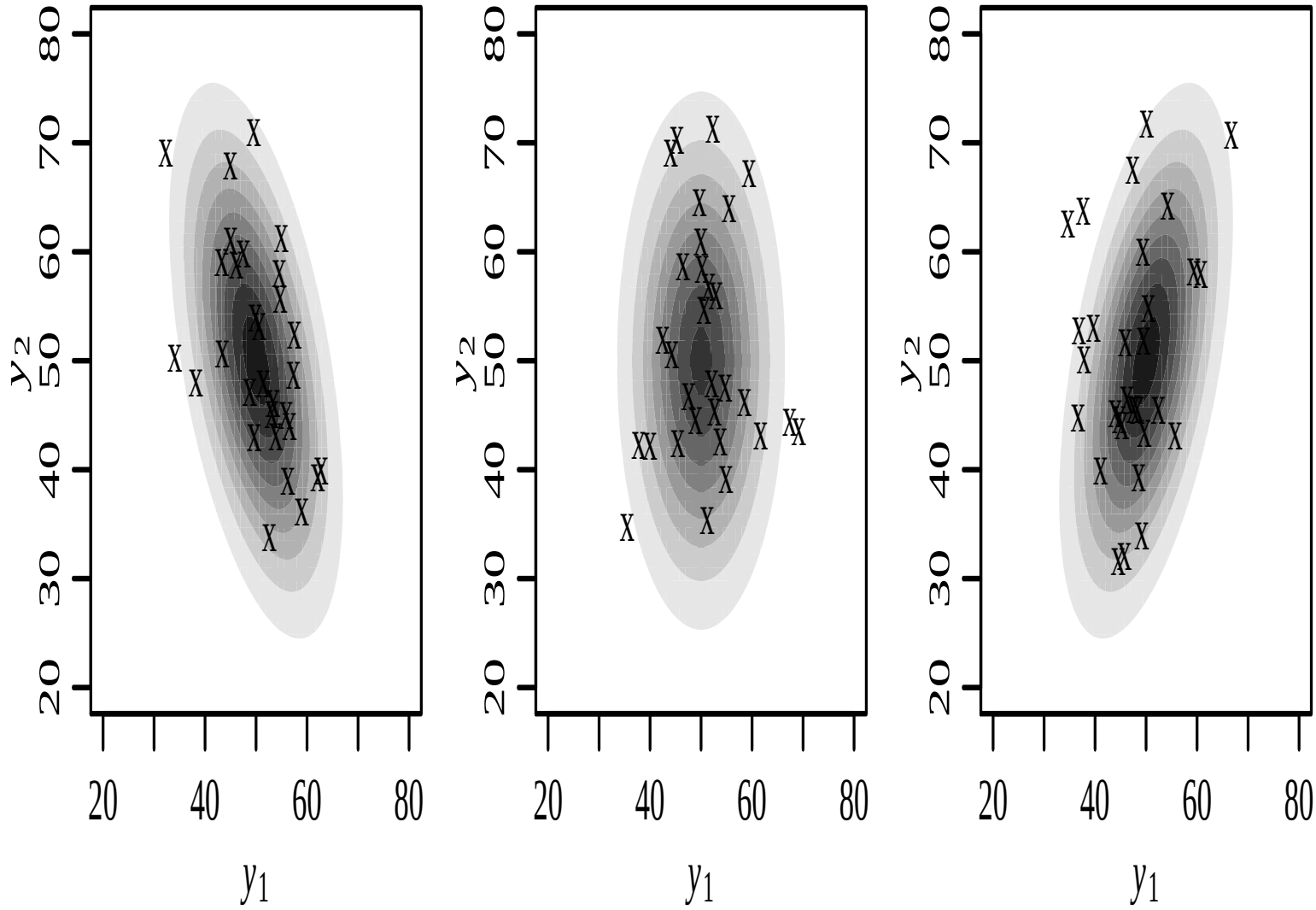
The marginal distribution of each variable Y_j is a univariate normal distribution, with mean θ_j and variance σ_j^2.

$Y_1 \sim N(\theta_1, \sigma_1^2)$
$Y_2 \sim N(\theta_2, \sigma_2^2)$
$\Rightarrow$ not necessarily to be 1

- det()
- t()
- solve()

The multivariate normal density

dark → higher joint density



$\boldsymbol{\theta} = (50, 50)$, $\sigma_1^2 = 64$, $\sigma_2^2 = 144$, $\rho_{1,2} = -0.5, 0, 0.5$

A semiconjugate prior distribution for the mean

Λ : Lambda

Consider a multivariate normal prior for θ

multi. var. normal

$$p(\theta) = MVN(\mu_0, \Lambda_0)$$ ← we know

Sampling model: $\{\mathbf{Y}_1, ..., \mathbf{Y}_n | \theta, \Sigma\}$ i.i.d MVN (θ, Σ)

we have data sampling model (assuming Σ is known)

(conjugate prior for θ)

Derive the posterior distribution $p(\theta | \mathbf{y}_1, ..., \mathbf{y}_n, \Sigma)$. What is the posterior mean and posterior variance of θ? Compare to the univariate case.

(show this later)

follows $MVN(\mu_n, \Lambda_n)$

$p(\theta) = MVN(\mu_0, \Lambda_0)$

$p(y_1, \dots, y_n \mid \theta, \Sigma) = MVN(\theta, \Sigma)$

$p(\theta) = (2\pi)^{-\frac{p}{2}} |\Lambda_0|^{-\frac{1}{2}} \exp\left(\frac{-1}{2}(\theta - \mu_0)^T \Lambda_0^{-1} (\theta - \mu_0)\right)$

$p(y_1, \dots, y_n \mid \theta, \Sigma) = \prod_{i=1}^{n} (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} (y_i - \theta)^T \Sigma^{-1} (y_i - \theta)\right)$

want to find.

$p(\theta \mid y_1, \dots, y_n, \Sigma) \propto ?$

Let $A_0 = \Lambda_0^{-1}$

$b_0 = \Lambda_0^{-1} \mu_0$

$= p(\theta) \cdot p(y_1, \dots, y_n \mid \theta, \Sigma)$

$= \exp\left[-\frac{1}{2}(\theta - \mu_0)^T \Lambda_0^{-1} (\theta - \mu_0) + \sum (y_i - \theta)^T \Sigma^{-1} (y_i - \theta)\right]$

$\propto \exp\left[-\frac{1}{2}\theta^T A_0 \theta + \theta^T b_0 + \theta^T \Sigma^{-1} \bar{y} \cdot n - \frac{1}{2}\theta^T \Sigma^{-1} \theta\right]$ (dropped terms irrelevant terms w.r.t. θ)

actually 2 $\frac{1}{2}$ terms

$A_1 = n\Sigma^{-1}$

$b_1 = n\Sigma^{-1} \cdot \bar{y}$

further define

$A_0 + A_1 = A_n$

$b_0 + b_1 = b_n$

$= \exp\left[-\frac{1}{2}\theta^T (A_0 + A_1)\theta + \theta^T (b_0 + b_1)\right]$

$= \exp\left(-\frac{1}{2}\theta^T A_n \theta + \theta^T b_n\right)$

So we again have a MVN.

We completed a square here.

$$\propto \exp\left(-\frac{1}{2}(\theta - A_n^{-1}b_n)^T A_n (\theta - A_n^{-1}b_n)\right)$$

$A_n^{-1}b_n$ ↓ mean

$$E(\theta \mid Y, \Sigma) = A_n^{-1}b_n = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})$$

$$Cov(\theta \mid Y, \Sigma) = A_n^{-1} = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$$

- How to generate random samples from MVN?

use "mvtnorm" package.

use "mvrnorm()" function to do so.

Multivariate normal with unknown mean and variance

In the univariate normal model setting, we used an inverse-gamma prior distribution on σ^2 which is the semi-conjugate prior.

The analogue in a multivariate normal model is the inverse-Wishart distribution:

integer positive

$p \times p$ matrix

"multivariate inverse-gamma"

constant (nothing to do with σ^2)

$$p(\Sigma|\nu_0, \mathbf{S_0}^{-1}) = \left[2^{\nu_0 p/2}\pi^{p(p-1)/4}|\mathbf{S}_0|^{-\nu_0/2}\prod_{j=1}^{p}\Gamma([\nu_0+1-j]/2)\right]^{-1} \times |\Sigma|^{-(\nu_0+p+1)/2} \times \exp\left\{\text{tr}(\mathbf{S_0}\Sigma^{-1})/2\right\}$$

So positive definite $\Rightarrow$ in order to be invertible.

trace : sum of diagonal elements

$$E[\Sigma] = \frac{1}{\nu_0 - p - 1}\mathbf{S_0}$$

if $\nu_0 = p+2$ i.e. a very weak informative prior then $E[\Sigma] \to S_0$

A prior distribution for the variance

To sample a covariance matrix Σ from an inverse-Wishart$(\nu_0, \mathbf{S_0}^{-1})$ distribution:

each z_i is a $p \times 1$ vector

1. sample $z_1, ..., z_{\nu_0} \overset{\text{iid}}{\sim} \text{multivariate normal}(\mathbf{0}, \mathbf{S_0^{-1}})$
2. calculate $\mathbf{Z}^T\mathbf{Z} = \sum_{i=1}^{\nu_0} z_i z_i^T$
3. set $\Sigma = (\mathbf{Z}^T\mathbf{Z})^{-1}$

↘ inverse of Wishart

- comparison with univariate case.

  $Z \sim N(0,1)$

  $Z^2 \sim$ Gamma

  $\frac{1}{Z^2} \sim$ inverse Gamma

- In R, use package MCMC

  riwish()

  to generate inverse-Wishart number.

Full conditional distribution of the covariance matrix

$$E[\Sigma | y_1, \dots, y_n, \theta] = \frac{1}{\nu_0 + n - p - 1}(S_0 + S_\theta)$$

$$= \frac{\nu_0 - p - 1}{\nu_0 + n - p - 1} \frac{1}{\nu_0 - p - 1} S_0 + \frac{n}{\nu_0 + n - p - 1} \frac{1}{n} S_\theta$$

Show that the conditional posterior distribution $p(\Sigma|\mathbf{y}_1, ..., \mathbf{y}_n, \theta)$ is inverse-Wishart$(\nu_0 + n, [\mathbf{S_0} + \mathbf{S}_\theta]^{-1})$. Interpret the parameters of this distribution. What is the posterior mean $\mathrm{E}[\Sigma|\mathbf{y}_1, ..., \mathbf{y}_n, \theta]$?

Hint: To simplify the sampling model expression, note that $\sum_{k=1}^{K} \mathbf{b}_k^T \mathbf{A} \mathbf{b}_k = tr(\mathbf{B}^T \mathbf{B} \mathbf{A})$, where $\mathbf{B}$ is the matrix whose kth row is $\mathbf{b}_k^T$.

$\Sigma \sim$ inverse-Wishart(ν_0, S_0^{-1})

$E(\Sigma) = \frac{1}{\nu_0 - p - 1} S_0$

$[\Sigma | y_1, \dots, y_n, \theta]$ $\sim$ inverse-Wishart $(\nu_0 + n, [S_0 + S_\theta]^{-1})$

$\nu_0 = p + 2 \quad S_0 = (\nu_0 - p - 1)\Sigma_0$

$p(\Sigma | y_1, \dots, y_n, \theta) \propto p(\Sigma | \nu_0, S_0^{-1}) \cdot p(y_1, \dots, y_n | \theta, \Sigma)$

prior: $|\Sigma|^{-(\nu_0 + p + 1)/2} \cdot e^{-\mathrm{tr}(S_0 \Sigma^{-1})/2}$ likelihood: $|\Sigma|^{-\frac{n}{2}} \cdot e^{-\sum_{i=1}^{n} (y_i - \theta)\Sigma^{-1}(y_i - \theta)^T}$

$\propto |\Sigma|^{-(\nu_0 + n + p + 1)/2} \exp\{-\mathrm{tr}([S_0 + S_\theta]\Sigma^{-1})/2\}$

Gibbs sampling of the mean and covariance

$p(\theta | y_1, \ldots, y_n, \Sigma) \sim MVN(\mu_n, \Lambda_n)$

$p(\Sigma | y_1, \ldots, y_n, \theta) \sim Inv\text{-}Wishart(\nu_n, S_n^{-1})$

Full conditionals for θ & Σ respectively

Using the conditional distributions of θ and Σ derived above, construct a Gibbs sampler to provide us with an MCMC approximation to the joint distribution $p(\theta, \Sigma | \mathbf{y}_1, ..., \mathbf{y}_n)$

Example - reading comprehension

Model the 22 pairs of scores as i.i.d samples from a multivariate normal distribution. The exam was designed to give average scores of around 50 out of 100. The scores cannot be less than 0 or greater than 100 so set the prior variance such that $Pr(\theta_j \notin [0, 100]) \approx 0.05$.

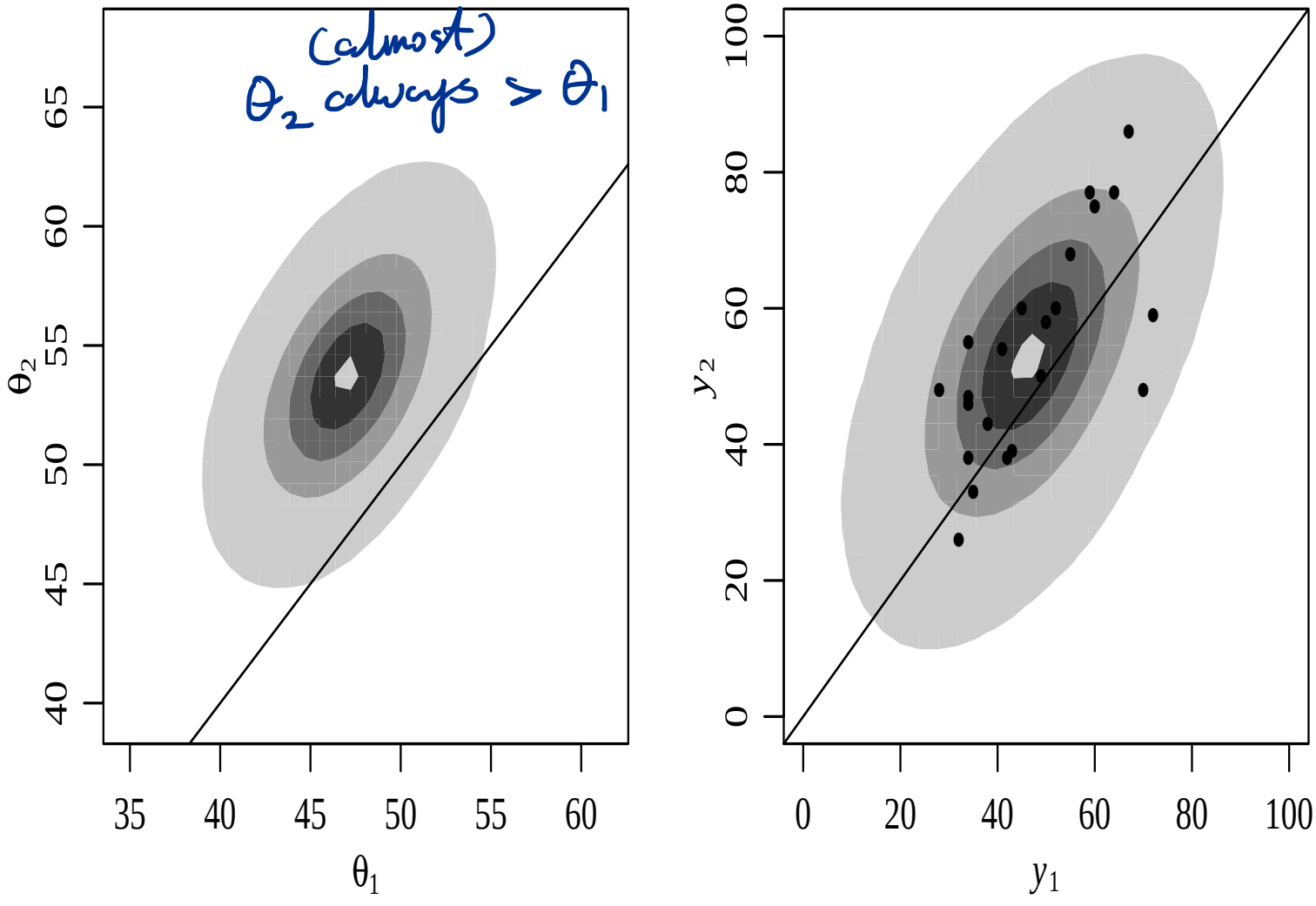
The two exams are measuring similar things, so set the prior correlation to be 0.5.

Take $\mathbf{S}_0$ to be the same as Λ_0, but only loosely center Σ around this value by taking $\nu_0 = p + 2 = 4$.

Q1: What is the posterior probability that the average score on the second exam is greater than the average score on the first exam? Estimate the average improvement in the population.

Q2: What is the probability that a randomly selected child will score higher on the second exam than on the first? Compare you answer to Q1 above and note any difference.

Example - reading comprehension



Example - reading comprehension

```
Y<-Y.reading

mu0<-c(50,50)
L0<-matrix( c(625,312.5,312.5,625),nrow=2,ncol=2)

nu0<-4
S0<-matrix( c(625,312.5,312.5,625),nrow=2,ncol=2)

n<-dim(Y)[1] ; ybar<-apply(Y,2,mean)
ybar
Sigma
Sigma<-cov(Y) ; THETA<-SIGMA<-NULL
YS<-NULL
set.seed(1)
```

Example - reading comprehension

```
for(s in 1:5000)
{ ###update theta
  Ln<-solve( solve(L0) + n*solve(Sigma) )
  mun<-Ln%*%( solve(L0)%*%mu0 + n*solve(Sigma)%*%ybar )
  theta<-rmvnorm(1,mun,Ln)
  ###

  ###update Sigma
  Sn<- S0 + ( t(Y)-c(theta) )%*%t( t(Y)-c(theta) )
  Sigma<-rinvwish(1,nu0+n,solve(Sn))

 ###
  YS<-rbind(YS,rmvnorm(1,theta,Sigma))
  ### save results
  THETA<-rbind(THETA,theta) ; SIGMA<-rbind(SIGMA,c(Sigma))
  ###
  cat(s,round(theta,2),round(c(Sigma),2),"\n")
}
```

Example - reading comprehension

```
> quantile(  SIGMA[,2]/sqrt(SIGMA[,1]*SIGMA[,4]),
      prob=c(.025,.5,.975) )
     2.5%       50%     97.5%
0.3977817 0.6883967 0.8476822
> quantile(   THETA[,2]-THETA[,1], prob=c(.025,.5,.975) )
     2.5%       50%     97.5%
 1.140972  6.688711 11.772220
> mean( THETA[,2]-THETA[,1])
[1] 6.615545
> mean( THETA[,2]>THETA[,1])
[1] 0.9912018
> mean(YS[,2]>YS[,1])
[1] 0.7092581
```

how much improved

significantly improved (on average)

for individual student, new teaching method does something but prob. is not that HIGH.

variability!
① predictive draws of θ
② sampling

(posterior predictive draw has more variability)

Exercise

The data set `Pima.tr` contains data on women of Pima Indian heritage and living near Phoenix, Arizona, who were tested for diabetes according to World Health Organization criteria. The data were collected by the US National Institute of Diabetes and Digestive and Kidney Diseases. There are 532 records in the data set. The variables in the data set are:

```
npreg - number of pregnancies
glu - plasma glucose concentration in an oral glucose
     tolerance test.
bp - diastolic blood pressure (mm Hg).
skin - triceps skin fold thickness (mm).
bmi - body mass index (weight in kg/(height in m)\^2).
ped - diabetes pedigree function.
age - age in years.
type - Yes or No, for diabetic according to WHO criteria.
```

Model the continuous variables using a multivariate normal distribution.

Estimate the correlations between each pair of variables.

Missing data and imputation

Consider the following data set on health-related measurements on 200 women of Pima Indian heritage living near Phoenix Arizona.

	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

We manually set some values ton NA to practice imputation.

Survey failure.
Privacy.

1. Missing at random. The entry does not depend on the missing data.
2. Missing at not random. Fo example, income too high, they don't wanna tell.

The 'NAs' stand for "not available". What are some reasons for missing data in this data set??

Missing data and imputation

Please don't brutally delete the whole entry due to one missing data. Otherwise the analysis would be. Less powerful.

How should we conduct parameter estimation in the presence of missing data?

Consider multivariate normal data, so we want posterior inference on θ and Σ which depends on the sampling model $\prod_{i=1}^{n} p(\mathbf{y}_i|\boldsymbol{\theta}, \Sigma)$, but components of $\mathbf{y}_i$ are missing.

Instead, do imputation, fill up the missing value with the mean.

Drawback: if we fill up the missing with a fixed value, that means we ignored the uncertainty.

Also, mean could be a very poor estimate.

That's why we introduce the multiple imputation with Bayesian method.

Introduction to multiple imputation

Let $\mathbf{O}_{i,j}$ be the matrix in which $o_{i,j} = 1$ if $Y_{i,j}$ is observed and $o_{i,j} = 0$ if $Y_{i,j}$ is missing. (i=1,...,n; j=1,...,p). The data matrix $\mathbf{Y}$ consists of two parts:

- $\mathbf{Y}_{\text{obs}} = \{y_{i,j} : o_{i,j} = 1\}$
- $\mathbf{Y}_{\text{mis}} = \{y_{i,j} : o_{i,j} = 0\}$

We want to draw posterior inference from the distribution $p(\theta, \Sigma|\mathbf{Y}_{\text{obs}})$.

Introduction to multiple imputation

[Rubin (1978)]. Multiple imputations in sample surveys - a phenomenological Bayesian approach to non response. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 20-34.]

$$p(\theta|\mathbf{Y}_{\text{obs}}) = \int p(\theta, \mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}})d\mathbf{Y}_{\text{mis}}$$
$$= \int p(\theta|\mathbf{Y}_{\text{mis}}, \mathbf{Y}_{\text{obs}})p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}})d\mathbf{Y}_{\text{mis}}$$

So the posterior distribution of θ can be simulated by first drawing the missing values $\mathbf{Y}_{\text{mis}}^{(d)}$ from their joint posterior *predictive* distribution $p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}})$, using the posterior draws of $p(\mathbf{Y}_{\text{mis}})$ to complete the data set, and then drawing θ from its complete-data posterior distribution.

Multivariate normal model with missing data

We can treat $\mathbf{Y}_{\text{mis}}$ as unknown quantities (like θ and Σ) that we need to estimate. If we can assume that our data follow a multivariate normal distribution, we can simply add one step to the Gibbs sampler for the multivariate normal model with complete data.

Given starting values $\left\{\Sigma^{(0)}, \mathbf{Y}_{\text{mis}}^{(0)}\right\}$, generate $\left\{\theta^{(s+1)}, \Sigma^{(s+1)}, \mathbf{Y}_{\text{mis}}^{(s+1)}\right\}$ from $\left\{\theta^{(s)}, \Sigma^{(s)}, \mathbf{Y}_{\text{mis}}^{(s)}\right\}$ by

1. sampling $\theta^{(s+1)}$ from $p(\theta|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(s)}, \Sigma^{(s)})$
2. sampling $\Sigma^{(s+1)}$ from $p(\Sigma|\mathbf{Y}_{\text{obs}}, \mathbf{Y}_{\text{mis}}^{(s)}, \theta^{(s+1)})$
3. sampling $\mathbf{Y}_{\text{mis}}^{(s+1)}$ from $p(\mathbf{Y}_{\text{mis}}|\mathbf{Y}_{\text{obs}}, \theta^{(s+1)}, \Sigma^{(s+1)})$ (like posterior predictive distribution.)

Sample missing data one by one.

Multivariate normal model - conditional distributions

Let $\mathbf{y} \sim MVN(\theta, \Sigma)$ and let $\mathbf{y}_{[a]}$ and $\mathbf{y}_{[b]}$ be a partition of $\mathbf{y}$. Then the distribution of $\mathbf{y}_{[b]}$ conditional on $\mathbf{y}_{[a]}$ is still follows

$$\{\mathbf{y}_{[b]}|\mathbf{y}_{[a]}, \theta, \Sigma\} \sim MVN(\theta_{b|a}, \Sigma_{b|a})$$

where

$\theta_{b|a} = \theta_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\mathbf{y}_{[a]} - \theta_{[a]})$

$\Sigma_{b|a} = \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}$

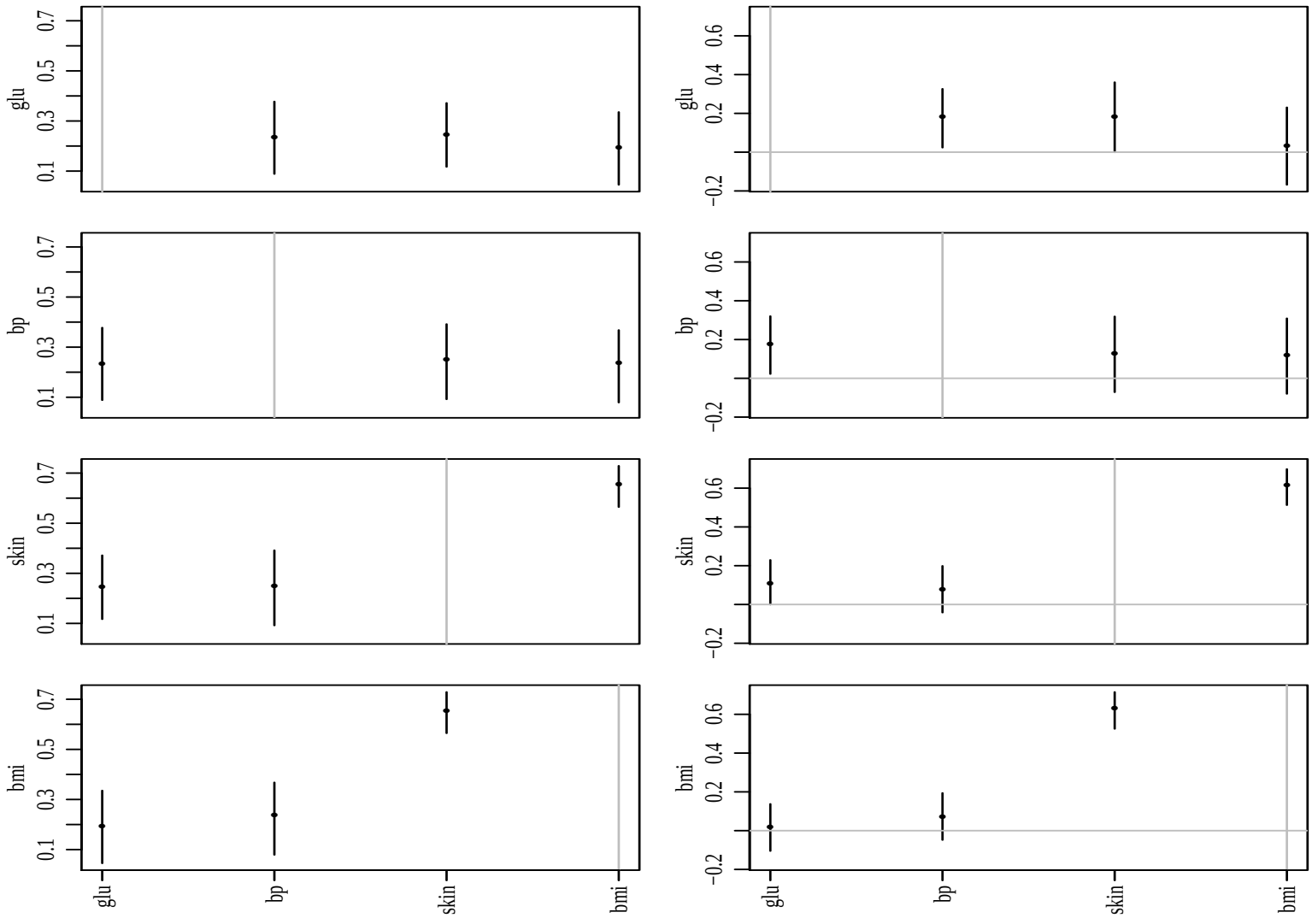
$$\theta = \begin{pmatrix} \theta_a \\ \theta_b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{[a,a]} & \Sigma_{[a,b]} \\ \Sigma_{[b,a]} & \Sigma_{[b,b]} \end{pmatrix}$$

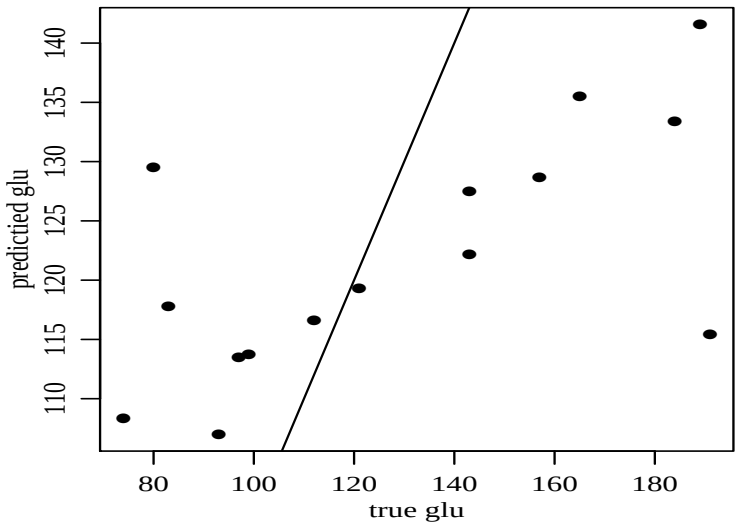
Health data example

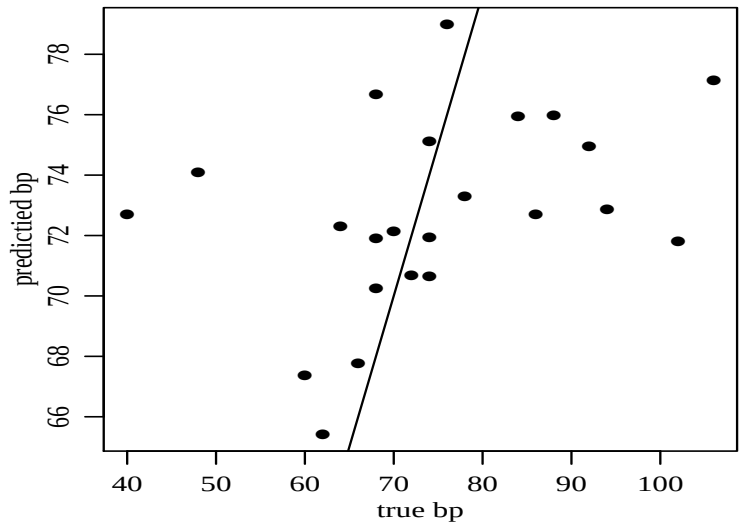
Let's implement the Gibbs sampler for multivariate normal data to impute the missing values in the health data set. Assess the relationships between the variables in the health data set using your imputed data.

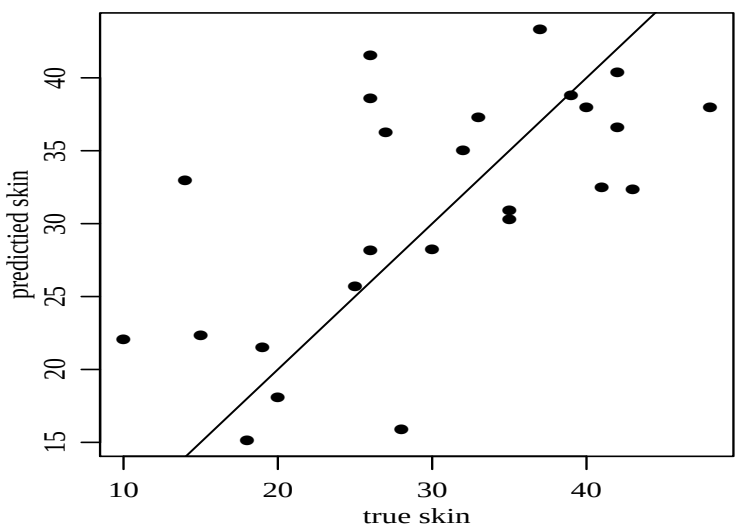
Health data example

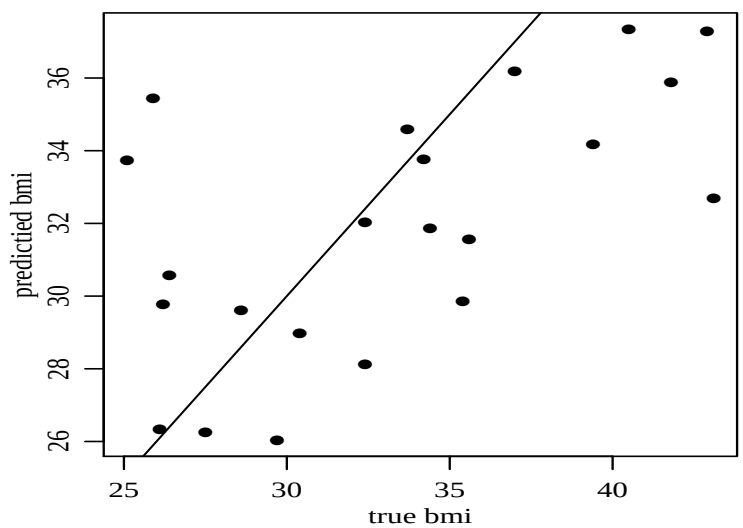


Health data example









Health data example

Suppose in addition to the four continuous variables (`glu, bp, skin, bmi`) , there is a fifth variable (called `med`) which $=1$ if the woman takes chronic medication and $= 0$ otherwise. This variable also has missing values.

Modify the algorithm on slide 21 to allow for imputation of missing values of `med`.