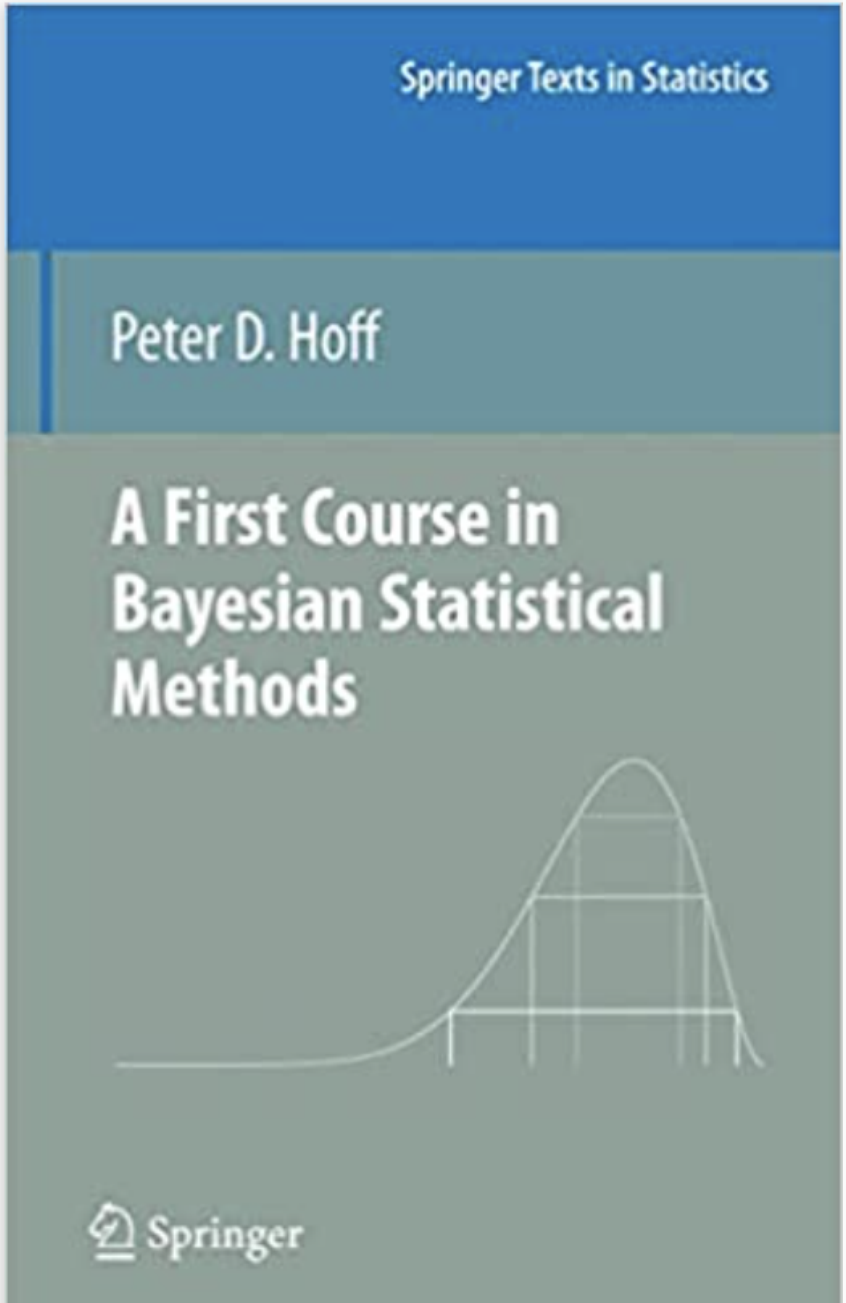
The Multivariate Normal Model

June 1, 2021

Missing data and imputation

References


Springer Texts in Statistics
Peter D. Hoff
A First Course in Bayesian Statistical Methods
Springer

Pima Dataset

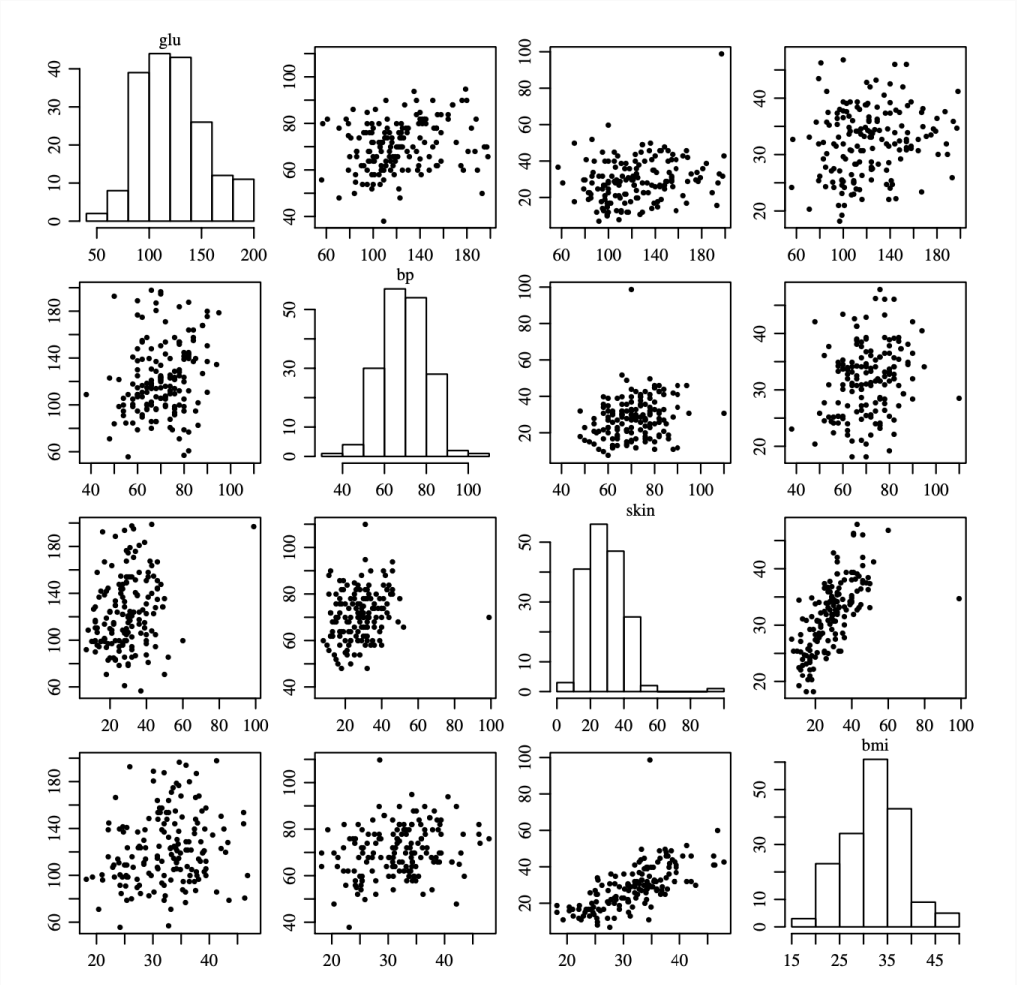


Figure 1: Univariate histograms and bivariate scatterplots for four variables taken from a dataset involving health-related measurements on 200 women of Pima Indian heritage living near Phoenix, Arizona. The four variables are `glu` (blood plasma glucose concentration), `bp` (diastolic blood pressure), `skin` (skin fold thickness), and `bmi` (body mass index).

Pima Dataset

	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

Figure 2: Entries for the first ten subjects in the dataset. The `NA`'s stand for "not available."

Description of problem

How to do parameter estimation in the presence of missing data?

> We cannot do parameter estimation, because we cannot compute the likelihood $\prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$.

Two common approaches taken by software packages:

1. Throw away all subjects with missing data

Description of problem

How to do parameter estimation in the presence of missing data?

> We cannot do parameter estimation, because we cannot compute the likelihood $\prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$.

Two common approaches taken by software packages:

1. Throw away all subjects with missing data
   - ✗ Discards a potentially large amount of useful information.

Description of problem

How to do parameter estimation in the presence of missing data?

> We cannot do parameter estimation, because we cannot compute the likelihood $\prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$.

Two common approaches taken by software packages:

1. Throw away all subjects with missing data
   - ✗ Discards a potentially large amount of useful information.
2. Impute the population mean or some other fixed value.

Description of problem

How to do parameter estimation in the presence of missing data?

We cannot do parameter estimation, because we cannot compute the likelihood $\prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{\theta})$.

Two common approaches taken by software packages:

1. Throw away all subjects with missing data

   ✗ Discards a potentially large amount of useful information.

2. Impute the population mean or some other fixed value.

   ✗ Assumes certainty about these values, when in fact we have not observed them.

Missing at random (MAR)

Let $\boldsymbol{O}_i = (O_1, ..., O_p)^T$ be a binary vector such that

- $O_{ij} = 1 \implies Y_{ij}$ is observed
- $O_{ij} = 0 \implies Y_{ij}$ is missing

Definition

We say the missing data are *missing at random* if $\boldsymbol{O}_i$ and $\boldsymbol{Y}_i$ are conditionally independent given the model parameters $\boldsymbol{\theta}$ and the distribution of $\boldsymbol{O}_i$ does not depend on $\boldsymbol{\theta}$.

Missing at random (MAR)

Let $\boldsymbol{O}_i = (O_1, ..., O_p)^T$ be a binary vector such that

- $O_{ij} = 1 \implies Y_{ij}$ is observed
- $O_{ij} = 0 \implies Y_{ij}$ is missing

Definition

We say the missing data are *missing at random* if $\boldsymbol{O}_i$ and $\boldsymbol{Y}_i$ are conditionally independent given the model parameters $\boldsymbol{\theta}$ and the distribution of $\boldsymbol{O}_i$ does not depend on $\boldsymbol{\theta}$.

Remark. This is one of the three types of missingness. In gist:

- Missing completely at random (MCAR) - missingness is independent of all data
- Missing at random (MAR) - missingness is independent of observed data
- Missing not at random (MNAR) - missingness depends on missing values (and perhaps observed data)

The likelihood in the presence of MAR data

When the data is missing at random, the sampling probability (density) for the data from observational unit i is given by

$$p(\boldsymbol{o}_i, \{y_{ij} : o_{ij} = 1\} \mid \boldsymbol{\theta}) \overset{(1)}{=} p(\boldsymbol{o}_i)\, p(\{y_{ij} : o_{ij} = 1\} \mid \boldsymbol{\theta})$$
$$= p(\boldsymbol{o}_i) \int p(y_{i1}, ..., y_{ip} \mid \boldsymbol{\theta}) \prod_{y_{ij}:o_{ij}=0} dy_{ij}$$

where in (1) we applied the definition of MAR.

The likelihood in the presence of MAR data

When the data is missing at random, the sampling probability (density) for the data from observational unit i is given by

$$p(\boldsymbol{o}_i, \{y_{ij} : o_{ij} = 1\} \mid \boldsymbol{\theta}) \overset{(1)}{=} p(\boldsymbol{o}_i)\, p(\{y_{ij} : o_{ij} = 1\} \mid \boldsymbol{\theta})$$
$$= p(\boldsymbol{o}_i) \int p(y_{i1}, ..., y_{ip} \mid \boldsymbol{\theta}) \prod_{y_{ij}:o_{ij}=0} dy_{ij}$$

where in (1) we applied the definition of MAR.

✓ So in the presence of MAR data, the correct thing to do is *integrate over* the missing data to obtain the marginal probability (density) of the observed data.

Utilization in multivariate normal models

In the case of multivariate normal models (so $\theta = (\mu, \Sigma)$) , the integration is easy: Multivariate normals have normal marginals.

Example

Suppose $\boldsymbol{y}_i = (y_{i1}, \texttt{NA}, y_{i3}, \texttt{NA})^T$, so $\boldsymbol{o}_i = (1, 0, 1, 0)^T$.

Then

$$\begin{aligned} p(\boldsymbol{o}_i, y_{i1}, y_{i3} \mid \boldsymbol{\mu}, \Sigma) &= p(\boldsymbol{o}_i)\, p(y_{i1}, y_{i3} \mid \boldsymbol{\mu}, \Sigma) \\ &= p(\boldsymbol{o}_i) \int p(\boldsymbol{y}_i \mid \boldsymbol{\mu}, \Sigma)\, dy_2\, dy_4 \end{aligned}$$

The marginal density $p(y_{i1}, y_{i3} \mid \boldsymbol{\theta})$ is simply a bivariate normal density with mean $(\mu_1, \mu_3)^T$ and covariance matrix made up of $(\sigma_1^2, \sigma_{13}, \sigma_3^2)$.

TODO

- Inference: Show how to adjust Gibbs sampling in the presence of missing data (see Hoff pp. 117-pp.118; also make sure the notation, and the use of semiconjugate vs conjugate prior, aligns with how I set up the multivariate normal initially – earlier on in this section of the course)
- Correlations - discuss how to construct the posterior correlation matrix from the Gibbs samples (and note this is another example of the Bayesian paradigm yielding unlimited access to posterior functionals of interest, without doing any extra inferential work). Show the specific values on pp.119, and the left hand side of Fig 7.4. This is good to show because it is something you'd probably want out of a normal model anyways, and also becuase it is needed for the next point.
- Show true values vs posterior expectations of the missing data (Hoff Fig 7.5) Mention that we get better predictions for skin and bmi, due to their higher correlations. Highlight how much better the posterior expectation is than a flat fixed value.