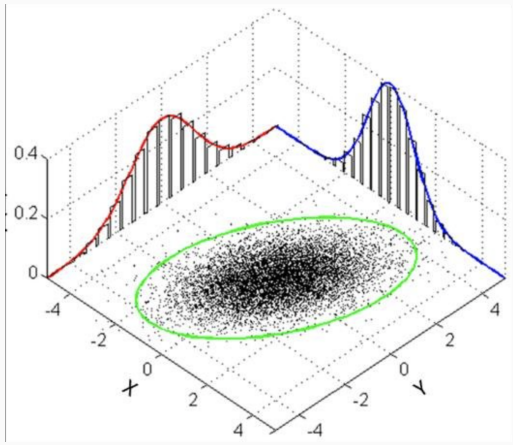
The Multivariate Normal Model

June 4, 2021

Table of contents

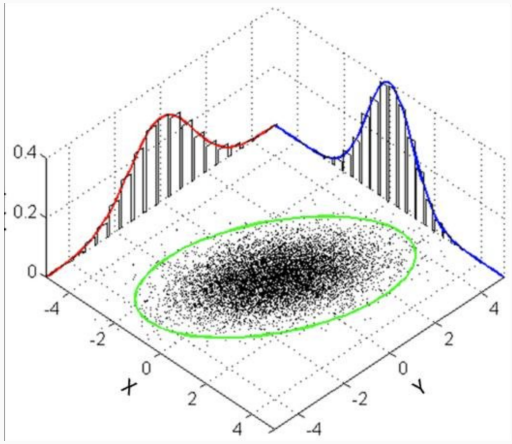
Overview

Some motivations for the normal



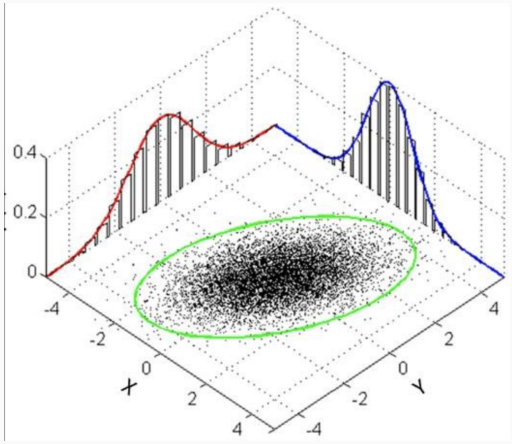
- *Maximum entropy* among all distributions with a given mean μ and variance Σ.

Some motivations for the normal



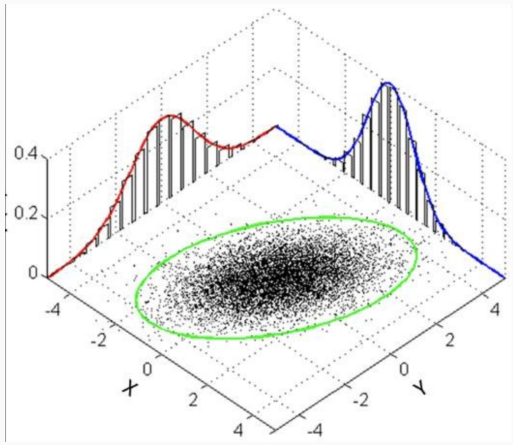
- *Maximum entropy* among all distributions with a given mean μ and variance Σ.
- Characterized by independence of sample mean and sample variance.
  (Bayesian take: ask if your beliefs about the sample mean are independent from those about the sample variance.)

Some motivations for the normal



- *Maximum entropy* among all distributions with a given mean μ and variance Σ.
- Characterized by independence of sample mean and sample variance.
  (Bayesian take: ask if your beliefs about the sample mean are independent from those about the sample variance.)
- Sample averages are generally approximately normally distributed due to the Central Limit Theorem.

Some motivations for the normal



- *Maximum entropy* among all distributions with a given mean μ and variance Σ.
- Characterized by independence of sample mean and sample variance.
  (Bayesian take: ask if your beliefs about the sample mean are independent from those about the sample variance.)
- Sample averages are generally approximately normally distributed due to the Central Limit Theorem.
- Sufficient statistics are sample mean and variance; so will consistently estimate population mean and variance even for non-normal distributions.

Why Bayesian normal?

- Prior information often exists and can be taken into account.
  - Population-level info: see the typing example, PIMA Indians example
  - Nature (e.g. support) of data: see the reading comprehension example

Why Bayesian normal?

- Prior information often exists and can be taken into account.
  - Population-level info: see the typing example, PIMA Indians example
  - Nature (e.g. support) of data: see the reading comprehension example
- ML estimates of covariance matrices have large variance.
  - Problem can be especially bad in certain contexts (e.g., small data, high-dimensions, missing data)
  - Spherical prior provides regularization
  - Posterior asymptotically concentrates around maximum likelihood (ML) solution

Why Bayesian normal?

- Prior information often exists and can be taken into account.
  - Population-level info: see the typing example, PIMA Indians example
  - Nature (e.g. support) of data: see the reading comprehension example
- ML estimates of covariance matrices have large variance.
  - Problem can be especially bad in certain contexts (e.g., small data, high-dimensions, missing data)
  - Spherical prior provides regularization
  - Posterior asymptotically concentrates around maximum likelihood (ML) solution
- Inference is no harder than for frequentist models
  - Easy, cheap updates (a conjugate prior exists)
  - Supports online learning
  - Fits nicely in more complex models
  - Nice hyperparameter interpretation

Conjugate inference

A conjugate prior

TODO: Fill in

Application: Modeling typing dynamics

See powerpoint slides.

Semi-conjugate inference

Semi-conjugate Bayesian MVN

Consider the following model with a normal sampling distribution and *semi-conjugate* prior

$$\begin{aligned}
\boldsymbol{\mu} &\sim \mathcal{N}_d(\boldsymbol{m}_0, \boldsymbol{V}_0) \\
\Sigma &\sim \mathcal{W}^{-1}(\nu_0, \Psi_0) \\
\boldsymbol{x}_i \mid \boldsymbol{\mu}, \Sigma &\overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \Sigma), \quad i = 1, ..., N
\end{aligned}$$

We define $\boldsymbol{x} := (\boldsymbol{x}_1, \ldots \boldsymbol{x}_N)$, where each $\boldsymbol{x}_i \in \mathbb{R}^d$.

Semi-conjugate Bayesian MVN

Consider the following model with a normal sampling distribution and *semi-conjugate* prior

$$\begin{aligned}
\boldsymbol{\mu} &\sim \mathcal{N}_d(\boldsymbol{m}_0, \boldsymbol{V}_0) \\
\Sigma &\sim \mathcal{W}^{-1}(\nu_0, \Psi_0) \\
\boldsymbol{x}_i \mid \boldsymbol{\mu}, \Sigma &\overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \Sigma), \quad i = 1, ..., N
\end{aligned}$$

We define $\boldsymbol{x} := (\boldsymbol{x}_1, \ldots \boldsymbol{x}_N)$, where each $\boldsymbol{x}_i \in \mathbb{R}^d$.

Fully conjugate vs semi-conjugate MVNs

This model is different than the model with the fully conjugate (Normal-Inverse-Wishart) prior on the pair $(\boldsymbol{\mu}, \Sigma)$. The conditionally conjugate prior lacks closed-form posterior updating, but is also more expressive. It is also easier to extend upwards.

Semi-conjugate models generally

Conjugate models

Conjugacy can be defined as follows (gelman2013bayesian). If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\theta \mid y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

Semi-conjugate models

Conditional conjugacy (sometimes called semi-conjugacy) can be defined similarly (gelman2013bayesian). If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for $\theta \mid \phi$, then the class $\mathcal{P}$ is *conditionally conjugate* for $\mathcal{F}$ if

$$p(\theta \mid \phi, y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta, \phi) \in \mathcal{F} \text{ and } p(\cdot \mid \phi) \in \mathcal{P}$$

Semi-conjugate models generally

Conjugate models

Conjugacy can be defined as follows (*gelman2013bayesian*). If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for θ, then the class $\mathcal{P}$ is *conjugate* for $\mathcal{F}$ if

$$p(\theta \mid y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

Semi-conjugate models

Conditional conjugacy (sometimes called semi-conjugacy) can be defined similarly (*gelman2013bayesian*). If $\mathcal{F}$ is a class of sampling distributions and $\mathcal{P}$ is a class of prior distributions for $\theta \mid \phi$, then the class $\mathcal{P}$ is *conditionally conjugate* for $\mathcal{F}$ if

$$p(\theta \mid \phi, y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta, \phi) \in \mathcal{F} \text{ and } p(\cdot \mid \phi) \in \mathcal{P}$$

In other words, a family of prior distributions for a parameter is called conditionally conjugate if the conditional posterior distribution (often called the *complete conditional*), given the data and all other parameters in the model, is also in that class.

We ♡ semi-conjugacy

Why are conditionally conjugate models of interest? The posterior distributions for conditionally conjugate models are easily approximated with Gibbs sampling or Mean Field Variational Inference – the former samples from the complete conditional, whereas the latter takes variational expectations with respect to the natural parameter of the complete conditional.

Complete conditionals for the Bayesian MVN

We sample from the posterior by iteratively sampling from the *complete conditionals*:

$$\boldsymbol{\mu} \mid \Sigma, \boldsymbol{x} \sim \mathcal{N}_d(\boldsymbol{m}, \boldsymbol{V})$$

where

$$\boldsymbol{m} = \left(\boldsymbol{V}_0^{-1} + N\Sigma^{-1}\right)^{-1}\left(\boldsymbol{V}_0^{-1}\boldsymbol{m}_0 + N\Sigma^{-1}\bar{\boldsymbol{x}}\right)$$

$$\boldsymbol{V} = \left(\boldsymbol{V}_0^{-1} + N\Sigma^{-1}\right)^{-1}$$

and

$$\Sigma \mid \boldsymbol{\mu}, \boldsymbol{x} \sim \mathcal{W}^{-1}(\nu, \Psi)$$

where

$$\nu = \nu_0 + N$$

$$\Psi = \Psi_0 + \sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

Complete conditionals: Interpretation

These complete conditionals have nice interpretations:

- **Complete conditional for** $(\mu \mid \Sigma, \boldsymbol{x})$: On the precision scale, $\boldsymbol{V}$ is the sum of the prior precision matrix $\boldsymbol{V}_0^{-1}$ and N copies of the precision for each observation, Σ^{-1}. Similarly, $\boldsymbol{m}$ is the precision-weighted convex combination of $\boldsymbol{m}_0$, the prior mean, and the empirical average, $\bar{\boldsymbol{x}}$.
- **Complete conditional for** $(\Sigma \mid \mu, \boldsymbol{x})$: The covariance was estimated from ν observations with a sum of pairwise deviation products Ψ.

Complete conditionals: Proof

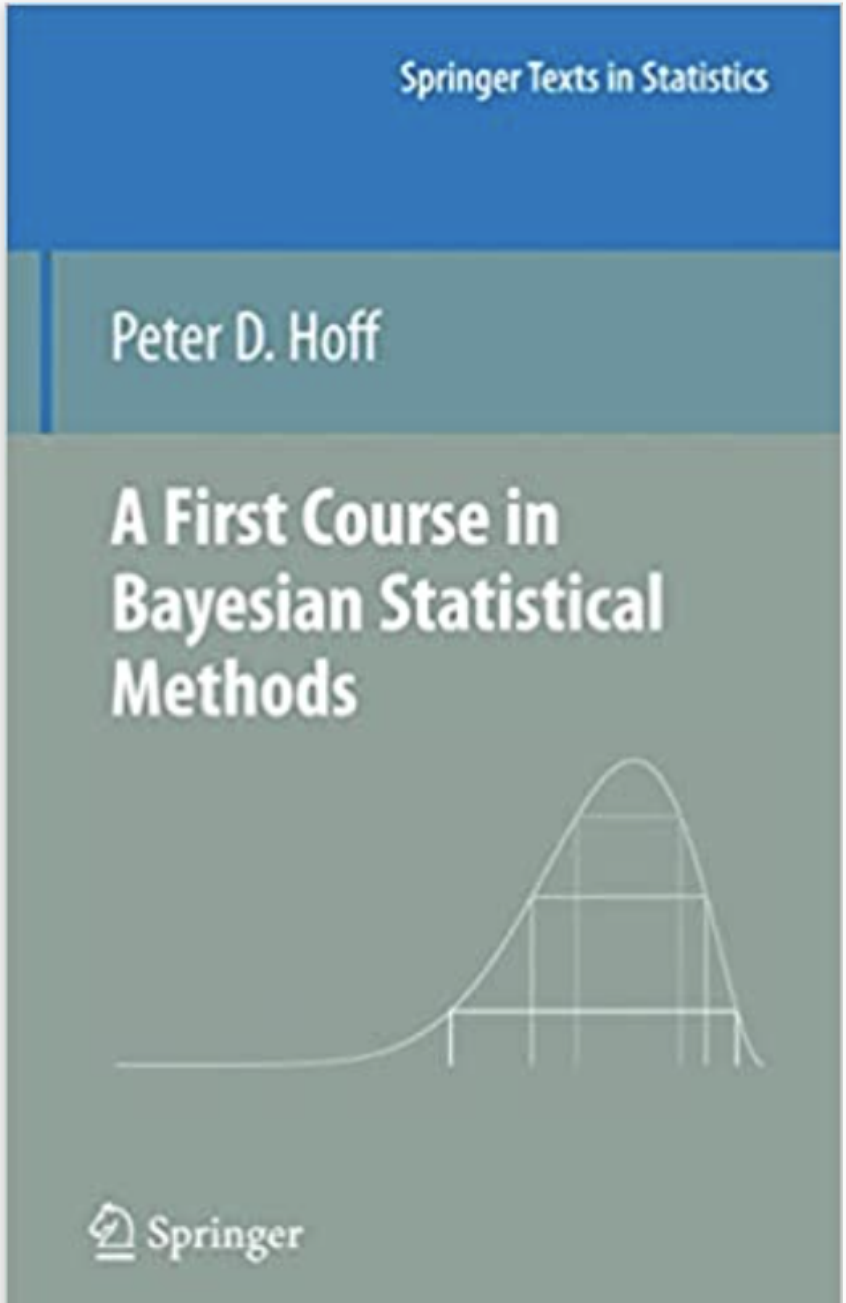
See exponential family notes.

Application: Reading Comprehension

See ipython notebook.

Missing data and imputation

References


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

Pima Dataset

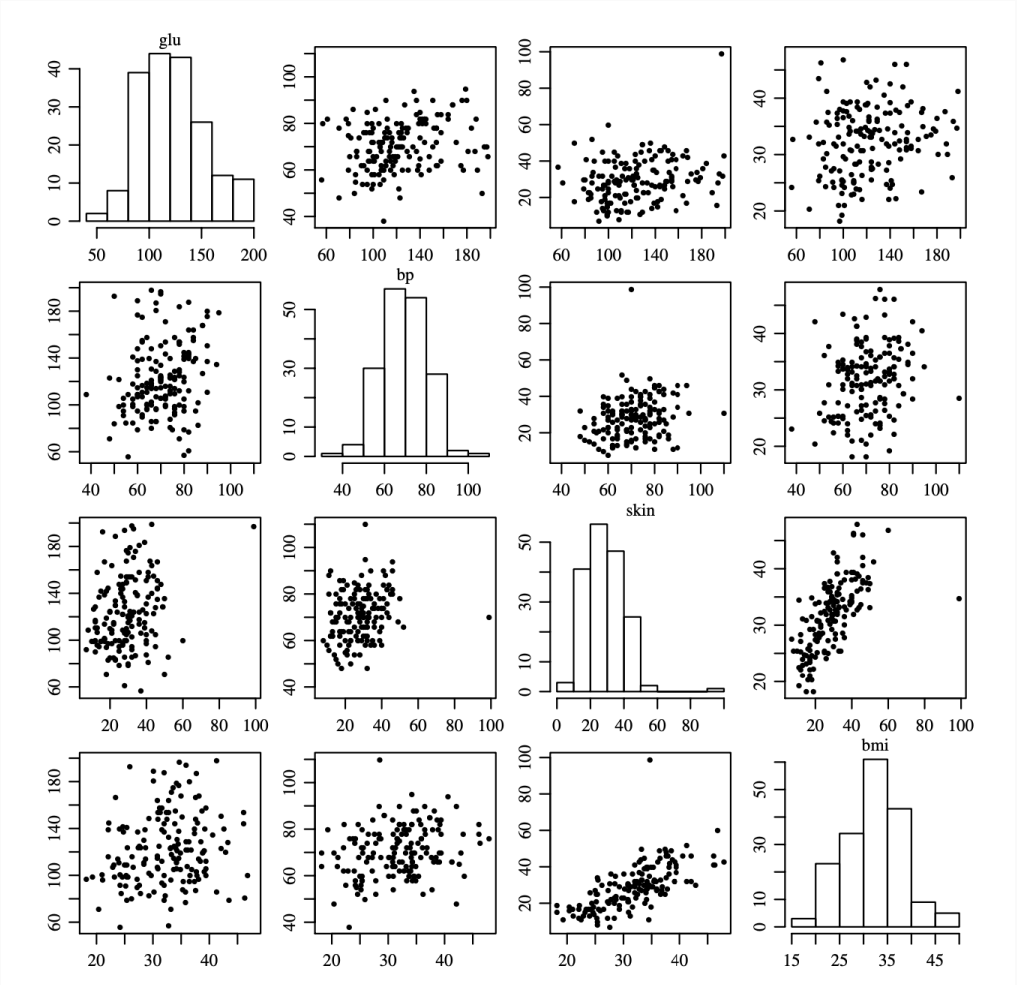


Figure 1: Univariate histograms and bivariate scatterplots for four variables taken from a dataset involving health-related measurements on 200 women of Pima Indian heritage living near Phoenix, Arizona. The four variables are `glu` (blood plasma glucose concentration), `bp` (diastolic blood pressure),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