The Multivariate Normal Model

June 9, 2021

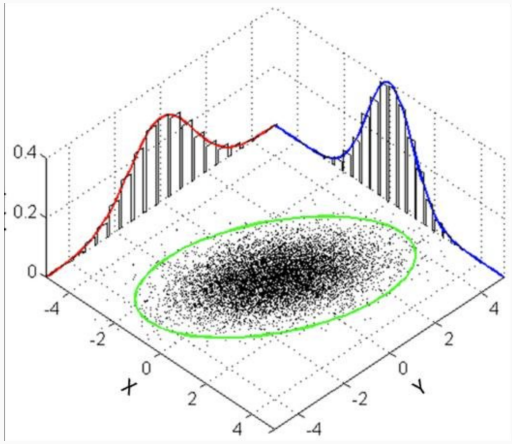
Table of contents

Overview

Overview

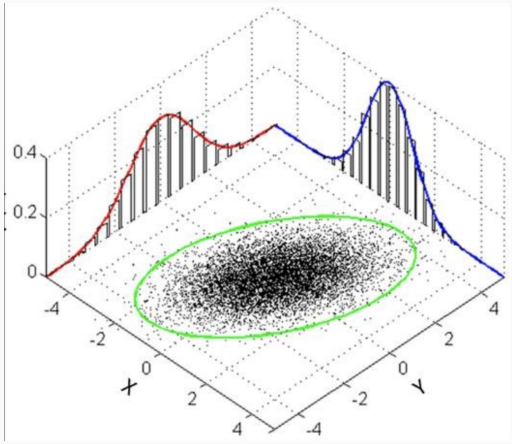
Motivations

Some motivations for the normal



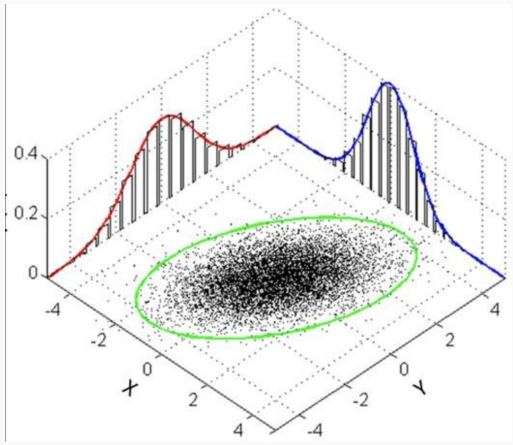
- *Maximum entropy* among all distributions with a given mean μ and variance Σ.

Some motivations for the normal



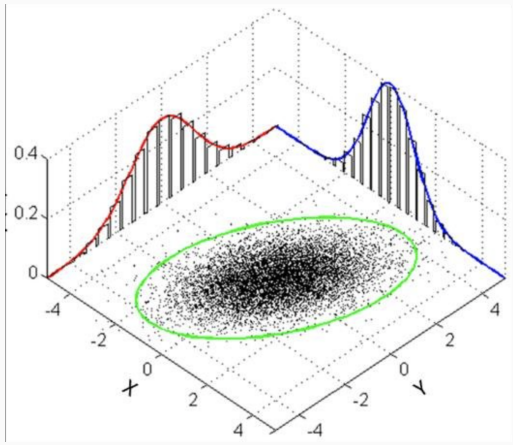
- *Maximum entropy* among all distributions with a given mean μ and variance Σ.
- Characterized by independence of sample mean and sample variance.
  (Bayesian take: ask if your beliefs about the sample mean are independent from those about the sample variance.)

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 (Previous example: disease prevalence. Forthcoming example: biometrics and PIMA Indians)
  - Nature (e.g. support) of data (Forthcoming example: reading comprehension)

Why Bayesian normal?

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

Why Bayesian normal?

- Prior information often exists and can be taken into account.
  - Population-level info (Previous example: disease prevalence. Forthcoming example: biometrics and PIMA Indians)
  - Nature (e.g. support) of data (Forthcoming example: reading comprehension)
- ML estimates of covariance matrices can have large variance.
  - Problem can be especially bad in certain contexts (e.g., small data, high-dimensions, missing data)
  - Spherical prior provides regularization
  - Posterior still asymptotically concentrates around maximum likelihood (ML) solution
- Inference can be as easy or easier than in frequentist models
  - Easy, cheap updates (esp. when using a conjugate prior)
  - Supports online learning
  - Fits nicely in more complex models

Overview

Building blocks

Review: Exponential families

We have observed that the Bernoulli and the Poisson distributions are both exponential families.

Exponential family

An *exponential family* is a set of probability distributions whose probability density functions have the following form

$$p(x \mid \theta) = h(x) \exp\{\eta(\theta)^T t(x) - a(\theta)\} \tag{1}$$

where we refer to h as the base measure, η as the natural parameter, t as the sufficient statistics, and a as the log normalizer.

Multivariate normal is an exponential family

We can write the density of a multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ distribution in exponential form. *(Try it!)*

Multivariate normal is an exponential family

We can write the density of a multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ distribution in exponential form. *(Try it!)*

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

$$= (2\pi)^{-d/2} \exp\left\{ \underbrace{-\frac{1}{2}}_{-\frac{1}{2}\mathrm{vec}(\Sigma^{-1})^T} \underbrace{\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}}_{\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)} + (\Sigma^{-1}\boldsymbol{\mu})^T \boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \tfrac{1}{2}\log|\Sigma^{-1}| \right\} \tag{2}$$

Note: the underbrace representation is given by $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}) = \mathrm{tr}(\Sigma^{-1}\boldsymbol{x}\boldsymbol{x}^T) = \mathrm{vec}(\Sigma^{-1})^T\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)$.

Multivariate normal is an exponential family

We can write the density of a multivariate normal $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ distribution in exponential form. *(Try it!)*

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

$$= (2\pi)^{-d/2} \exp\left\{ \underbrace{-\frac{1}{2}}_{-\frac{1}{2}\mathrm{vec}(\Sigma^{-1})} \underbrace{\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}}_{\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)} + (\Sigma^{-1}\boldsymbol{\mu})^T \boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu} + \tfrac{1}{2}\log|\Sigma^{-1}| \right\} \tag{2}$$

Note: the underbrace representation is given by $\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^T\Sigma^{-1}\boldsymbol{x}) = \mathrm{tr}(\Sigma^{-1}\boldsymbol{x}\boldsymbol{x}^T) = \mathrm{vec}(\Sigma^{-1})^T\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T)$.

- **Natural parameter**: $\eta(\boldsymbol{\mu}, \Sigma) = \left(-\frac{1}{2}\mathrm{vec}(\Sigma^{-1}),\ \Sigma^{-1}\boldsymbol{\mu} \right)$,
- **Sufficient statistics**: $t(\boldsymbol{x}) = \left(\mathrm{vec}(\boldsymbol{x}\boldsymbol{x}^T), \boldsymbol{x}\right)$,
- **Log normalizer**: $a(\boldsymbol{\mu}, \Sigma) = -\frac{1}{2}\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\log|\Sigma^{-1}|$.
- **Base measure**: $h(x) = (2\pi)^{-d/2}$

Qualitative points

The natural parametrization of the MVN

From Equation (2), we see that that the natural parameters of the MVN are

- the *precision* Σ^{-1} , and
- the *precision-weighted mean* $\Sigma^{-1}\boldsymbol{\mu}$.

Are you exponentiating a quadratic?

We also see that if a random vector $\boldsymbol{x}$ has a density on $\mathbb{R}^d$ that satisfies

$$p(\boldsymbol{x}) \propto \exp\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{x}^T\boldsymbol{b}\}$$

for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$, then $\boldsymbol{x}$ must be multivariate normal. Moreover its natural parameters are precision $\boldsymbol{A}$ and precison-weighted mean $\boldsymbol{b}$. In other words, the covariance is $\boldsymbol{A}^{-1}$ and the mean is $\boldsymbol{A}^{-1}\boldsymbol{b}$.

Self-help guide for deriving Gibbs samplers

Inverse Wishart Distribution

The Inverse Wishart is a distribution on symmetric, positive definite matrices. The Inverse Wishart distribution, denoted $\mathcal{W}^{-1}(\nu, \Psi)$, has density

$$p(\Sigma) \propto |\Sigma|^{-(\nu+d+1)/2} \exp\left[-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}\Psi)\right] \tag{3}$$

where $\Sigma \succ 0$ and $\nu > d-1$ to have a proper prior. The expected value of an Inverse Wishart random variable parametrized as in (3) is given by $\mathbb{E}[\Sigma] = \frac{\Psi}{\nu-d-1}$.

Interpreting the parameters of the Inverse Wishart

Note that the parameters of the Inverse Wishart can be interpreted (as per conjugacy; we will see this below) in the following way: the covariance was estimated from ν observations with a residual sum of squares (a.k.a. sum of pairwise deviation products) Ψ.

Inverse Wishart Distribution

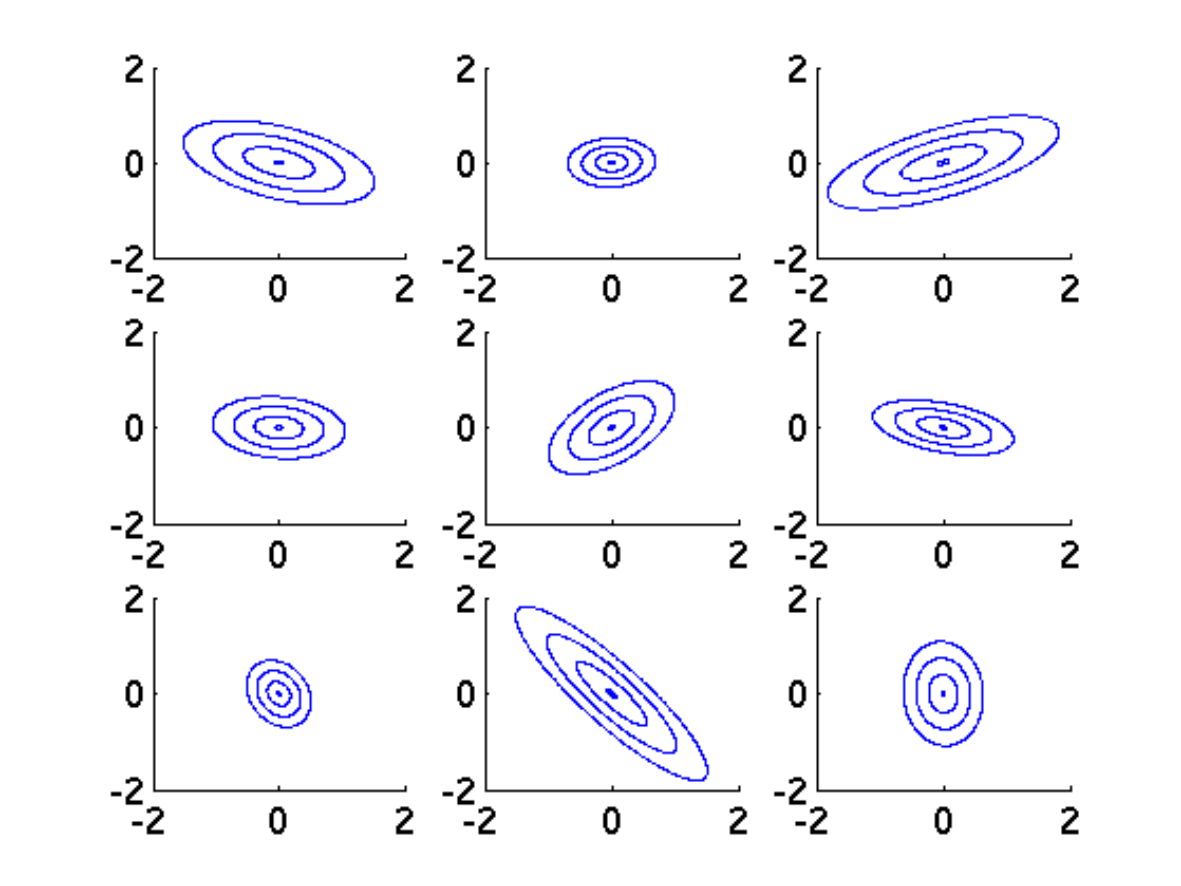


Image Credit: Michael Hughes, Tufts University

Sampling from the Inverse Wishart

A sample Σ from the $\mathcal{W}^{-1}(\nu, \Psi)$ distribution can be obtained by the following scheme:

1. Sample $\boldsymbol{z}_1, ..., \boldsymbol{z}_\nu \overset{\text{iid}}{\sim} \mathcal{N}(0, \Psi^{-1})$
2. Calculate $\boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^{\nu} \boldsymbol{z}_i \boldsymbol{z}_i^T$.
3. Set $\Sigma = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$.

The intuition is that the Inverse Wishart models covariance matrices as an inverse sum of squares.

Conjugate inference

A fully conjugate formulation

Since the MVN is an EF, we can form a fully conjugate model.

Fully conjugate Bayesian MVN (has closed-form posteriors and predictive posteriors!)

Given observations $\boldsymbol{x} := (\boldsymbol{x}_1, \ldots \boldsymbol{x}_N)$, where each $\boldsymbol{x}_i \in \mathbb{R}^d$, we take

$$\boldsymbol{\mu}, \Sigma \sim \text{NIW}_d(\alpha_0, \boldsymbol{\mu}_0, \nu_0, \Psi_0)$$

$$\boldsymbol{x}_i \mid \boldsymbol{\mu}, \Sigma \overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \Sigma), \quad i = 1, ..., N$$

A fully conjugate formulation

Since the MVN is an EF, we can form a fully conjugate model.

Fully conjugate Bayesian MVN (has closed-form posteriors and predictive posteriors!)

Given observations $\boldsymbol{x} := (\boldsymbol{x}_1, \ldots \boldsymbol{x}_N)$, where each $\boldsymbol{x}_i \in \mathbb{R}^d$, we take

$$\boldsymbol{\mu}, \Sigma \sim \text{NIW}_d(\alpha_0, \boldsymbol{\mu}_0, \nu_0, \Psi_0)$$
$$\boldsymbol{x}_i \mid \boldsymbol{\mu}, \Sigma \overset{\text{iid}}{\sim} \mathcal{N}_d(\boldsymbol{\mu}, \Sigma), \quad i = 1, ..., N$$

The Normal-Inverse-Wishart prior

$$\boldsymbol{\mu}, \Sigma \sim \text{NIW}_d(\alpha_0, \boldsymbol{\mu}_0, \nu_0, \Psi_0)$$

which means

$$\Sigma \sim \mathcal{W}_d^{-1}(\nu_0, \Psi_0)$$
$$\boldsymbol{\mu} \mid \Sigma \sim \mathcal{N}_d(\boldsymbol{\mu}_0, \frac{1}{\alpha_0}\Sigma)$$

Application: Modeling typing dynamics

See powerpoint slides.

Semi-conjugate inference

Semi-conjugate Bayesian normal

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 models: Gist

Main idea: A family of prior distributions for a parameter is called *semi-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 family.

Comparison

Fully conjugate model

- $(\boldsymbol{\mu}, \Sigma) \mid$ rest is Normal-Inverse-Wishart

Semi-conjugate model

- $\boldsymbol{\mu} \mid$ rest is Normal
- $\Sigma \mid$ rest is Inverse Wishart

Semi-conjugate models: Gist

Main idea: A family of prior distributions for a parameter is called *semi-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 family.

Comparison

Fully conjugate model

- $(\boldsymbol{\mu}, \Sigma) \mid$ rest is Normal-Inverse-Wishart

Semi-conjugate model

- $\boldsymbol{\mu} \mid$ rest is Normal
- $\Sigma \mid$ rest is Inverse Wishart

Relative evaluation of semi-conjugate model

Semi-conjugate models: Gist

Main idea: A family of prior distributions for a parameter is called *semi-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 family.

Comparison

Fully conjugate model

- $(\boldsymbol{\mu}, \Sigma) \mid$ rest is Normal-Inverse-Wishart

Semi-conjugate model

- $\boldsymbol{\mu} \mid$ rest is Normal
- $\Sigma \mid$ rest is Inverse Wishart

Relative evaluation of semi-conjugate model

✗ lacks closed-form posterior updating

Semi-conjugate models: Gist

Main idea: A family of prior distributions for a parameter is called *semi-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 family.

Comparison

Fully conjugate model

- $(\boldsymbol{\mu}, \Sigma) \mid$ rest is Normal-Inverse-Wishart

Semi-conjugate model

- $\boldsymbol{\mu} \mid$ rest is Normal
- $\Sigma \mid$ rest is Inverse Wishart

Relative evaluation of semi-conjugate model

✗ lacks closed-form posterior updating

✓ more expressive

Semi-conjugate models: Gist

Main idea: A family of prior distributions for a parameter is called *semi-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 family.

Comparison

Fully conjugate model

- $(\boldsymbol{\mu}, \Sigma) \mid$ rest is Normal-Inverse-Wishart

Semi-conjugate model

- $\boldsymbol{\mu} \mid$ rest is Normal
- $\Sigma \mid$ rest is Inverse Wishart

Relative evaluation of semi-conjugate model

✗ lacks closed-form posterior updating

✓ more expressive

✓ easier to embed in more complex models

Semi-conjugate models: Definition

Review: Conjugate models

Conjugacy was 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

Semi-conjugate (sometimes called conditionally-conjugate) models 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

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

We ♡ semi-conjugacy

Why are semi-conjugate models of interest?

The posterior distributions for conditionally conjugate models are easily approximated

- with **Gibbs sampling**, by sampling from the complete conditionals
- with **mean-field variational inference**, by taking variational expectations of the natural parameters of the complete conditional.

The complete conditional for μ

We use the exponential family representation of the MVN to represent the prior in terms of its natural parameters

$$p(\boldsymbol{\mu}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T \Big(\underbrace{\boldsymbol{V}_0^{-1}}_{\text{prior precision}} \Big)\boldsymbol{\mu} + \boldsymbol{\mu}^T \Big(\underbrace{\boldsymbol{V}_0^{-1}\boldsymbol{m}_0}_{\text{prior precision-weighted mean}} \Big) \right\} \tag{4}$$

And similarly, we write the likelihood $L(\boldsymbol{\mu}) = p(\boldsymbol{y} \mid \boldsymbol{\mu}) = \prod_{i=1}^n p(\boldsymbol{y}_i \mid \boldsymbol{\mu})$ as

$$\begin{aligned} L(\boldsymbol{\mu}) &\propto \exp\left\{ -\frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right\} \\ &= \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T \Big(\underbrace{n\Sigma^{-1}}_{\text{data precision}} \Big)\boldsymbol{\mu} + \boldsymbol{\mu}^T \Big(\underbrace{\Sigma^{-1} n\bar{\boldsymbol{x}}}_{\text{data precision-weighted mean}} \Big) \right\} \end{aligned} \tag{5}$$

Can you finish the derivation? (Hint: Remember Jimi!)

The complete conditional for μ

We use the exponential family representation of the MVN to represent the prior in terms of its natural parameters

$$p(\boldsymbol{\mu}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T \Big(\underbrace{\boldsymbol{V}_0^{-1}}_{\text{prior precision}} \Big)\boldsymbol{\mu} + \boldsymbol{\mu}^T \Big(\underbrace{\boldsymbol{V}_0^{-1}\boldsymbol{m}_0}_{\text{prior precision-weighted mean}} \Big) \right\} \tag{4}$$

And similarly, we write the likelihood $L(\boldsymbol{\mu}) = p(\boldsymbol{y} \mid \boldsymbol{\mu}) = \prod_{i=1}^n p(\boldsymbol{y}_i \mid \boldsymbol{\mu})$ as

$$\begin{aligned} L(\boldsymbol{\mu}) &\propto \exp\left\{ -\frac{1}{2}\sum_{i=1}^n (\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right\} \\ &= \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T \Big(\underbrace{n\Sigma^{-1}}_{\text{data precision}} \Big)\boldsymbol{\mu} + \boldsymbol{\mu}^T \Big(\underbrace{\Sigma^{-1} n\bar{\boldsymbol{x}}}_{\text{data precision-weighted mean}} \Big) \right\} \end{aligned} \tag{5}$$

Can you finish the derivation? (Hint: Remember Jimi!) By Bayes' law, and combining like terms,

$$\begin{aligned} p(\boldsymbol{\mu} \mid \boldsymbol{x}, \Sigma) &\propto \underbrace{p(\boldsymbol{\mu})}_{\text{prior}} \underbrace{p(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma)}_{\text{likelihood}} \\ &= \exp\left\{ -\frac{1}{2}\boldsymbol{\mu}^T \Big(\underbrace{\boldsymbol{V}_0^{-1} + n\Sigma^{-1}}_{\text{posterior precision}} \Big)\boldsymbol{\mu} + \boldsymbol{\mu}^T \Big(\underbrace{\boldsymbol{V}_0\boldsymbol{m}_0 + \Sigma^{-1} n\bar{\boldsymbol{x}}}_{\text{posterior precision-weighted mean}} \Big) \right\} \end{aligned}$$

which reveals that the posterior is normal, along with the particular form of its parameter (covariance $\boldsymbol{A}^{-1}$ and mean $\boldsymbol{A}^{-1}\boldsymbol{b}$).

Complete conditionals for the Bayesian MVN

The complete conditions under natural parametrization

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

where

$$\underbrace{\boldsymbol{V}^{-1}}_{\text{posterior precision}} = \underbrace{\boldsymbol{V}_0^{-1}}_{\text{prior precision}} + \underbrace{N\Sigma^{-1}}_{\text{data precision}}$$

$$\underbrace{\boldsymbol{V}^{-1}\boldsymbol{m}}_{\text{posterior precision-weighted mean}} = \underbrace{\boldsymbol{V}_0^{-1}\boldsymbol{m}_0}_{\text{prior precision-weighted mean}} + \underbrace{N\Sigma^{-1}\bar{\boldsymbol{x}}}_{\text{data precision-weighted mean}}$$

and

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

where

$$\underbrace{\nu}_{\text{posterior sample size}} = \underbrace{\nu_0}_{\text{prior sample size}} + \underbrace{N}_{\text{sample size}}$$

$$\underbrace{\Psi}_{\text{posterior RSS}} = \underbrace{\Psi_0}_{\text{prior RSS}} + \underbrace{\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T}_{\text{data RSS}}$$

Note: RSS = Residual Sum of Squares

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 the data precision matrix (which is 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.

Gibbs Sampler (in standard parametrization)

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

$$\boldsymbol{V} = \Bigg(\underbrace{\boldsymbol{V}_0^{-1}}_{\text{prior precision}} + \underbrace{N\Sigma^{-1}}_{\text{data precision}} \Bigg)^{-1}$$

$$\boldsymbol{m} = \boldsymbol{V} \Bigg(\underbrace{\boldsymbol{V}_0^{-1}\boldsymbol{m}_0}_{\text{prior precision-weighted mean}} + \underbrace{N\Sigma^{-1}\bar{\boldsymbol{x}}}_{\text{data precision-weighted mean}} \Bigg)$$

and

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

where

$$\nu = \nu_0 + N$$

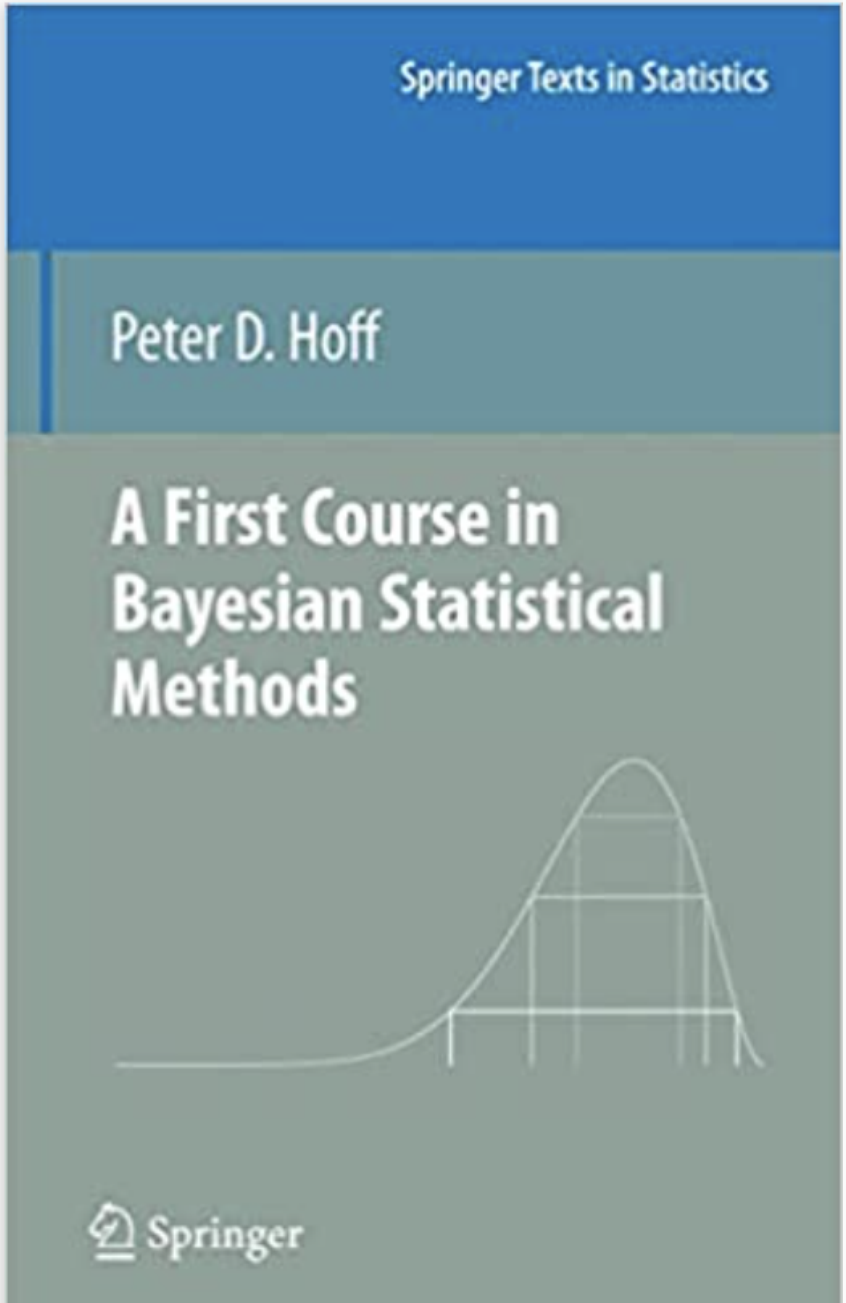
$$\Psi = \Psi_0 + \sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T$$

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

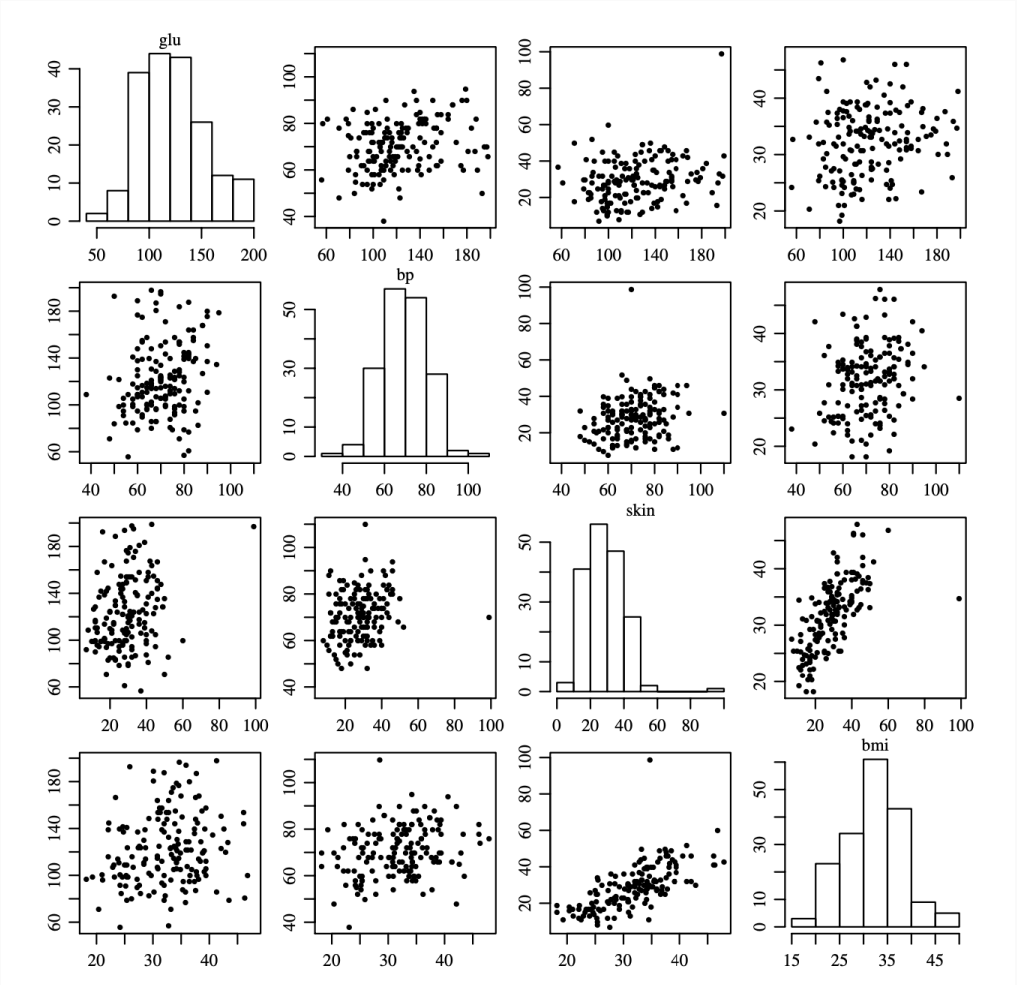


Figure 1: Univariate histograms and bivariate scatterplots for four variables taken from a dataset involving health-related measurements on 200 women of Pima Indian heritage living near Phoenix, Arizona. The four variables are `glu` (blood plasma glucose concentration), `bp` (diastolic blood pressure), `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 $(\sigma_1^2, \sigma_{13}, \sigma_3^2)$.

Gibbs sampling with missing data

Complete data

If $\boldsymbol{Y}$ is the $n \times p$ matrix in which $o_{i,j} = 1$ if $Y_{i,j}$ is observed and $o_{i,j} = 0$ if $Y_{i,j}$ is missing, then $\boldsymbol{Y}$ has two parts

- $\boldsymbol{Y}_{\text{obs}} := \{y_{i,j} : o_{i,j} = 1\}$, the data that we observe, and
- $\boldsymbol{Y}_{\text{miss}} := \{y_{i,j} : o_{i,j} = 0\}$, the data that we do not observe.

Gibbs sampler

A Gibbs sampling scheme for approximating the posterior is given by:

1. Sampling $\boldsymbol{\mu}^{(s+1)}$ from $p(\boldsymbol{\mu} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{miss}}^{(s)}, \Sigma^{(s)})$;
2. Sampling $\Sigma^{(s+1)}$ from $p(\Sigma \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{miss}}^{(s)}, \boldsymbol{\mu}^{(s+1)})$;
3. Sampling $\boldsymbol{Y}_{\text{miss}}^{(s+1)}$ from $p(\boldsymbol{Y}_{\text{miss}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\mu}^{(s+1)}, \Sigma^{(s+1)})$;

Gibbs sampling with missing data

Complete data

If $\boldsymbol{Y}$ is the $n \times p$ matrix in which $o_{i,j} = 1$ if $Y_{i,j}$ is observed and $o_{i,j} = 0$ if $Y_{i,j}$ is missing, then $\boldsymbol{Y}$ has two parts

- $\boldsymbol{Y}_{\text{obs}} := \{y_{i,j} : o_{i,j} = 1\}$, the data that we observe, and
- $\boldsymbol{Y}_{\text{miss}} := \{y_{i,j} : o_{i,j} = 0\}$, the data that we do not observe.

Gibbs sampler

A Gibbs sampling scheme for approximating the posterior is given by:

1. Sampling $\boldsymbol{\mu}^{(s+1)}$ from $p(\boldsymbol{\mu} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{miss}}^{(s)}, \Sigma^{(s)})$;
2. Sampling $\Sigma^{(s+1)}$ from $p(\Sigma \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{Y}_{\text{miss}}^{(s)}, \boldsymbol{\mu}^{(s+1)})$;
3. Sampling $\boldsymbol{Y}_{\text{miss}}^{(s+1)}$ from $p(\boldsymbol{Y}_{\text{miss}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\mu}^{(s+1)}, \Sigma^{(s+1)})$;

The first two steps are the same as before! The third step is covered in the next slide. Any guesses?

Sampling the missing data

$$
\begin{aligned}
p(\boldsymbol{Y}_{\text{miss}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\mu}, \Sigma) &\propto p(\boldsymbol{Y}_{\text{miss}}, \boldsymbol{Y}_{\text{obs}} \mid \boldsymbol{\mu}, \Sigma) \\
&= \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}}, \boldsymbol{y}_{i,\text{obs}} \mid \boldsymbol{\mu}, \Sigma) \\
&\propto \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}} \mid \boldsymbol{y}_{i,\text{obs}}, \boldsymbol{\mu}, \Sigma)
\end{aligned}
$$

How to proceed?

Sampling the missing data

$$
\begin{aligned}
p(\boldsymbol{Y}_{\text{miss}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\mu}, \Sigma) &\propto p(\boldsymbol{Y}_{\text{miss}}, \boldsymbol{Y}_{\text{obs}} \mid \boldsymbol{\mu}, \Sigma) \\
&= \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}}, \boldsymbol{y}_{i,\text{obs}} \mid \boldsymbol{\mu}, \Sigma) \\
&\propto \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}} \mid \boldsymbol{y}_{i,\text{obs}}, \boldsymbol{\mu}, \Sigma)
\end{aligned}
$$

How to proceed? We apply standard results about conditional distributions formed from partitions of multivariate normals:

$$
\begin{aligned}
\boldsymbol{y}_{[b]} \mid \boldsymbol{y}_{[a]}, \boldsymbol{\mu}, \Sigma &\sim \mathcal{MVN}\left(\boldsymbol{\mu}_{b|a}, \Sigma_{b|a}\right), \quad \text{where} \\
\boldsymbol{\mu}_{b|a} &= \boldsymbol{\mu}_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\boldsymbol{y}_{[a]} - \boldsymbol{\mu}_{[a]}) \\
\Sigma_{b|a} &= \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}
\end{aligned}
$$

Some macroscopic properties:

Sampling the missing data

$$
\begin{aligned}
p(\boldsymbol{Y}_{\text{miss}} \mid \boldsymbol{Y}_{\text{obs}}, \boldsymbol{\mu}, \Sigma) &\propto p(\boldsymbol{Y}_{\text{miss}}, \boldsymbol{Y}_{\text{obs}} \mid \boldsymbol{\mu}, \Sigma) \\
&= \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}}, \boldsymbol{y}_{i,\text{obs}} \mid \boldsymbol{\mu}, \Sigma) \\
&\propto \prod_{i=1}^{n} p(\boldsymbol{y}_{i,\text{miss}} \mid \boldsymbol{y}_{i,\text{obs}}, \boldsymbol{\mu}, \Sigma)
\end{aligned}
$$

How to proceed? We apply standard results about conditional distributions formed from partitions of multivariate normals:

$$
\begin{aligned}
\boldsymbol{y}_{[b]} \mid \boldsymbol{y}_{[a]}, \boldsymbol{\mu}, \Sigma &\sim \mathcal{MVN}\left(\boldsymbol{\mu}_{b|a}, \Sigma_{b|a}\right), \quad \text{where} \\
\boldsymbol{\mu}_{b|a} &= \boldsymbol{\mu}_{[b]} + \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}(\boldsymbol{y}_{[a]} - \boldsymbol{\mu}_{[a]}) \\
\Sigma_{b|a} &= \Sigma_{[b,b]} - \Sigma_{[b,a]}(\Sigma_{[a,a]})^{-1}\Sigma_{[a,b]}
\end{aligned}
$$

Some macroscopic properties:

- The conditional mean, $\boldsymbol{\mu}_{b|a}$, starts off at the unconditional mean, $\boldsymbol{\mu}_{[b]}$, but then is modified by $(\boldsymbol{y}_{[a]} - \boldsymbol{\mu}_{[a]})$ in a way that depends on the covariance $\Sigma_{[b,a]}$.
- The conditional variance $\Sigma_{b|a}$ is less than the unconditional variance $\Sigma_{[b,b]}$

Posterior Correlations

To each covariance matrix there corresponds a correlation matrix $\boldsymbol{C}$ given by

$$\boldsymbol{C} := \left\{ c_{jk} : c_{jk} = \Sigma_{[j,k]} / \sqrt{\Sigma_{[j,j]} \Sigma_{[k,k]}} \right\}$$

Simply taking the mean across samples, we obtain the approximation

$$\mathbb{E}[\boldsymbol{C} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] = \begin{bmatrix} 1.00 & 0.23 & 0.25 & 0.19 \\ 0.23 & 1.00 & 0.25 & 0.24 \\ 0.25 & 0.25 & 1.00 & 0.65 \\ 0.19 & 0.24 & 0.65 & 1.00 \end{bmatrix}$$

Notes:

- Bayesian paradigm again yielding unlimited access to posterior functionals of interest, without doing any extra inferential work!
- Correlations are generally of interest for multivariate normal models, but they are *especially* relevant to imputation.

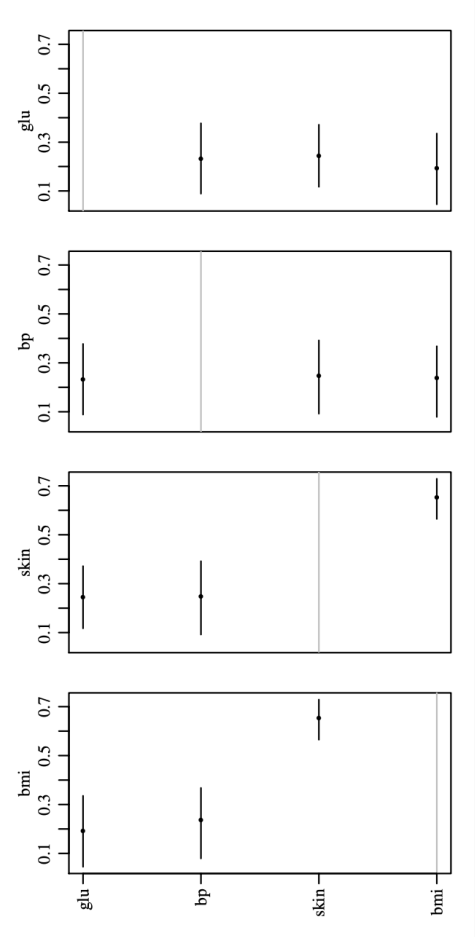


Figure 3: 95% posterior confidence intervals for correlations

Intelligent imputations

The posterior expectation gives a much better imputation than some flat fixed value.

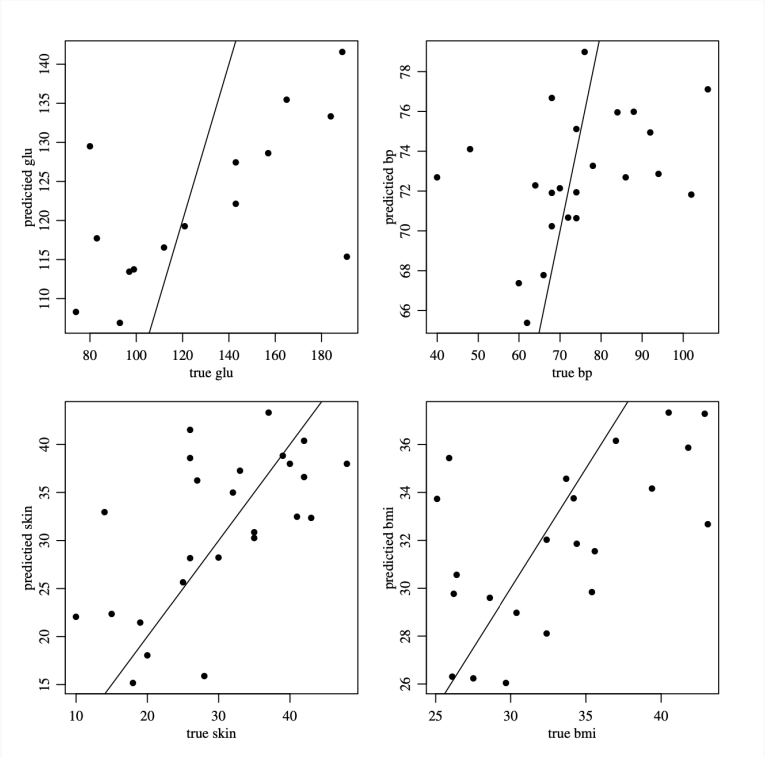


Figure 4: True values of the missing data vs. posterior expectations

Intelligent imputations

The posterior expectation gives a much better imputation than some flat fixed value.

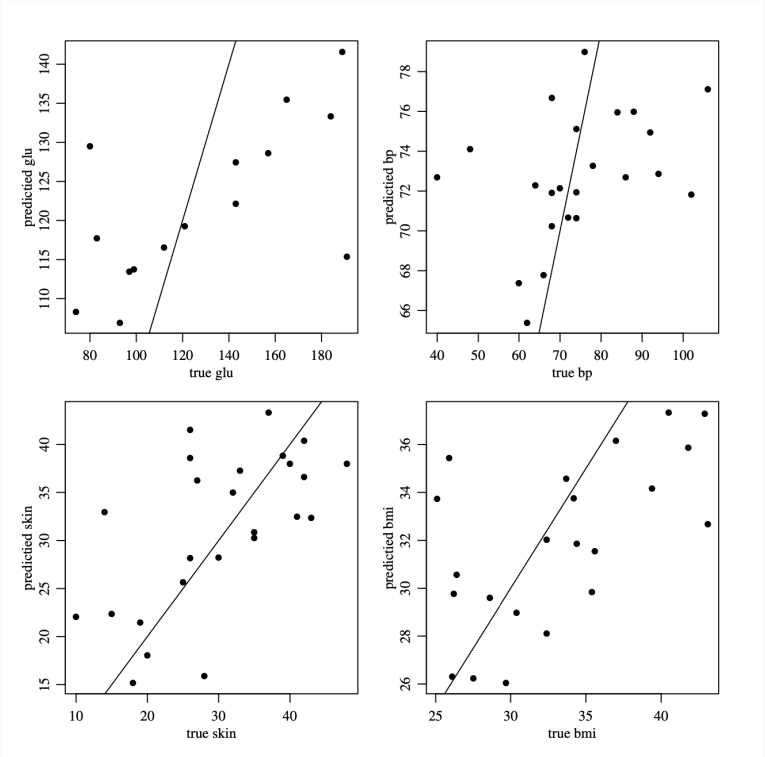


Figure 4: True values of the missing data vs. posterior expectations

Imputations are especially good for `skin` and `bmi`, due to their higher correlations.