INF367: Spring 2020
Exercise 10

Instructions:

The solutions are returned electronically via MittUiB by Monday 10.00. Grades are awarded for effort so scanned notes are fine if you solve exercises by hand (no need to make fancy latex files).

Students are encouraged to write computer programs to derive solutions whenever appropriate.

Tasks

Gibbs sampling for missing data

Suppose random variables X_i follow a bivariate normal distribution $X_i \sim N(\mu, \Sigma)$ where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

We have n observations. However, in some observations the first coordinate value is missing. Let us denote the observed values as X. That is, X consists of n observations X_i where either $X_i = (X_{i1}, X_{i2})$ (both coordinates observed) or $X_i = (?, X_{i2})$ (first coordinate unobserved). Furthermore, let Z denote the unobserved values. That is, $Z_i \in Z$ if $X_i = (?, X_{i2})$.

We assume that the covariance matrix Σ is known and we want to compute the posterior of the mean vector μ, that is, $P(\mu \mid X)$. We can approximate this posterior by sampling from $P(\mu, Z \mid X)$ and then marginalizing Z out. In this case, marginalization is easy: just ignore the Z values.

Do the following tasks:

1. Derive and implement a Gibbs sampler.
2. Load data `missing_data.csv`. Note that some values are missing. Learn the posterior distribution of μ assuming that the known correlation coefficient is $\rho = 0.8$.

Hints:

For a Gibbs sampler, we need to find the full conditional distributions $P(\mu \mid X, Z)$ and $P(Z \mid X, \mu)$.

You are free to use known properties of multivariate Gaussians.