ECE368: Probabilistic Reasoning

Aman Bhargava

January-April 2021

Contents

0.1 Introduction and Course Information

Course Information

- Professors: Prof. Saeideh Parsaei Fard and Prof. Foad Sohrabi
- Course: Engineering Science, Machine Intelligence Option
- Term: 2021 Winter

Main Course Topics

- Vector, temporal, and spatial models.
- Classification and regression model training.
- Bayesian statistics, frequentist statistics.

Chapter 1

Review Topics

See ECE286 notes for further reference

1.1 Review of Probability Functions

Probability Mass Function: For *discrete random variables*, $P_X(x)$denotes the probability that random variable X takes on value x.

Probability Density Function: For *continuous random variables*, the probability $\Pr\{X \in [x_1, x_2]\}$ is given by $\int_{x_1}^{x_2} f_X(x)dx$.

Joint PMF's and PDF's are similarly defined.

Marginal Probability Distributions: Given joing PMF $P_{X,Y}(x, y)$ or PDF $f_{X,Y}(x, y)$, we can **marginalize** them as follows:

$$P_X(x) = \sum_{y \in Y} P_{X,Y}(x, y) \tag{1.1}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \tag{1.2}$$

Conditional Probability Functions:

$$P_{Y|X}(y, x) = \frac{P_{X,Y}(x, y)}{P_X(x)} \tag{1.3}$$

Prior Probability: Probability **before** an additional observation is made (hence *prior*). Example: $P_X(x)$.

Posterior Probability: Probability **after** an observation is made (hence *poste*rior). Example: $P_{X|Y}(x, y)$.

Bayes Rule:

$$P(B|A) = P(A|B)\frac{P(B)}{P(A)} \tag{1.4}$$

1.2 Expectation, Correlation, and Independence

Expectation Value: $\mathbb{E}[x] = \sum_{x \in X} P_X(x) = \int_{-\infty}^{\infty} x f_X(x) dx$

Law or Large Numbers: $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_i = \mathbb{E}[X]$

Variance:

$$\begin{aligned} \mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[x])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned} \tag{1.5}$$

Covariance:

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}_{XY}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \end{aligned} \tag{1.6}$$

Correlation Coefficient:

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \tag{1.7}$$

- $\rho_{XY} \in [-1, 1]$
- $\rho > 0$ indicates positive correlation (line of best fit has positive slope).
- $\rho < 0$ indicates negative correlation.
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ **iff** X, Y **are uncorrelated**.

Independence

Theorem 1 *Independence Random variables X, Y are independent* ***iff***

$$P_{XY}(x, y) = P_X(x) \cdot P_Y(y) \tag{1.8}$$

This also means that $\rho_{XY} = 0$, $P(X|Y) = P(X)$, etc.

1.3 Laws of Large Numbers

Weak Law: Sample mean converges to the mean.

Strong Law: If $\{x_i\}$ are **independent, identically distributed** (i.i.d.) random variables with mean μ, then the **probability of** the sample mean $= \mu$ is 1 as $n \to \infty$.

Chapter 2

Parameter Estimation

2.1 Estimation Terminology

- $\hat{\theta}_n$ is an **estimator** of some unknown parameter θ.
- **Estimation Error:** $\hat{\theta}_n - \theta$
- **Bias** of estimator: $\mathbb{E}[\hat{\theta}_n] - \theta$
  - **Unbiased** estimator: Bias$= 0 = \mathbb{E}[\hat{\theta}_n] - \theta$.
  - **Asymptotically Unbiased:** $\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_n] = \theta$ for all θ.
- **Consistency:** Estimator is consistent if $\lim_{n\to\infty} \hat{\theta}_n = \theta$.

2.2 Maximum Likelihood Estimation

Framing: Let random variable $\vec{X} = [X_1, X_2, ..., X_n]$ be defined by either

1. Joint PMF $P_{\vec{X}}(\vec{x}; \theta)$
2. Joint PDF $f_{\vec{X}}(\vec{x}; \theta)$

$\vec{x}$ is a series of measurements.

Maximum Likelihood Estimation: The ML estimate of model parameter θ is

$$\hat{\theta}_n = \text{argmax}_\theta P_{\vec{X}}(\vec{x}; \theta) \tag{2.1}$$

Independent, identically distributed case: If each $x_i \in \vec{x}$ are independent and identically distributed, then

$$P_{\vec{X}}(\vec{x};\theta) = \prod_{i=1}^{n} P_X(x_i;\theta) \tag{2.2}$$

Which we can convert to a summation by taking the **log-likelihood** (recall that logarithm is monotonically increasing, so maximizing log-likelihood is equialent to maximizing likelihood).

$$\hat{\theta}_n = \arg\max_{\theta}(\sum_{i=1}^{n} \log P_X(x_i;\theta)) \tag{2.3}$$

2.3 Frequentist vs. Bayesian Statistics

Frequentist: In **classical statistics**, probability is taken to be approximately equal to the **frequency of events**. Model parameters are assumed to have some deterministic, fixed value (even though they might be unknown).

Bayesian Statistics: Model parameters are treated as **random variables** with their own distributions.

- Generally the more modern approach.
- We are most interested in the **joint probability distribution** of model parameters and model arguments (e.g., $f_x(x,\theta)$).
- **Main criticism:** probabilities are assigned to unrepeatable events (arguably violates the definition of probability as the limit of event frequency).

2.4 Maximum a Posteri Estimation (MAP)

$$\begin{aligned}\hat{\theta}_{map} &= \arg\max_{\theta} f_{\theta|x}(\theta|x) \\ &= \arg\max_{\theta} f_{X|\theta}(x|\theta)\frac{f_\theta(\theta)}{f_X(x)}\end{aligned} \tag{2.4}$$

Where $f_\theta(\theta)$ is the **prior distribution** of model parameter.

- If $f_\theta(\theta)$ is uniform, we will still get the same answer as a **maximum likelihood** estimation.

2.4.1 Picking a Prior Distribution

Best Practice: Pick a distribution of the same form as $f_{X|\theta}(x|\theta)$ (called "conjugate pair").

Beta Distribution: Used for **binomial distribution**.

- Binomial distribution:

$$P_{X=k|\theta} = \binom{n}{k}\theta^k(1-\theta)^{n-k} \tag{2.5}$$

  where

  - θ: Probability of success on each Bernoulli trial.
  - n: Total number of trials.
  - k: Total number of successful trials.

- **Beta Distribution:**

$$f_\theta(\theta;\alpha,\beta) = \begin{cases} c\theta^{\alpha-1}(1-\theta)^{\beta-1} & \text{for } \theta \in [0,1] \\ 0 & \text{else} \end{cases} \tag{2.6}$$

  Where

  - α, β are customizable parameters.
  - $c = [\Gamma(\alpha+\beta)]/[\Gamma(\alpha)\Gamma(\beta)]$
  - $\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}du$
  - $\Gamma(x+1) = x\Gamma(x)$ for all $x \in \mathbb{R}$.
  - $\Gamma(n+1) = n!$ for integer n.
  - $\therefore c = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}$ for integer α, β.
  - $\mu_f = \mathbb{E}[f_\theta(\theta)] = \frac{\alpha}{\alpha+\beta}$
  - Maximum likelihood $\arg\max_\theta f_\theta(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$

2.5 Conditional Expectation Estimator

Key Idea: Find the **expected value** for the estimator given your observations.

$$\hat{\theta}_{conditionalexpectation} = \mathbb{E}[\theta|\vec{X} = \vec{x}] = \int_{-\infty}^{\infty} \theta f_{\theta|\vec{x}}(\theta|\vec{x}) \tag{2.7}$$

2.6 Bayesian Least Mean Square Estimator (LMS)

Key Idea: To estimate random variable model parameter θ, we find

$$\hat{\theta}_{LMS} = \arg\min_{\hat{\theta}} \mathbb{E}[(\theta - \hat{\theta})^2] \tag{2.8}$$

- $\hat{\theta}_{LMS} = \mathbb{E}[\theta]$ achieves the goal.
- **Equivalently:** We can also find

$$\hat{\theta}_{LMS} = \arg\min_{\hat{\theta}} (\mathbb{E}[\theta - \hat{\theta}])^2 \tag{2.9}$$

Chapter 3

Hypothesis Testing

Goal: Given two hypotheses H_0, H_1 and observation $\boldsymbol{x} = (x_0, \ldots, x_n)$, we wish to decide which hypothesis is **better**.

Error Types: These are generally with respect to H_0, or the "null hypothesis".

- **Type I:** False **rejection**. We reject hypothesis H_i despite it being the correct one.
- **Type II:** False **acceptance**. We accept hypothesis H_i despite it being false.

3.1 Likelihood Ratio Test

$$\mathbb{L}(\boldsymbol{x}) = \frac{P_x(\boldsymbol{x}; H_1)}{P_x(\boldsymbol{x}; H_0)} \lessgtr z \tag{3.1}$$

If $\mathbb{L}(\boldsymbol{x}) > z$, we **accept** H_1. Else, we accept H_0. $z = 1$ corresponds to the maximum likelihood decision rule.

Neyman-Pearson Lemma: We let α represent the probability of **false rejection** and β represent the probability of **false acceptance** (i.e., type I and type II error probability respectively).

$$P(\mathbb{L}(\boldsymbol{x}) > z; H_0) = \alpha \tag{3.2}$$

$$P(\mathbb{L}(\boldsymbol{x}) \leq z; H_1) = \beta \tag{3.3}$$

There is a direct tradeoff between α and β – they are inversely proportional (i.e., we cannot get better overall confidence "for free" with the same data).

There is equivalence between selecting some likelihood cutoff z and some γ for $\boldsymbol{x} \lessgtr \gamma$.

3.2 Bayesian Hypothesis Testing

The likelihood ratio test is essentially a **maximum likelihood** method. Bayesian hypothesis testing is equivalent to **maximum a posteri** methods.

- **Goal:** Choose the most probable hypothesis *given the data*.
- **Given:** Hypotheses $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_M\}$, data $\boldsymbol{x} = (x_1, \ldots, x_n)$.
- **Method:** Select the optimal hypothesis based on

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta \in \boldsymbol{\theta}} P(\theta|\boldsymbol{x}) \\ &= \arg\max_{\theta \in \boldsymbol{\theta}} P(\boldsymbol{x}|\theta) P(\theta) \end{aligned} \tag{3.4}$$

By maintaining $\boldsymbol{x}$ as a free variable in these computations, the rejector regions $\mathcal{R}$ for $\boldsymbol{x}$ can be found relatively easily.

3.3 Gaussian Vector Distribution

Scalar Gaussian Normal Distribution:

$$\begin{aligned} f_x(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp[\frac{-(x-\mu)^2}{2\sigma^2}] \\ &\Leftrightarrow x \sim \mathcal{N}(\mu, \sigma^2) \end{aligned} \tag{3.5}$$

Gaussian Vector

$$f_{\vec{x}}(\vec{x}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp[\frac{-1}{2}(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})] \tag{3.6}$$

Where

- D is the dimension of the vectors x.
- $\vec{\mu} \in \mathbb{R}^D$ is the mean vector.

- $\vec{\mu} = \mathbb{E}[\vec{x}]$

- $\Sigma \in \mathbb{R}^{D \times D}$ is the covariance matrix. It is **positive semidefinite**.
  - $\Sigma = \mathbb{E}[(\vec{x} - \vec{\mu})(\vec{x} - \vec{\mu})^T]$

3.3.1 Eigen Analysis of Gaussian Vectors

From ECE367 : All PSD matrices A have **orthogonal** eigenvectors. We can also arbitrarily scale them to be ortho**normal**.

$$A = Q\Lambda Q^T \tag{3.7}$$

Where each column of Q is an eigenvector and $\Lambda = \text{diag}(\text{eigen values})$. $QQ^T = I$ and $Q^T = Q^{-1}$ by orthonormalcy.

Applying to Gaussian Vectors: We note the term $\frac{-1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})$. Since Σ is PSD we can decompose it into

$$\begin{aligned} \Sigma &= Q\Lambda Q^T \\ \Sigma^{-1} &= Q\Lambda^{-1}Q^T \end{aligned} \tag{3.8}$$

We define a helper variable $\vec{y}$ as follows:

$$\begin{aligned} &\vec{y} \equiv Q^T(\vec{x} - \vec{\mu}) \\ &\Rightarrow \vec{y}^T = (\vec{x} - \vec{\mu})^T Q \\ &\Rightarrow (\vec{x} - \vec{\mu})^T\Sigma^{-1}(\vec{x} - \vec{\mu}) = \vec{y}^T Q^T\Sigma^{-1}Q\vec{y} \\ &= \vec{y}^T\Lambda^{-1}\vec{y}^T \end{aligned} \tag{3.9}$$

The big payoff is that $\vec{y}$ is a random variable with a **diagonal** covariance matrix (i.e., independent components). $\vec{y}$ also has zero mean!

$$\begin{aligned} f_y(\vec{y}) &= \frac{1}{(2\pi)^{D/2}|\Lambda|^{1/2}} \exp(\frac{-1}{2}\vec{y}^T\Lambda^T\vec{y}) \\ &= \prod_{i=1}^{D} \frac{1}{\sqrt{2\pi\lambda_i}} \exp[\frac{-y_i^2}{2\lambda_i}] \end{aligned} \tag{3.10}$$

3.4 Gaussian Estimation

Given: Observations $x_i \in \mathbb{R}$ where we know that were generated by $x_i = \theta + w_i$ and $w_i \sim \mathcal{N}(0, \sigma_i^2)$. In other words, each x_i is a "measure" of the same mean value θ, but there is some zero-mean noise w_i with variance σ_i^2. Note that each data point has its own variance.

Goal: Estimate θ.

3.4.1 Maximum Likelihood

$$\begin{aligned}\hat{\theta}_{ML} &= \arg\max_{\theta} f_x(\vec{x};\theta) \\ &= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp[\frac{-1(x-\theta)^2}{2\sigma_i^2}] \\ \hat{\theta}_{ML} &= \frac{\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}}\end{aligned} \tag{3.11}$$

Intuitively, this corresponds to a weighted sum of each x_i. Weight is proportional to $\frac{1}{\sigma_i^2}$, which corresponds to "certainty" in the validity of the data point in representing θ.

3.4.2 MAP Estimation

We assume a **conjugate prior** distribution for $\theta \sim \mathcal{N}(x_0, \sigma_0^2)$. Conveniently enough, these x_0, σ_0 are functionally identical to just having another data point!

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} f_\theta(\theta) f(x|theta) \\ &= \dots \\ \hat{\theta}_{MAP} &= \frac{\sum_{i=0}^{n} \frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n} \frac{1}{\sigma_i^2}}\end{aligned} \tag{3.12}$$

Note that the sums now start at zero to incorporate the prior distribution's information!

Chapter 4

Statistical Machine Learning

4.1 Naive Bayesian Classifier

Technically this was in the first half, but it fits well to motivate LDA and QDA.

Goal: Let $\theta \in \{1, 2\}$ be the class of data points $\vec{x} \in \mathbb{R}^n$. We wish to find $P(\theta|\vec{x})$.

Naive Bayesian Assumption: Each component of $\vec{x}$ is **independent** with respect to class θ. By the probability axioms relating to independence,

$$P_{\vec{x}|\theta}(\vec{x}|\theta) = \prod_{i=1}^{n} P_{x_i|\theta}(x_i|\theta) \tag{4.1}$$

Classification Equations:

$$P(\theta|\vec{x}) = \frac{P(\theta)\prod_{i=1}^{n} P_{x_i|\theta}(x_i|\theta)}{P(\vec{x})} \tag{4.2}$$

Where $P(\vec{x}) = \sum_\theta [P(\theta)\prod_{i=1}^{n} P_{x_i|\theta}(x_i|\theta)]$

$$P(\theta = 1)\prod_{i=1}^{n} P_{x_i|\theta}(x_i|\theta = 1) \lessgtr P(\theta = 2)\prod_{i=1}^{n} P_{x_i|\theta}(x|\theta = 2) \tag{4.3}$$

Bag of words: For classifying text, we choose a set of W of the most common words. In our vectors $\vec{x}$, each index x_i corresponds to whether word

w_i appears in the given text.

$$P_{x_i|\theta}(x_i = w_d|\theta = j) = \frac{\text{occurrences of } w_i \text{ in set } j}{\text{total words in set } j} \tag{4.4}$$

Laplace Smoothing: Even if we perform the computation in the log domain, we might run into a $P_{x_i|theta}(x_i|\theta) = 0$ if we lack a datapoint (word occurrence) for a given class. Laplace smoothing simply appends the vocabulary set W to each class (i.e., all words in the set occur at least once a priori).

4.2 Linear Discriminant Analysis (LDA)

Goal: Create algorithm LDA : $\vec{x}_i \to c_i$ where $\vec{x}_i$ is a data point and $c_i \in C$ is the class it belongs to.

General Methodology: We use **Bayesian hypothesis testing** with variable classes we model with normal distributions. An important simplifying assumption is that each class c has different $\vec{\mu}_c$ **BUT** they all have the same Σ. This is what makes our decision boundaries **linear**!

Classification Equations:

1. $\hat{y}(\vec{x}) = \arg\max_{c \in C} \beta_c^T \vec{x} + \gamma_c$ where
   - $\beta_c = \Sigma^{-1}\mu_c$
   - $\pi_c = P(c)$ (prior probability of class c)
   - $\gamma_c = \log(\pi_c) - \frac{1}{2}\mu_c^T \Sigma^{-1} \mu_c$
2. Posterior probability of class c:

$$P(c|\vec{x}) = \frac{\exp(\beta_c^T x + \gamma_c)}{\sum_{c' \in C} \exp(\beta_{c'}^T x + \gamma_{c'})} \tag{4.5}$$

4.3 Quadratic Discriminant Analysis (QDA)

Goal: Same as LDA.

Assumptions: We are given μ_c, Σ_c, π_c for each class, and each $\vec{x}_i \sim \mathcal{N}(\mu_c, \Sigma_c)$ for some $c \in C$.

Classification Equations:

$$\hat{y}(\vec{x}) = \arg\max_{c \in C}[\log(\tilde{\pi}_c) - \frac{1}{2}(x-\mu_c)^T\Sigma_c^{-1}(x-\mu_c)] \tag{4.6}$$

Where $\tilde{\pi}_c = \frac{1}{(2\pi)^{D/2}|\Sigma_c|^{1/2}}\pi_i$

$$P(c|x) = \pi_c \frac{1}{(2\pi)^{D/2}|\Sigma_c|^{1/2}} \exp(\frac{-1}{2}(x-\mu_c)^T\Sigma^{-1}(x-\mu_c))\frac{1}{P(x)} \tag{4.7}$$

Where $P(x)$ is given by

$$\begin{aligned} P(x) &= \sum_{c' \in C} P(x|c')P(c') \\ &= \sum_{c' \in C} \mathcal{N}(x; \mu_c, \Sigma_c)P(c') \end{aligned} \tag{4.8}$$

Calculating μ_c, Σ_c The maximum likelihood estimation of these are as follows:

$$\hat{\mu}_c = \frac{1}{n}\sum_{i=1}^{n} x_i \ \forall x_i \in c \tag{4.9}$$

$$\hat{\Sigma}_c^{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T \ \forall x_i \in c \tag{4.10}$$

PROBLEM: $\hat{\Sigma}_c^{ML}$ is a biased estimator. For the $D = 1$ case, $\mathbb{E}[\hat{\sigma}_c^{ML}] = \frac{n-1}{n}\sigma^2$. To correct, we use the following:

$$\hat{\Sigma}_c^{corrected} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_c)(x_i - \hat{\mu}_c)^T \ \forall x_i \in c \tag{4.11}$$

The primary problem with these sorts of classifiers is that Σ has high dimensionality. If we use the same Σ for all classes, we get LDA. If we force Σ_c to be diagonal, we get a naive Bayesian classifier.

4.4 General Bayesian Inference on Gaussian Vectors

Given: We know that $\vec{z} \sim \mathcal{N}(\vec{\mu}, \Sigma)$. We know the value of only part of the $\vec{z}$ vector, and want to determine the probability distribution and ML/MLE/MMSE estimation for the remaining indices we don't know.

$$\vec{z} = \begin{bmatrix} \vec{x} \\ \vec{y} \end{bmatrix} \tag{4.12}$$

We are given the value of $\vec{y}$ and want to know $\vec{x}$.

Key Tools:

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \tag{4.13}$$

Where $\Sigma_{xy} = \mathbb{E}[(x - \mu_x)(x - \mu_x)^T]$.

- *The maximum of a Gaussian is at the mean!*
- $f_{x|y} \sim \mathcal{N}$
- $\hat{x}_{MAP} = \hat{x}_{MMSE,LMS} = \mathbb{E}[\vec{x}|\vec{y}]$

Results:

$$f_{x|y}(x|y) \sim \mathcal{N}(\mu_{x|y}, \Sigma_{x|y}) \tag{4.14}$$

Where

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \tag{4.15}$$

$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \tag{4.16}$$

Which is pretty cool, especially considering that $\hat{x}_{MAP,MMSE} = \mu_{x|y}$ is now linear with $\vec{y}$. We can think of $\Sigma_{x|y}$ as the "remaining uncertainty" in $\vec{x}$ after we receive information $\vec{y}$.

4.5 Linear Gaussian Systems

Given: $P(x) \sim \mathcal{N}(\mu_x, \Sigma_x)$. $y = Ax + b + z$ where

- A is a known matrix.
- b is a known vector.
- $z \sim \mathcal{N}(0, \Sigma_z)$
- x, z are independent.

Goal: We observe y. What are our estimates $\hat{x}_{MAP}$, $\hat{x}_{MMSE}$ that correspond to this particular y?

Solution: Unsurprisingly, x, y have Gaussian distributions.

$$\hat{x}_{MAP,MMSE} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \tag{4.17}$$

Where

$$\mu_y = A\mu_x + b \tag{4.18}$$

$$\Sigma_{xy} = \Sigma_x A^T \tag{4.19}$$

$$\Sigma_{yy} = A\Sigma_x A^T + \Sigma_z \tag{4.20}$$

$$\Sigma_{x|y} = (\Sigma_x^{-1} + A^T\Sigma_z^{-1}A)^{-1} \tag{4.21}$$