156 – Machine Learning

University of California, Los Angeles

Duc Vu

Summer 2021

This is math 156 – Machine Learning, an introductory course on mathematical models for pattern recognition and machine learning. It's instructed by Professor Zosso, and we meet weekly on MWTh from 9:00 am to 10:50 am. The textbook used for the class is *Pattern Recognition and Machine Learning* by *Bishop*. You can find the other course notes through my blog site. Any error appeared in this note is my responsibility and please email me if you happen to notice it.

Contents

List of Theorems

List of Definitions

§1 Lec 1: Jun 21, 2021

§1.1 Introduction & Probability Review

According to Wikipedia, **Machine Learning** is a scientific discipline that deals with the construction and study of algorithms that can learn from data.

$$\text{Input(data)} \to \boxed{\text{Model}} \to \text{Output(Predictions/Decisions)}$$

From §1.2 of the book, let's review a bit on probability.

- Discrete random variable X, value $\{x_i\}$

$$\text{prob}(X = x_i) = p(x_i) = \frac{n_i}{N}$$

and

$$\sum_i \text{prob}(X = x_i) = \sum_i p(x_i)$$

For multiple random variables, $X, Y \in \{x_i\} \times \{y_i\}$

1. $\text{prob}(X = x_i, Y = y_i) = \frac{n_{ij}}{N} = p(x_i, y_i)$ – joint probability
2. $\text{prob}(X = x_i) = \sum_j \text{prob}(X = x_i, Y = y_j)$ – marginal probability
3. $\text{prob}(X = x_i | Y = y_j) =$ conditional

$$\underbrace{p\left(x_i|y_j\right)}_{\text{conditional}} \cdot \underbrace{p(y_j)}_{\text{marginal}} = \underbrace{p(x_i, y_j)}_{\text{joint}}$$

$\implies$ product rule

Bayes' Rule:

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

- Continuous random variable $X \in \mathbb{R}$

$$\text{prob}(X = x_i) = 0 \text{ in general}$$

So we consider probability densities instead where

$$p(x) \geq 0$$

s.t. $p(x)$ can be greater than 1. In addition,

$$\int_{-\infty}^{\infty} p(x) = 1$$

Within a neighborhood $a \leq b$, we have

$$\text{prob}(a \leq x \leq b) = \int_a^b p(x)\, dx$$

Sum rule:

$$\int \underbrace{p(x, y)}_{\text{joint pdf}}\, dy = \underbrace{p(x)}_{\text{marginal pdf}}$$

Product rule:

$$p(x, y) = p(y|x)p(x) = p(x|y)p(y)$$

Bayes' Rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Expectations & Covariances

Expectations:

Definition 1.1 — Expectation is defined as

$$\begin{aligned}\mathbb{E}[f] &:= \sum_i p(x_i)f(x_i) \\ \text{or} &:= \int_{\mathbb{R}} p(x)f(x)\,dx\end{aligned}$$

"Average value of a function $f : \mathbb{R} \to \mathbb{R}$ under a probability distribution $p(x)$"

In practice, we need to estimate p from data.

$$\text{Sampling Approximation: } \mathbb{E}[f] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

Definition 1.2 — Marginal expectation is defined as

$$\mathbb{E}_x[f](y) := \sum_x p(x)f(x,y)$$

Conditional expectation:

$$\mathbb{E}_x[f|y] := \sum_x p(x|y)f(x)$$

Covariances:

Definition 1.3 — Variance is defined as

$$\begin{aligned}\text{var}[f] &:= \mathbb{E}\left[(f(x) - \mathbb{E}[f])^2\right] \\ &= \mathbb{E}[f^2] - \mathbb{E}[f]^2\end{aligned}$$

Covariance (random variables) is defined as

$$\begin{aligned}\text{cov}[x,y] &:= \mathbb{E}\left[(x - \mathbb{E}[x])(y - \mathbb{E}[y])\right] \\ &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

For vectors $\vec{x}, \vec{y} \in \mathbb{R}^D$, the covariance matrix is

$$\mathbb{E}\left[(\vec{x} - \mathbb{E}[\vec{x}])(\vec{y} - \mathbb{E}[\vec{y}])^\top\right]$$

Question 1.1. How does this fit in within the context of machine learning?

In machine learning, there are usually two approaches to find the "optimal prediction"

- Frequentist approach: maximize likelihood

$$\max_w p(D|w)$$

- Bayesian approach: maximize posterior

  $$\text{posterior through Bayes': } p(w|D) = \frac{p(D|w) \cdot p(w)}{p(D)}$$

  s.t.

  $$\max_w p(w|D) \sim p(D|w) \cdot p(w)$$

  where D represents data, and w is parameters.

Gaussian noise model:

$$p\left(t_n|x_n, w, \beta\right) = N\left(t_n|y(x_n, w), \frac{1}{\beta}\right)$$

Given training data $\{(x,t)\}$, we can determine optimal parameters w, β by

1. Frequentist: maximize likelihood

   $$p(t|x, w, \beta) \overset{\text{i.i.d}}{=} \prod_{n=1}^{N} N\left(t_n|y(x_n|w), \beta^{-1}\right)$$

2. include a prior: $p(w|\alpha) = N(w|0, \alpha^{-1})$

   $$\implies \text{posterior: } p\left(w|x, t, \alpha, \beta\right) \propto p\left(t|x, w, \beta\right) p(w|\alpha)$$

   Then, we can estimate

   $$\min_w \left\{\frac{\beta}{2} \sum_{n=1}^{N} \left(y(x_n, w) - t_n\right)^2 + \frac{\alpha}{2} w^\top w\right\}$$

3. Fully Bayesian: not just point estimates $\implies$ predictive distribution

   $$p\left(t_i|x_i, x, t\right) = \int \underbrace{p\left(t_i|x_i, w\right)}_{\text{model}} \underbrace{p\left(w|x, t\right)}_{\text{posterior}} dw$$

§1.2 Gaussian Distribution

Definition 1.4 (Gaussian Distribution) — The 1-D Gaussian distribution is defined as

$$N\left(x|\mu, \sigma^2\right) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the mean and σ^2 is the variance.
For D-dimensional,

$$N\left(\vec{x}|\vec{\mu}, \sum\right) := \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\sum|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\top \sum^{-1}(x-\mu)}$$

where Σ is the covariance matrix and $|\Sigma|$ is the determinant of Σ.

Consider $x \in \mathbb{R}^D$, $x \sim N$. Assume

$$x = \begin{bmatrix} x_a \\ x_b \end{bmatrix}$$

where x_a is unknown and x_b is given component.

$$x \sim N\left(\begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}\right)$$

Note that

$$\Sigma = \Sigma^\top$$

Also, we define the precision matrix Λ as

$$\begin{aligned}\Lambda &\coloneqq \Sigma^{-1} \\ &= \begin{bmatrix}\Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}\end{aligned}$$

Unfortunately, $\Lambda_{aa} \neq \Sigma_{aa}^{-1}$ and similar result applies for b.

Question 1.2. What can we say about $p(x_a|x_b)$?

Use product rule:

$$p\left(x_a|x_b\right) \cdot p(x_b) = p(x_a, x_b)$$

where $p(x_b)$ is a constant w.r.t. x_a

$$\implies p(x_a|x_b) \propto p(x_a, x_b)$$

Let's look at quadratic form in exponential only.

$$\begin{aligned}-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) = -\frac{1}{2}(x_a-\mu_a)^\top\Lambda_{aa}(x_a-\mu_a) - \frac{1}{2}(x_a-\mu_a)^\top\Lambda_{ab}(x_b-\mu_b) \\ -\frac{1}{2}(x_b-\mu_b)^\top\Lambda_{ba}(x_a-\mu_a) - \frac{1}{2}(x_b-\mu_b)^\top\Lambda_{bb}(x_b-\mu_b)\end{aligned}$$

Also,

$$\text{other side } = -\frac{1}{2}x_a^\top\Sigma_{a|b}^{-1}x_a + x_a^\top\Sigma_{a|b}^{-1}\mu_{a|b} + \text{const}$$

- Quadratic terms need to match

$$\begin{aligned}-\frac{1}{2}x_a^\top\Sigma_{a|b}^{-1}x_a &= -\frac{1}{2}x_a^\top\Lambda_{aa}x_a \\ \implies \Sigma_{a|b}^{-1} &= \Lambda_{aa}\end{aligned}$$

- Linear terms in x_a

$$\begin{aligned}x_a^\top\Sigma_{a|b}^{-1}\mu_{a|b} &= x_a^\top\Lambda_{aa}\mu_{a|b} \\ \Lambda_{aa}\mu_{a|b} &= \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b) \\ \implies \mu_{a|b} &= \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b-\mu_b)\end{aligned}$$

Note that

$$\begin{aligned}\Lambda_{aa} &= \left(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1} \\ \Lambda_{ab} &= -\Lambda_{aa}\Sigma_{ab}\Sigma_{bb}^{-1}\end{aligned}$$

Thus,

$$\begin{cases}\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b-\mu_b) \\ \Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\end{cases}$$

§2 Lec 2: Jun 23, 2021

§2.1 Gaussian Distribution (Cont'd)

Let's start with a set of observations:

$$X = \{\vec{x}_1, \dots, \vec{x}_N\} \quad N \text{ data points where each } \vec{x}_n \in \mathbb{R}^D$$

and each $\vec{x}_n \sim N(\mu, \Sigma)$. As usual, there are two approach to this.

- Maximum likelihood: given the data, what μ, Σ are most probable/likely?

  $$\max_{\mu,\Sigma} p\left(X|\mu,\Sigma\right)$$

  Model assumption: $\vec{x}_n$ are i.i.d (independently, identically distributed). From i.i.d, we have

  $$\begin{aligned} p\left(X|\mu,\Sigma\right) &= \prod_{n=1}^{N} p\left(\vec{x}_n|\mu,\Sigma\right) \\ &= \prod_{n=1}^{N} N\left(\vec{x}_n|\mu,\Sigma\right) \end{aligned}$$

  This is tricky to do, so let's minimize the negative log likelihood

  $$\begin{aligned} \min_{\mu,\Sigma} -\ln p\left(X|\mu,\Sigma\right) &= -\ln \prod_{n=1}^{N} \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x_n-\mu)^\top \Sigma^{-1}(x_n-\mu)} \\ &= \cancel{-N\ln\frac{1}{(2\pi)^{\frac{D}{2}}}} - N\ln\frac{1}{|\Sigma|^{\frac{1}{2}}} + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^\top\Sigma^{-1}(x_n-\mu) \\ &= \frac{N}{2}\ln|\Sigma| + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^\top\Sigma^{-1}(x_n-\mu) + C \end{aligned}$$

  As the domain is unbounded (unconstrained optimization problem) and objective function is convex, so to find optimal μ, we set $\frac{d}{d\mu} = 0$. Then

  $$\begin{aligned} \frac{1}{2}\sum_{n=1}^{N}\Sigma^{-1}(x_n-\mu) &= 0 \\ \sum_{n=1}^{N}\Sigma^{-1}x_n &= N\Sigma^{-1}\mu \\ \implies \mu &= \frac{1}{N}\sum_{n=1}^{N}x_n \end{aligned}$$

- Maximum a posteriori (MAP)

  $$\max_{\mu} p\left(\mu,\Sigma|X\right) \overset{\text{Bayes'}}{\implies} \max_{\mu} p\left(X|\mu,\Sigma\right)\cdot p\left(\mu\right)$$

e.g., $p(\mu|\mu_0, \Sigma_0) = N(\mu|\mu_0, \Sigma_0)$. We have

$$-\ln p(X|\mu, \Sigma) \cdot p(\mu|\mu_0, \Sigma_0)$$

$$\min_{\mu} \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^\top \Sigma^{-1} (x_n - \mu) + \frac{1}{2} (\mu - \mu_0)^\top \Sigma_0^{-1} (\mu - \mu_0)$$

$$\frac{d}{d\mu} = 0 : \sum_{n=1}^{N} \Sigma^{-1} (x_n - \mu) + \Sigma_0^{-1} (\mu - \mu_0) = 0$$

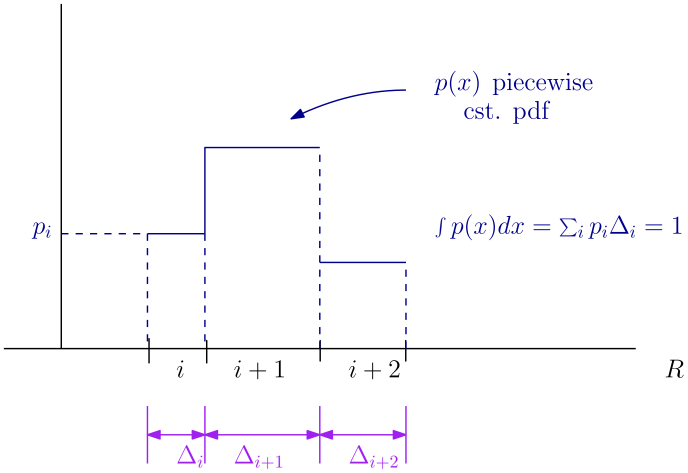
$$\implies \mu_{\text{MAP}} = \left(N\Sigma^{-1} + \Sigma_0^{-1}\right)^{-1} \left(N\Sigma^{-1}\overline{x} + \Sigma_0^{-1}\mu_0\right)$$

§2.2 Non-parametric Probability Density Function (Estimation)

Let's consider the following

- Histograms
- partition domain of x into distinct bins of width $\triangle_i$
- count number of observations n_i of x falling into bin i
- divide by $N, \triangle_i$ to get a pdf.

$$p_i = \frac{n_i}{N\Delta_i} \text{ is density over bin } i$$



We often partition the domain uniformly, i.e., $\Delta_i = \Delta$

refer to fig 2.24 in textbook for other cases

Consider a region $R \subseteq \mathbb{R}^D$. The probability of a randomly chosen point will fall into R (according to pdf of $p(x)$ is

$$p = \int_R p(x)\, dx$$

Collect N samples; a fraction K of which will fall into R. So $K \sim \text{Binomial}(N, p)$

$$\mathbb{E}\left[\frac{K}{N}\right] = p$$

$$\text{var}\left[\frac{K}{N}\right] = \frac{p(1-p)}{N}$$

$$\text{var}\left[\frac{K}{N}\right] \xrightarrow[N\to\infty]{} 0$$

For large N, $\frac{K}{N} \approx P \implies K \approx N \cdot P$. Also, we want R big so that there are plenty of points in there. On the other hand, we want R small s.t. $p(x) \sim$ constant over R where $p = p(x)V$ in which V is the volume of R. Thus,

$$p(x) = \frac{K}{NV}$$

For histogram: we fix V and measure $\frac{K}{N}$. For the kernel, it's essentially the same but bin locations are not predefined.
Kernel Approach: If we want to know $p(x)$ at arbitrary x, we put a bin of predefined size around x then count $\frac{K}{N}$ for that bin.
Pick a smooth kernel, e.g., the Gaussian

$$p_h(x) \coloneqq \frac{1}{N} \sum_{n=1}^{N} \frac{1}{(2\pi h^2)^{\frac{D}{2}}} e^{-\frac{\|x-x_n\|_2^2}{2h^2}}$$

where h is standard deviation of Gaussian. Recall from 131BH that this is a convolution.

$$(f * g)(x) \coloneqq \int f(y) g(x-y)\, dy$$

So $k * \sum \delta(-x_n)$. More general,

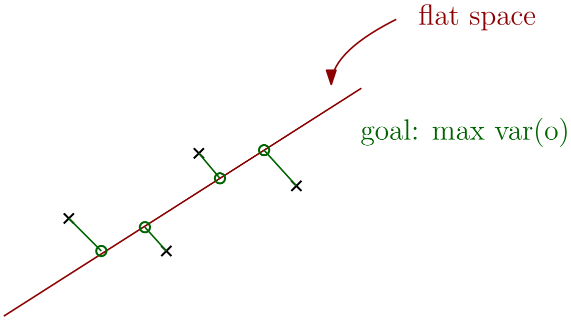
$$\begin{cases} k(u) \geq 0 \\ \int k(u)\, du = 1 \end{cases}$$

is sufficient criteria to be a kernel for kernel density estimation (KDE).

§3 Lec 3: Jun 24, 2021

§3.1 Principal Component Analysis

Maximum Variance Formulation: consider $\{x_n\}$, $n = 1, \ldots, N$, $x_n \in \mathbb{R}^D$. The goal is to project x onto a flat space with dimension $M \ll D$ while maximizing the variance of the projected data.



Let's start with $M = 1$ (a line) defined by a single vector $\vec{u} \in \mathbb{R}^D$ with unit norm, i.e.,

$$u_1^\top u_1 = \langle u_1, u_1 \rangle = \|u_1\|_2^2 = 1$$

Define: $\overline{x} = \frac{1}{N}\sum_{n=1}^N x_n$. Note that the variance before projection is

$$\text{var} = \frac{1}{N}\sum_{n=1}^N (x_n - \overline{x})^2$$

and after projection is

$$\text{var} = \frac{1}{N}\sum_{n=1}^N \left(u_1^\top x_n - u_1^\top \overline{x}\right)^2 = u_1^\top S u_1$$

with

$$\begin{aligned} S &= \frac{1}{N}\sum_{n=1}^N (x_n - \overline{x})(x_n - \overline{x})^\top \\ &= \text{cov}(x) \end{aligned}$$

Our optimization goal is

$$\max_{u_1} u_1^\top S u_1 \quad \text{s.t.} \quad u_1^\top u_1 = 1$$

This is a constrained optimization problem – let's introduce Lagrange multipliers for constraint:

$$\max_{u_1, \lambda_1} \left\{ \underbrace{u_1^\top S u_1 + \lambda_1 (1 - u_1^\top u_1)}_{=:L[u_1, \lambda_1]} \right\}$$

We have

$$\begin{aligned} \frac{\partial L}{\partial u_1} : \; & 2Su_1 - 2\lambda_1 u_1 = 0 \\ & Su_1 = \lambda_1 u_1 \end{aligned}$$

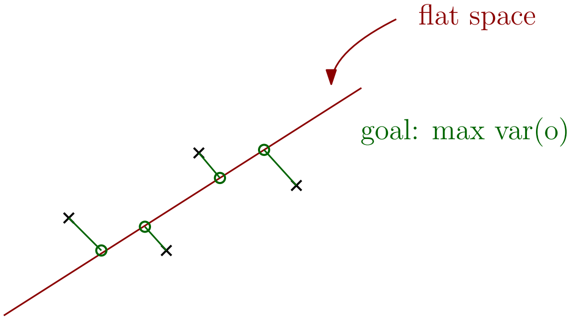
So, the eigen-problem: (λ_1, u_1) is eigenpair of S.

$$\text{var} = u_1^\top S u_1 = u_1^\top (\lambda_1 u_1) = \lambda_1 u_1^\top u_1 = \lambda_1$$

$\implies$ we need to pick the dominant eigenpair of S. So if we want to project onto a flat with $M > 1$, we can simply pick $u_1, \ldots, u_n$ as the M leading eigenvectors of S where all u_i are orthogonal and

$$\text{var} = \sum_{i=1}^{N} \lambda_i$$

Minimum Error Formulation:



Goal: introduce as little distortion as possible.
Consider: $\{u_i\}, i = 1, \ldots, D$ orthonormal basis of $\mathbb{R}^D$

$$\implies u_i^\top u_j = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & \text{otherwise} \end{cases}$$

Then each data point x_n has unique expansion in that basis

$$x_n = \sum_{i=1}^{D} \alpha_{ni} u_i \qquad \alpha_{ni} \in \mathbb{R}$$

where

$$\begin{aligned} x_n^\top u_j = u_j^\top x_n &= u_j^\top \sum_{i=1}^{D} \alpha_{ni} u_i \\ &= \sum_{i=1}^{D} \alpha_{ni} u_j^\top u_i = \alpha_{nj} \\ \implies x_n &= \sum_{i=1}^{D} \left(x_n^\top u_i\right) u_i \end{aligned}$$

As we project to a flat, we need only the first M terms

$$\tilde{x}_n = \sum_{i=1}^{M} z_{ni} u_i + \sum_{i=M+1}^{D} b_i u_i$$

Now, we choose z_{ni}, u_i, b_i so as to minimize the distortion.

$$J = \frac{1}{N} \sum_{n=1}^{N} \|x_n - \tilde{x}_n\|_2^2$$

The results we should've obtained are

1. $z_{ni} = x_n^\top u_i$, $i = 1, \ldots, M$
2. $b_i = \overline{x}^\top u_i$, $i = M+1, \ldots, D$

We can substitute these into the expression of $\tilde{x}_n$ as follow

$$\tilde{x}_n = \sum_{i=1}^{M} \left(x_n^\top u_i\right) u_i + \sum_{i=M+1}^{D} \left(\overline{x}^\top u_i\right) u_i$$

$$x_n - \tilde{x}_n = \sum_{i=M+1}^{D} \left(x_n^\top u_i - \overline{x}^\top u_i\right) u_i$$

In addition, the error term can be written as

$$J = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=M+1}^{D} \left(x_n^\top u_i - \overline{x}^\top u_i\right)^2 = \sum_{i=M+1}^{D} u_i^\top S u_i$$

So the problem now becomes

$$\min_{u_i, i=M+1,\ldots,D} \sum_{i=M+1}^{D} u_i^\top S u_i \quad \text{s.t.} \quad u_i^\top u_i = 1$$

Analogous to the case of maximum variance, we "throw away" the weakest eigenpairs of S.

§3.2 High-Dimensional PCA

Assume we have N data points with D dimensions and $\overline{x} = 0$. Then, $S = \frac{1}{N} x^\top x$

$$X = \begin{bmatrix} \rule{1.5cm}{0.4pt} \\ \rule{1.5cm}{0.4pt} \\ \rule{1.5cm}{0.4pt} \\ \\ \end{bmatrix}$$

where each x_n is a row of X. As $\overline{x} = 0$, rows sum up to 0.
Let's examine the eigenvalues of $x^\top x$ v.s. eigenvalues of $xx^\top$.

$$\frac{1}{N} x^\top x u_i = \lambda_i u_i$$

$$\frac{1}{N} x x^\top (x u_i) = \lambda_i (\underbrace{x u_i}_{v_i})$$

$$\frac{1}{N} x x^\top v_i = \lambda_i v_i$$

§3.3 Probabilistic PCA

Consider $x_n \in \mathbb{R}^D$ where

$$x_n = Wz + \mu + \varepsilon$$

where $z \in \mathbb{R}^M$ is latent variable and μ is mean and ε is noise & $\varepsilon \sim N(0, \sigma^2 I)$; z is the coordinates within the lower-dim flat, and W is the basis of the flat. The probabilistic formulation is

$$p(z) = N\left(z | 0, I\right)$$

$\Longrightarrow$ latent variable $\sim$ zero-mean, unit variance Gaussian. The conditional distribution $x|z$ is again Gaussian

$$p(x|z) = N\left(x| \underbrace{Wz + \mu}_{\text{nozzle location}}, \underbrace{\sigma^2 I}_{\text{spray size}}\right)$$

Resulting point cloud is governed by predictive density $p(x)$.

$$p(x) = \int \underbrace{p(x|z) \cdot p(z)}_{p(x,z)} \, dz$$

Claim 3.1. $p(x)$ is Gaussian, too.

$$\begin{aligned} p(x) &= N\left(x|\mu, C\right) \\ C &= WW^\top + \sigma^2 I \in \mathbb{R}^{D \times D} \end{aligned}$$

Proof. Sufficient statistics

$$\begin{aligned} \mathbb{E} &= \mathbb{E}\left[Wz + \mu + \varepsilon\right] \\ &= \mathbb{E}[Wz] + \mu + \mathbb{E}[\varepsilon] \\ &= W\mathbb{E}[z] + \mu = \mu \end{aligned}$$

For the covariance,

$$\begin{aligned} \operatorname{cov}[x] &= \mathbb{E}\left[(x-\mu)(x-\mu)^\top\right] \\ &= \mathbb{E}\left[\left(Wz + \mu + \varepsilon - \mu\right)\left(Wz + \mu + \varepsilon - \mu\right)^\top\right] \\ &= \mathbb{E}\left[(Wz+\varepsilon)(Wz+\varepsilon)^\top\right] \\ &= \mathbb{E}\left[\left(Wz(Wz)^\top\right) + Wz\varepsilon^\top + \varepsilon(Wz)^\top + \varepsilon\varepsilon^\top\right] \\ &= \mathbb{E}\left[Wzz^\top W^\top\right] + \mathbb{E}\left[Wz\varepsilon^\top\right] + \mathbb{E}\left[\varepsilon z^\top W^\top\right] + \mathbb{E}\left[\varepsilon\varepsilon^\top\right] \\ &= W\mathbb{E}\left[zz^\top\right]W^\top + \cancel{W\mathbb{E}\left[z\varepsilon^\top\right]} + \cancel{\mathbb{E}\left[\varepsilon z^\top\right]W^\top} + \mathbb{E}\left[\varepsilon\varepsilon^\top\right] \\ &= WW^\top + \sigma^2 I \end{aligned}$$

Remark 3.1. $\mathbb{E}\left[z\varepsilon^\top\right] = 0 = \mathbb{E}\left[\varepsilon z^\top\right]$ because z is independent from ε.

□

Note: Redundancy w.r.t. rotations in latent space (lack of uniqueness). Let $\tilde{W} = WQ$ where Q is orthonormal.

$$\begin{aligned} C &= \tilde{W}\tilde{W}^\top + \sigma^2 I \\ &= W\underbrace{QQ^\top}_{I}W^\top + \sigma^2 I \\ &= WW^\top + \sigma^2 I \end{aligned}$$

To evaluate $p(x) = N\left(x|\mu, C\right)$. We need C^{-1}.

$$C^{-1} = \sigma^{-2}I - \sigma^2 WM^{-1}W^\top$$

for $M = W^\top W + \sigma^2 I \in \mathbb{R}^{M \times M}$.

§3.4 Maximum Likelihood PCA

We need to learn W, μ, σ^2 from given data. By i.i.d,

$$\begin{aligned} p\left(X|W,\mu,\sigma^2\right) &= \prod_{n=1}^{N} p\left(x_n|W,\mu,\sigma^2\right) \\ \implies \ln p\left(X|W,\mu,\sigma^2\right) &= \sum_{n=1}^{N} \ln N\left(x_n|\mu, WW^\top + \sigma^2 I\right) \\ &= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|C| - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^\top C^{-1}(x_n-\mu) \end{aligned}$$

where $C = WW^\top + \sigma^2 I$; $\frac{d}{d\mu} = 0 \to \mu = \overline{x}$.
W, σ^2 are more tricky but again

refer to Bishop's paper

$$W = \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix}$$

where u_i are leading eigenvectors of S.