SPR Day 5

Univariate Gaussians

Readings: Bishop PRML Sec. 1.2.4
Gaussian r.v.
ML estimators for Gaussian

Bishop PRML Sec. 2.3.1-2.3.2
Gaussian properties
Conditionals and marginals

Outline:

(1) Univariate Gaussian distribution
PDF and parameters

(2) ML estimation of Gaussian parameters

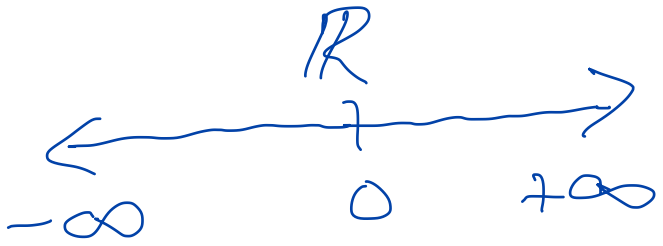
(3) Biased vs. unbiased estimators

(4) Useful properties:
- Linear transform of a G. is G.
- Sum of iid G. is G.
- Product of G. is NOT G. but PDF product is a Gaussian PDF

(5) Covariance

Univariate Gaussian

Random variable X

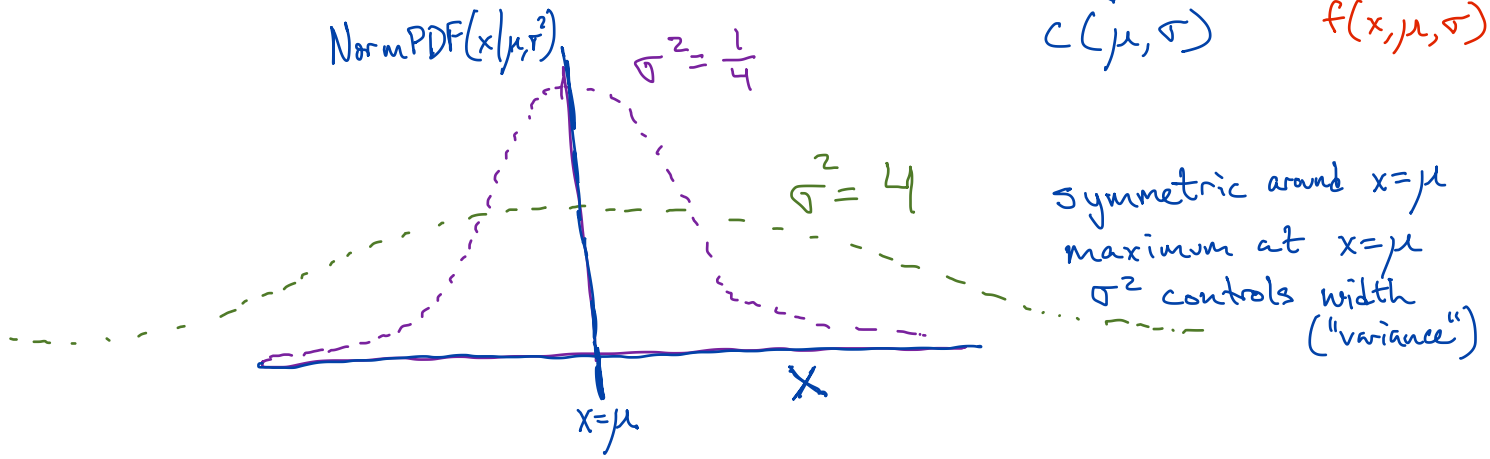
Sample space 

Parameters "mean" $\mu \in \mathbb{R}$

"variance" $\sigma^2 > 0$

PDF

$$\text{NormPDF}(x \mid \mu, \sigma^2) = \underbrace{\frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma}}_{c(\mu,\sigma)} \underbrace{e^{-\frac{1}{2}\frac{1}{\sigma^2}(x-\mu)^2}}_{f(x,\mu,\sigma)}$$



symmetric around $x=\mu$
maximum at $x=\mu$
σ^2 controls width ("variance")

Note the identity: $\int f(x,\mu,\sigma)\,dx = (2\pi)^{1/2}\sigma = \frac{1}{c(\mu,\sigma)}$ *

Computing the mean

Expected value of X (aka mean of r.v. X) is:

$$E[X] = \int_{-\infty}^{\infty} x \cdot \text{NormPDF}(x \mid \mu, \sigma^2)\, dx \qquad \text{by defn of expectation}$$

$$= \int_{-\infty}^{\infty} x \cdot \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{\sigma} e^{-\frac{1}{2}\frac{1}{\sigma^2}(x-\mu)^2} dx \qquad \text{by defn of PDF}$$

Change of variables: $t = \frac{x-\mu}{\sigma} \iff x = \sigma t + \mu$

$dt = \frac{1}{\sigma} dx \iff dx = \sigma\, dt$

$$= \frac{1}{(2\pi)^{1/2}} \int_{t=\frac{-\infty-\mu}{\sigma}=-\infty}^{+\infty} (\sigma t + \mu) e^{-\frac{1}{2}t^2} dt$$

by change of vars from calculus: $dt = \frac{1}{\sigma} dx$

$$= \frac{1}{(2\pi)^{1/2}} \cdot \left[\underbrace{\int_{-\infty}^{\infty} \sigma t e^{-\frac{1}{2}t^2} dt}_{\text{(A)}} + \underbrace{\int_{-\infty}^{\infty} \mu e^{-\frac{1}{2}t^2} dt}_{\text{(B)}} \right]$$

(A) is an odd function

$f(t) = t e^{-\frac{1}{2}t^2}$

$f(-t) = -t e^{-\frac{1}{2}(-t)^2} = -1 \cdot f(t)$

Integral of any odd function is zero because "areas" cancel

(B) uses identity * on previous page: $\int e^{-\frac{1}{2}t^2} dt = c(\mu=0, \sigma^2=1) = (2\pi)^{1/2}$

$$= \frac{1}{(2\pi)^{1/2}} \left[\underbrace{\sigma \cdot 0}_{\text{(A)}} + \mu \underbrace{(2\pi)^{1/2}}_{\text{(B)}} \right]$$

$$= \mu$$

Punchline: $E[X] = \mu$

Computing the variance

First, compute expected value of square

$$E[X^2] = \int_{-\infty}^{\infty} x^2 \, \text{NormPDF}(x \mid \mu, \sigma^2) \, dx$$

$$= \mu^2 + \sigma^2$$

after similar math as used on previous page for $E[X]$

Then, apply identity

$$\text{Var}[X] = E[X^2] - E[X]^2$$

$$= \mu^2 + \sigma^2 - \mu^2$$

$$= \sigma^2$$

substitute
$E[X] = \mu$
$E[X^2] = \mu^2 + \sigma^2$

Punchline:

$$\text{Var}[X] = E[(X-\mu)^2] = \sigma^2$$

ML Estimation for Gaussian

Consider observing N data measurements

$x_1, \cdots x_N$, where each $x_n \in \mathbb{R}$

If we assume these are indep. and identically distributed

$$p(x_1 \cdots x_N) = \prod_{n=1}^{N} p(x_n | \mu, \sigma^2) = \prod_{n=1}^{N} \text{NormPDF}(x_n | \mu, \sigma^2)$$

we can try to estimate the (unknown) parameters μ and σ^2 by maximizing likelihood

$$\mu_{ML}, \sigma_{ML} = \underset{\substack{\mu \in \mathbb{R} \\ \sigma^2 > 0}}{\text{argmax}} \sum_{n=1}^{N} \log \text{NormPDF}(x_n | \mu, \sigma^2)$$

$$= \text{argmax} \sum_{n=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[-\frac{1}{2} \frac{1}{\sigma^2} (x - \mu)^2 \right] \right]$$

$$= \text{argmax} \underbrace{ -\frac{N}{2} \log[2\pi] - N \log \sigma - \frac{1}{2} \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 }_{\text{call this } \ell(\mu, \sigma)}$$

(sum of squared distances)

To solve, we find local maxima via std. calculus methods...

$$\ell(\mu, \sigma) = -\frac{N}{2} \log[2\pi] - N \log \sigma - \frac{1}{2} \frac{1}{\sigma^2} SSD(x, \mu)$$

(sum of squared distances)

Sometimes useful to expand quadratic:

$$SSD = \sum_{n=1}^{N} (x_n - \mu)^2$$

$$= \sum_{n=1}^{N} (x_n^2 - 2 x_n \mu + \mu^2)$$

Solving for ML of mean parameter

Compute partial derivative wrt μ

$$\frac{\partial}{\partial \mu} \mathcal{L}(\mu, \sigma) = \frac{\partial}{\partial \mu} \left[-\frac{1}{2} \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n^2 - 2 x_n \mu + \mu^2) \right]$$

only SSD term of $\mathcal{L}$ depends on μ. drop the rest.

$$= -\frac{1}{2} \frac{1}{\sigma^2} \left(-2 \sum_n x_n + 2 \sum_n \mu \right)$$

$\frac{\partial}{\partial \mu}[x\mu] = x$

$\frac{\partial}{\partial \mu}[\mu^2] = 2\mu$

$$= \frac{1}{\sigma^2} \left(\sum_n x_n - N\mu \right)$$

Set to zero and solve

$$0 = \frac{1}{\sigma^2} \left(\sum_n x_n - N\mu \right)$$

solve for μ!

$$\boxed{\mu = \frac{1}{N} \sum_{n=1}^{N} x_n}$$

Punchline: ML estimation for μ sets μ equal to empirical mean of N observed datapoints

Solving for ML of variance parameter

$$\frac{\partial}{\partial \sigma} \mathcal{L}(\mu, \sigma^2) = \frac{\partial}{\partial \sigma}\left[-N \log \sigma - \frac{1}{2}\frac{1}{\sigma^2} SSD(x,\mu) \right]$$

$$= -\frac{N}{\sigma} - \frac{1}{2} SSD(x,\mu) \frac{\partial}{\partial \sigma}\left[\sigma^{-2}\right]$$

by derivative rules of powers $\frac{\partial}{\partial \sigma} \sigma^a = a \sigma^{a-1}$

$$= -\frac{N}{\sigma} - \frac{1}{2} SSD \cdot (-2) \sigma^{-3}$$

$$= -N\sigma^{-1} + SSD \cdot \sigma^{-3}$$

Set to zero and solve

$$0 = -N\sigma^{-1} + SSD\,\sigma^{-3}$$

multiply both sides by σ^3

$$0 = -N\sigma^2 + SSD$$

Thus,

$$\hat{\sigma}^2_{ML} = \frac{SSD(x,\mu)}{N} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\mu}_{ML})^2$$

and

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

Biased vs Unbiased Estimators

μ_{ML} and σ_{ML} are <u>estimators</u> of unknown parameters given N observations.

Consider following experiment:

1) Pick <u>true</u> values μ_T, σ_T
2) Pick a value of N
3) Draw N samples $x_1, \dots x_N$ s.t.
$$x_n \sim \mathcal{N}(\mu_T, \sigma_T^2)$$
then compute $\hat{\mu}_{ML}, \hat{\sigma}_{ML}$ from these samples
4) Repeat (3) R times, with $R \to \infty$

Question: What is
$$\bar{\mu} = \frac{1}{R}\sum_{r=1}^{R} \hat{\mu}_{ML}^{r} = \mathop{\mathbb{E}}_{x \overset{iid}{\sim} \mathcal{N}(\mu_T, \sigma_T)}\left[\mu_{ML}(x)\right]$$
$$\bar{\sigma}^2 = \frac{1}{R}\sum_{r=1}^{R} \sigma_{ML}^{2(r)} = \mathop{\mathbb{E}}_{x \overset{iid}{\sim} \mathcal{N}(\mu_T, \sigma_T)}\left[\sigma_{ML}^2(x)\right]$$

Turns out, $\bar{\mu} = \mu_T$ so we say μ_{ML} is <u>unbiased</u>

but $\bar{\sigma}^2 \neq \sigma_T^2$ so we say σ_{ML} is <u>biased</u>

derivation that σ_{ML} is biased estimator for finite N

$$\mathbb{E}\left[\sigma_{ML}^2(x_1 \cdots x_N)\right] = \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\left(x_n^2 - 2\mu_{ML}x_n + \mu_{ML}^2\right)\right]$$

$x_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[x_n^2] - 2\,\mathbb{E}[\mu_{ML}(x_1 \cdots x_N)\cdot x_n] + \mathbb{E}[\mu_{ML}^2]\right)$$

$$= \frac{1}{N}\left[\sum_{n=1}^{N}\mathbb{E}[x_n^2] - 2N\,\mathbb{E}\left[\mu_{ML}(x_1 \cdots x_N)\cdot\frac{\left(\sum_n x_n\right)}{N}\right] + N\,\mathbb{E}[\mu_{ML}^2]\right]$$

$$= \frac{1}{N}\left[\sum_{n=1}^{N}\mathbb{E}[x_n^2] - 2N\,\mathbb{E}[\mu_{ML}^2] + N\,\mathbb{E}[\mu_{ML}^2]\right]$$

$$= \left[\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[x_n^2]\right] - \mathbb{E}[\mu_{ML}^2]$$

$$= \frac{N}{N}\mu^2 + \sigma^2 - \mathbb{E}\left[\frac{1}{N}(x_1 + x_2 + \cdots x_N)\frac{1}{N}(x_1 + \cdots x_N)\right]$$

$$- \frac{1}{N^2}\,\mathbb{E}\left[\sum_n x_n^2 + \sum_{i\neq j} x_i x_j\right]$$

$$= \mu^2 + \sigma^2 - \frac{1}{N^2}\sum_n(\mu^2 + \sigma^2) - \frac{1}{N^2}\sum_{i\neq j}\mathbb{E}[x_i x_j]$$

$$= \mu^2 + \sigma^2 - \underbrace{\frac{1}{N^2}\sum \mu^2 + \sigma^2}_{N \text{ terms}} - \underbrace{\frac{1}{N^2}\sum_{i\neq j}\mu^2}_{N^2 - N \text{ terms}}$$

$$= \mu^2 - \frac{N}{N^2}\mu^2 - \frac{N^2 - N}{N^2}\mu^2 + \sigma^2 - \frac{N}{N^2}\sigma^2$$

$$= 0\cdot\mu + \frac{N^2 - N}{N^2}\sigma^2 = \frac{N-1}{N}\sigma^2$$

Assuming $x_1 \cdots x_N \sim N(\mu_T, \sigma_T^2)$

Punchline: μ_{ML} is an unbiased estimator of the mean for any finite sample:

$$\mathbb{E}[\mu_{ML}(x_1 \cdots x_N)] = \mu_T$$

σ_{ML} is a biased estimator of the variance for any finite sample

$$\mathbb{E}[\sigma_{ML}^2(x_1 \cdots x_N)] = \frac{N-1}{N}\sigma_T^2$$

When N is small, this might matter a lot

As N gets larger, $\frac{N-1}{N} \to 1.0$ and bias gradually disappears

Another problem w/ ML estimators:

If N too small, variance will be nonsensical

e.g. at $N=1$, $\sigma_{ML}^2 = \frac{(x_1 - \mu_{ML})^2}{N} = \frac{(x_1 - x_1)^2}{N} = \frac{0}{N} = 0$ zero!

Useful Gaussian properties

Assume that $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$

and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$

then we have

(1) Deterministic Linear Transforms of Gaussians are also Gaussian

Let $Z = aX + b$, for $a \neq 0$, $b \in \mathbb{R}$

then $Z \sim \mathcal{N}(a\mu_X + b, a^2\sigma_X^2)$

(2) Sums of Gaussian r.v.s are also Gaussian

Let $S = X + Y$.

then $S \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$

(3) Products of Gaussian r.v. are NOT Gaussian.

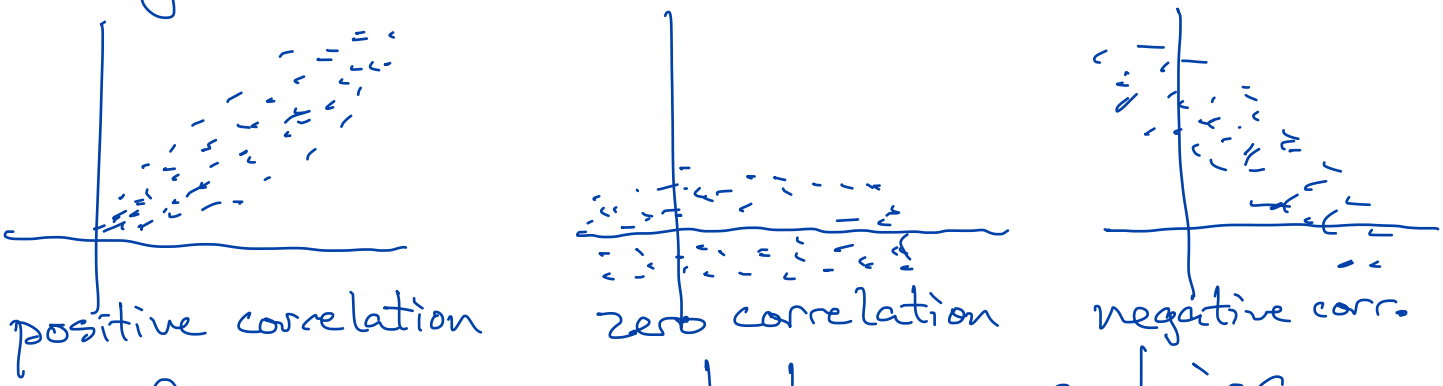
However, the product of Gaussian PDFs has Gaussian form.

$$\text{NormPDF}(x \mid \mu_x, \sigma_x^2)\,\text{NormPDF}(x \mid \mu_y, \sigma_y^2) \propto \text{NormPDF}(x \mid \mu, \sigma^2)$$

where $\mu = \frac{\sigma_x^2 \mu_x + \sigma_y^2 \mu_y}{\sigma_x^2 + \sigma_y^2}$ and $\sigma^2 = \frac{1}{\frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2}}$

Covariance

For any multivariate random variable $X \in \mathbb{R}^D$



We define covariance between entries i and j of vector X as:

$$\text{Cov}[x_i, x_j] = \mathbb{E}_{x \sim p(X)}\left[(x_i - \mathbb{E}[x_i])(x_j - \mathbb{E}[x_j]) \right]$$

in general, can be any real value (negative, zero, or positive)

Special case: $i = j$. Then $\text{Cov}[x_i, x_i] = \text{Var}[x_i] \geq 0$

Correlation is defined as:

$$\text{Corr}[x_i, x_j] = \frac{\text{Cov}[x_i, x_j]}{\sqrt{\text{Var}[x_i] \cdot \text{Var}[x_j]}}$$

always between -1 and 1

$$-1 \leq \text{corr} \leq 1$$

Covariance Matrix

For any random variable X of size D.

Define covariance matrix as

$$\Sigma = \text{Cov}(X) = \begin{bmatrix} \text{Var}[x_1] & \text{Cov}[x_1, x_2] & \cdots & \text{Cov}[x_1, x_D] \\ \text{Cov}[x_2, x_1] & \text{Var}[x_2] & \cdots & \text{Cov}[x_2, x_D] \\ \text{Cov}[x_3, x_1] & \text{Cov}[x_3, x_2] & \ddots & \vdots \\ \text{Cov}[x_D, x_1] & \text{Cov}[x_D, x_2] & \cdots & \text{Var}[x_D] \end{bmatrix}$$

$D \times D$ matrix

Will always be symmetric and positive-semidefinite

$\Sigma_{ij} = \Sigma_{ji}$ for all i, j

$a^T \Sigma a \geq 0$ for all $a \in \mathbb{R}^D$

Sometimes positive-definite: $a^T \Sigma a > 0$ for all $a \in \mathbb{R}^D$

(when $\text{Var}[x_i] > 0$ for every i and no variable is linear combo of others)

Lin Alg review

Σ is positive definite implies:

- matrix Σ is invertible
- all columns linearly independent
- all eigenvalues are positive