LECTURE NOTES ON MULTIVARIATE PROBABILITY [*]

A. A. ERGÜR, A. MOMIN

Abstract. This scribe contains lecture notes on multivariate probability, covering fundamental concepts including covariance matrices, multivariate Gaussian distributions, Principal Component Analysis (PCA), and the Expectation-Maximization (EM) algorithm. These topics are the core concept and the base for probabilistic modeling in ML and data science.

Key words. multivariate probability, covariance, Gaussian distribution, PCA, EM algorithm

[*] We thank Robbins family for supporting the Algorithmic Foundations of Data Science Course

1. Introduction. In this scribe, I will be discusing about some very basic ideas in multivariate probability, covering some basic statistics and distributions often used in ML and data analysis. I will discuss covariance and the covariance matrix to understand how a group of variables relate to one another. I will also cover the multivariate Gaussian distribution, which is a core component of probabilistic modelling. I will also discuss properties of the multivariate Gaussian distribution including conditional normal distributions and linear transformations. In addition I will discuss Principal Component Analysis (PCA), a popular example of dimensionality reduction, as well as the Expectation Maximization (EM) algorithm that is used to estimate parameters of models with unobserved variables. These concepts are important in understanding how models are used to understand complex data and to infer from the real-world data they are working with.

2. Covariance. Covariance is a measure in statistics that indicates how two random variables vary together. If two variables tend to increase together we say that their covariance is positive. If we see a variable increase and the other variable decrease than we say that it is a negative covariance. The covariance is defined as follows w.r.t two random variables X and Y.

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Where:

- $E[X]$ and $E[Y]$ are the expected values (means) of X and Y.
- The covariance represent the direction of the linear relationship b/w the variables.

In multivariate statistics, the covariance among many variables is captured in the covariance matrix. The covariance matrix Σ is symmetric for a set of random variables $\mathbf{X} = [X_1, X_2, \ldots, X_n]^T$ where:

- The diagonal entries represent the variances of each of the individual variables.
- The off-diagonal entries represent the covariances b/w pairs of the different variables.

For example, when we have a two-dimensional dataset, the covariance matrix may look like:

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{bmatrix}$$

The covariance matrix is very important to understand how multiple variables are connected in a multivariate distribution. Following is the image which shows correlation coefficient:

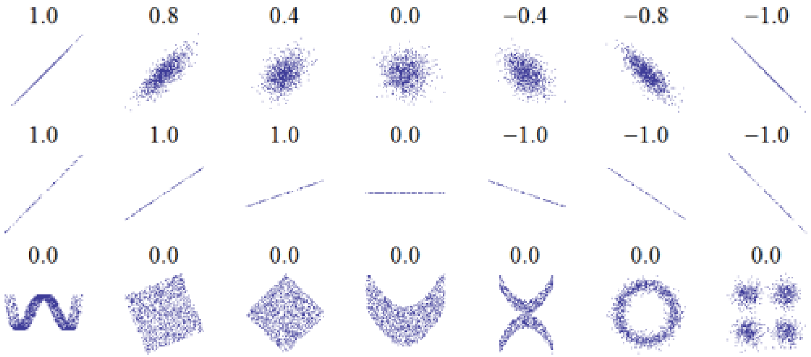


Fig. 1. *Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). (Note: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.) From [2].*

2.1. Univariate and Multivariate Covariance.

2.1.1. Univariate Case:. In this case, we can compute the covariance b/w 2 variables in such a way:

$$\mathrm{Cov}(x, y) = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

The sum of the variance for 2 random variable, let say x and y, would be related to individual variables variance and their covariance which can be shown as: and its matrix is shown as:

$$\mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(y) + 2\,\mathrm{Cov}(x,y)$$

2.1.2. Multivariate Case:. In this case, we have multiple variables unless like univariate. Let say we have 2 vectors $X = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ so convariance matrix of these 2 vectors would be:

$$\mathrm{Cov}(x,y) := \mathbb{E}[xy^T] - \mathbb{E}[x]\mathbb{E}[y^T]$$

and their matrix would look like:

$$\mathrm{Cov}(x,y) := \begin{bmatrix} \mathrm{Cov}(x_1,y_1) & \mathrm{Cov}(x_1,y_2) & \cdots & \mathrm{Cov}(x_1,y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(x_{n-1},y_1) & \cdots & \cdots & \mathrm{Cov}(x_{n-1},y_n) \end{bmatrix}$$

This is a real symmetric matrix because the covariance b/w x_i and y_i is similar to the covariance b/w y_j and x_j

3. Variance and Covariance Matrix. Variance is simply the case of covariance where a variable is conected with itself. It calculates how much a variable deviates from its mean:

$$Var(X) = Cov(X,Y)$$

The covariance matrix $\sum$ is important in multivariate distributions since it tells the dispersion and correlation characteristics of the data, and some of properties of the covariance matrix include:

- One of the property is that it is symmetric i.e: $\sum = \sum^T$
- It is positive semi-definite (PSD), which means that all of its eigenvalues are non-negative. This guarantees that the matrix is valid to be used for statistical models.

The covariance matrix is then used to quantify relationships b/w many variables in a dataset. This matrix can be used in many ways, including PCA, Gaussian models, and also in modelling relationships b/w features.

3.1. Properties of Covariance Matrix:. Now we will discuss about the random vector covariance matrix which is defined as:

$$\mathrm{Cov}(X) := \mathrm{Cov}(X,X) = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

and the matrix is represented as:

$$\mathrm{Cov}(X) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \ddots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \cdots & \cdots & \mathrm{Var}(X_n) \end{bmatrix}$$

When a matrix $A = BB^T$ it holds that: $\lambda \geq 0$, where λ is the eigenvalues of the matrix. This means that our matrix is PSD which make sure that all of the eigenvalue is non-negative and this property further can be useful in statistical model.

4. Multivariate Gaussian Distribution. In probability theory and statistics, the multivariate normal distribution, multivariate Gaussian distribution, or joint normal distribution is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions [1].

A multivariate Gaussian distribution for a random vector $\mathbf{x} \in R^n$ is defined by mean vector $\mu \in R^n$ and a covariance matrix $\Sigma \in R^{n\times n}$. The PDF of a multivariate Gaussian is:

$$P(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right)$$

over here:

- Mean vector of the distribution is μ.
- Covariance matrix is Σ
- The inverse of the covariance matrix is Σ^{-1}, and $|\Sigma|$ is the determinant of Σ.
- n is the dimensionality of the vector $\mathbf{x}$.

4.1. Independent Coordinates.

For the multivariate Gaussian when coordinates are independent,let say, for $X = (x_1, \ldots, x_n) \sim N(\mu, I)$:

- Mean: $\mathbb{E}[X] = \mu$
- Covariance: $\text{Cov}(X) = I$ (Identity matrix)

The density of this distribution is given by:

$$P_X(t) = \prod_{i=1}^{n} P_{X_i}(t_i) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-||t-\mu||_2^2/2}$$

Here, the components of X are independent, and the covariance matrix is the identity matrix I, which means that each variable has variance 1 and is uncorrelated with the others.

Following are the properties:

- Marginal Distributions: Any subset of the variables in a multivariate Gaussian is also Gaussian.
- Conditional Distributions: The conditional distribution of one variable given others is also Gaussian.
- Geometric Interpretation: The multivariate Gaussian can be visualized as a hyper-ellipsoid in the feature space, with the axes aligned along the eigenvectors of the covariance matrix.

Following figure show the 2D gaussian density (Surface Plot): Following figure depicts 2D Gaussian Density

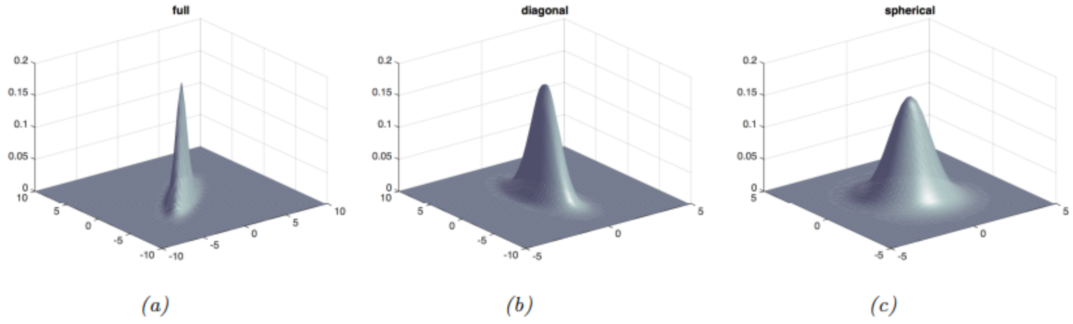


Fig. 2. *Visualization of a 2d Gaussian density as a surface plot. (a) Distribution using a full covariance matrix can be oriented at any angle. (b) Distribution using a diagonal covariance matrix must be parallel to the axis. (c) Distribution using a spherical covariance matrix must have a symmetric shape. Generated by [3]*

(Level Sets):

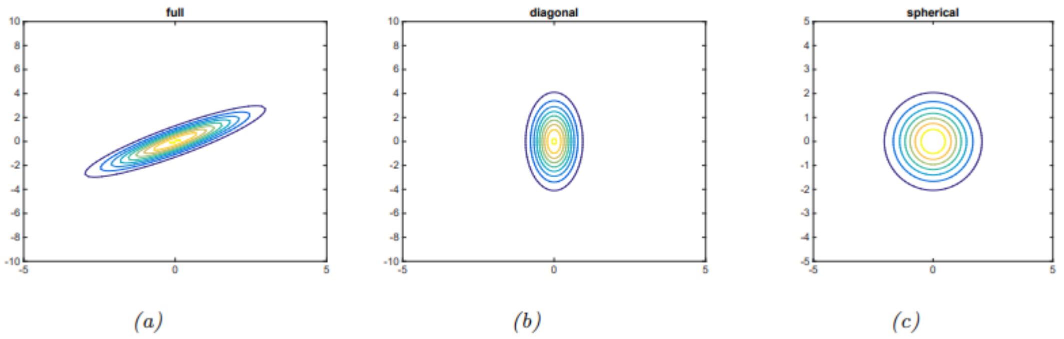


Fig. 3. *Visualization of a 2d Gaussian density in terms of level sets of constant probability density. (a) A full covariance matrix has elliptical contours. (b) A diagonal covariance matrix is an axis aligned ellipse. (c) A spherical covariance matrix has a circular shape. Generated by [3]*

5. Marginal and Conditional Distributions. For a jointly Gaussian random vector $y = (y_1, y_2)$, with mean vector and covariance matrix defined as:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \quad \Lambda = \Sigma^{-1} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A22 \end{pmatrix}$$

over here A is the precision matrix.
Marginals: The marginal distributions for y_1 and y_2 are given by:

$$p(y_1) = N(y_1|\mu_1, \Sigma_{11})$$

$$p(y_2) = N(y_2|\mu_2, \Sigma_{22})$$

Conditional: The conditional distribution of y_1 given y_2 is also Gaussian:

$$p(y_1|y_2) = N(y_1|\mu_{1|2}, \Sigma_{1|2})$$

Where:

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = A_{11}^{-1}$$

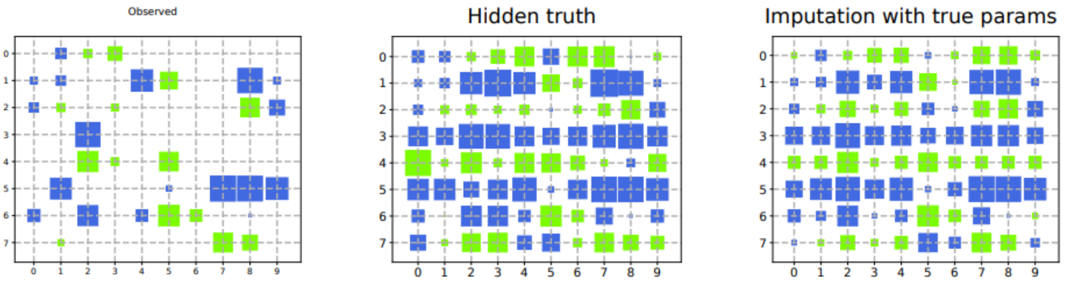
5.1. General Multivariate Gaussian. For a random vector $y \sim N(\tilde{\mu}, \Sigma)$:

$$y = AX + \tilde{\mu}$$

where $X \sim N(0, I)$, $A \in \mathbb{R}^{m \times n}$, and $\Sigma = AA^T$.
The density of this distribution is given by:

$$p(y) = \frac{1}{(\sqrt{2\pi})^m \sqrt{\det \Sigma}} e^{-\frac{1}{2}(y-\tilde{\mu})^T \Sigma^{-1}(y-\tilde{\mu})}$$

5.2. Imputing missing values. Missing value imputation uses observed dimensions and Gaussian correlations to infer unobserved entries, computing posterior means as optimal estimates. Following is the visualization representing data imputation using an MVN.



(a) (b) (c)

FIG. 4. *Illustration of data imputation using an MVN. Rows are features, columns are data samples (the transpose of the convention used in the text). (a) Visualization of the data matrix. Blank entries are missing (not observed). Blue are positive, green are negative. Area of the square is proportional to the value. (This is known as a Hinton diagram, named after Geoff Hinton, a famous ML researcher.) (b) True data matrix (hidden). (c) Mean of the posterior predictive distribution, based on partially observed data for that example (column), using the true model parameters. Generated by* [7]

6. PCA (Principal Component Analysis) and Linear Transformations. Principal component analysis (PCA) is a method for reducing the dimensionality of multivariate data with maximum retention of variance. PCA idea is to convert the data into a new coordinate system where the axes correspond to the directions of max variance.
Following are the steps to perform PCA:

- Calculate the data's covariance matrix.
- Calculate the covariance matrix's eigenvalues and eigenvectors.

- The eigenvectors must be sorted based on eigenvalue size in descending order since the larger the eigenvalue, the larger percent variance explained by that eigenvector.
- Lastly, projecting our data onto the eigenvectors (the principal components also known as axes) corresponding to the largest eigenvalues.

In multivariate Gaussian distributions, PCA is equivalent to maximising the variance of the data by finding the linear transformation of the data. The transformed variables (the principal components) will be uncorrelated.

Linear transformations of multivariate Gaussian data are a new multivariate Gaussian distribution with transformed mean and covariance:

$$\text{New Mean} = A\mu$$

$$\text{New Covariance} = A\Sigma A^T$$

Over here A is the transformation matrix. Following equation tell us about the relationship b/w data matrix and its covariance matrix along with SVD.

Data Matrix:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \quad (n \times N)$$

Covariance Matrix:

$$XX^T = \text{Covariance Matrix (real, symmetric, PSD)}$$

SVD

$$X^T X = U\Sigma U^T$$

$$XX^T = \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \operatorname{diag}(b_1, b_2, \ldots, b_n) \begin{bmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_n^T \end{bmatrix}$$

over here, the direction preseve the most of the variance where direction is $u_1, u_2, \ldots, u_d$.

6.1. Probabilistic PCA (PPCA).

Model:

- Latent variable $z \sim N(0, I_d)$ (Standard Gaussian distribution in the latent space).
- Observed data: $x|z \sim N(Wz + \mu, \sigma^2 I)$, where W is the matrix of weights, and μ is the mean of the observed data.
- Marginal distribution: $x \sim N(\mu, WW^T + \sigma^2 I)$, where the observed data follows a Gaussian distribution with a mean μ and covariance $WW^T + \sigma^2 I$.

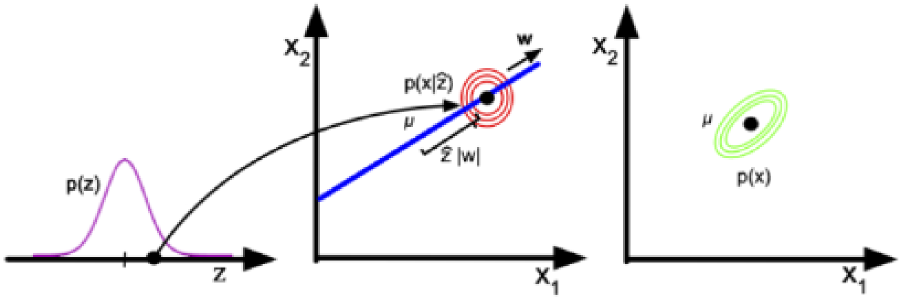


Fig. 5. *PPCA*

7. Maximum Likelihood Estimation (MLE).

Gaussian MLE:
For samples $x_1, \ldots, x_n \sim N(\mu, \sigma)$, the maximum likelihood estimators for the mean μ and variance σ^2 are given by:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Bias Correction: To correct for bias in the variance estimator, we use:

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \implies \text{Use } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

8. KL Divergence and MLE.

KL Divergence:
The Kullback-Leibler (KL) divergence between two probability distributions p and q is defined as:

$$KL(p||q) = H_{ce}(p, q) - H(p)$$

Where $H_{ce}(p, q)$ is the cross-entropy and $H(p)$ is the entropy of p.

MLE as KL Minimization:
Maximizing the likelihood is equal to minimizing the KL divergence b/w the data distribution P_D and model distribution $P_{y|\theta}$:

$$\arg\max_{\theta} \prod_{i=1}^{N} P_{y|\theta}(x_i)$$

This means that maximum likelihood estimation can be viewed as finding the parameter values that minimize the divergence b/w observed data distribution and model.

9. Conditional Distributions. It describes the behavior of a random variables subset, given the values of other variables. For example in multivariate Gaussians, the one subset of variable in conditional distribution given other is also a Gaussian distribution.. Lets say we have a joint distribution X $= [X_1, X_2]^T$ then the conditional distribution would be:

$$P(X_1|X_2) \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

over here:

- μ_1, μ_2 are the means of X_1 and X_2.
- $\Sigma_{11}, \Sigma_{12}, \Sigma_{21}, \Sigma_{22}$ are blocks of the covariance matrix

Its an important property in probabilistic graphical models and in Bayesian inference. Over here they modeled the conditional dependencies b/w random variables.

10. Expectation and Covariance of Linear Functions. The expectation and covariance are derived for a linear function of random variable $Y = AX$ as:

- Expectation: $E[Y] = AE[\mathbf{X}]$.
- Covariance: $\text{Cov}(Y) = A\,\text{Cov}(\mathbf{X})A^T$.

These results are useful for regression models and for thinking about how the data will be affected by linear transformations. For example, linear regression has coefficients for its linear model that can be estimated through the properties listed.

11. Multivariate Gaussian as a Building Block. The multivariate Gaussian is the core component in many probabilistic ML models. It is used in Bayesian models, generative models, and in Hidden Markov Models. In Gaussian Mixture Models, each component is a multivariate Gaussian, and the model is used for tasks like clustering.

12. Bayesian Inference and Linear Gaussian Systems.

Bayesian inference in a Linear Gaussian System models the relationship b/w observed data (y) and latent variables (z) using Gaussian distributions:

Linear Gaussian System:

$$p(z) = N(z|\mu_z, \Sigma_z)$$

$$p(y|z) = N(y|Wz + b, \Sigma_y)$$

Joint Distribution:

It represents the relationship b/w the latent variable z and the observed data y.

$$\mu = \begin{pmatrix} \mu_z \\ W\mu_z + b \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_z & \Sigma_z W^T \\ W\Sigma_z & \Sigma_y + W\Sigma_z W^T \end{pmatrix}$$

Posterior:

Lastly, posterior mean and covariance are computed to estimate the latent variables.

$$p(z|y) = N(z|\mu_{z|y}, \Sigma_{z|y})$$

$$\Sigma_{z|y}^{-1} = \Sigma_z^{-1} + W^T \Sigma_y^{-1} W$$

$$\mu_{z|y} = \Sigma_{z|y} \left[W^T \Sigma_y^{-1} (y - b) + \Sigma_z^{-1} \mu_z \right]$$

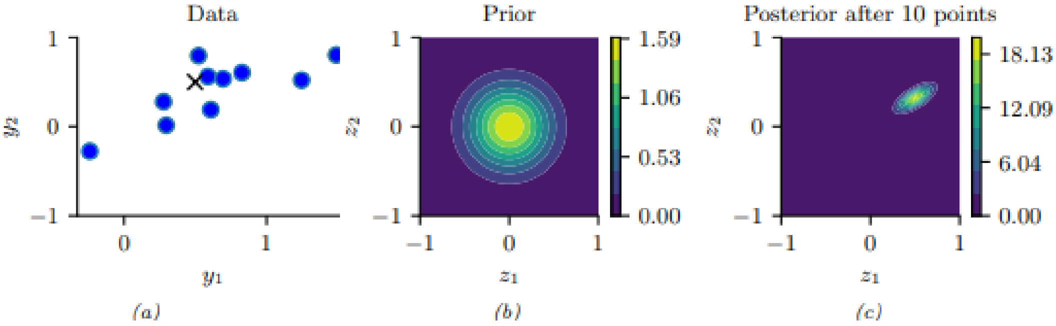
12.1. Inferring an Unknown Vector.

For an unknown vector $z \sim N(\mu_z, \Sigma_z)$ with observed data $D : y_1, y_2, \dots, y_N \overset{\text{i.i.d.}}{\sim} N(z, \Sigma_y)$:

- Likelihood and Posterior:

$$p(D|z) = \prod_{i=1}^{N} \mathcal{N}(y_i|z, \Sigma_y)$$

$$p(z|y_1, \dots, y_N) = \mathcal{N}(z|\hat{\mu}, \hat{\Sigma})$$



(a) (b) (c)

FIG. 6. *Illustration of Bayesian inference for a 2D Gaussian random vector z. (a) The data is generated from* $y_n \sim \mathcal{N}(z, \Sigma_y)$*, where* $z = [0.5, 0.5]^T$ *and* $\Sigma_y = 0.1 \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$*. We assume the sensor noise covariance* Σ_y *is known but z is unknown. The black cross represents z. (b) The prior is* $p(z) = \mathcal{N}(z|0, 0.1I_2)$*. (c) Posterior after observing 10 data points. Generated by [6]*

13. Expectation Maximization (EM) Algorithm.

13.1. General EM Algorithm.

An iterative approach known as Expectation Maximization (EM) Algorithm is used for estimating the parameters in the model who have hidden values. It is consist of 2 steps:

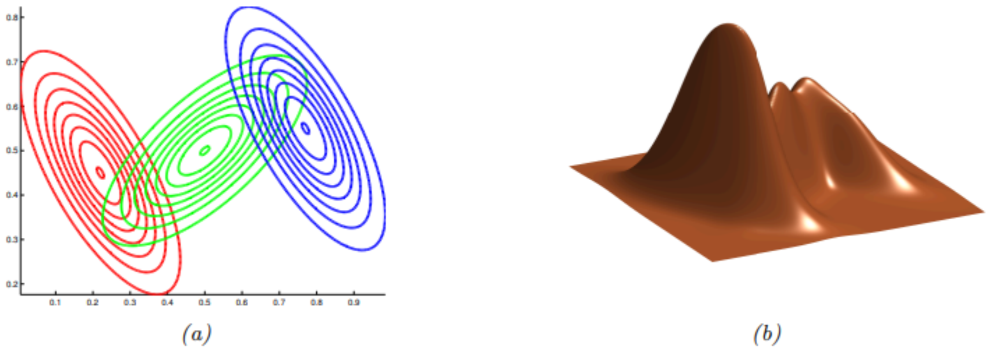
- Expectation (E-step): We compute the expected value for the hidden/latent variables based on the current estimate of the parameters.
- Maximization (M-step): Maximizing the likelihood function w.r.t the params, given the expected values of the latent variables from the E-step.

The EM algorithm is particularly useful when dealing with incomplete data or hidden variables, and is widely applied in Gaussian Mixture Models and Hidden Markov Models. EM algorithm is used when we are analyzing incomplete data or any hidden variables and its used in Gaussian Mixture Models and Hidden Markov Models. It is iteratively proceeds:

- E-step: Given the observed data, it estimate the distribution of the hidden variables .
- M-step: Using the current estimates of the hidden variables, we can update the model parameters by maximizing the likelihood.

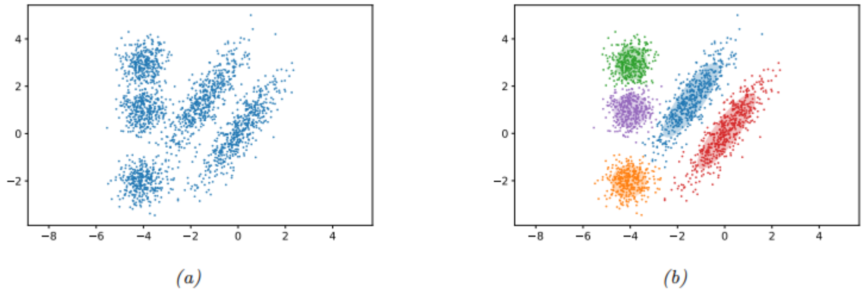
EM make sures that the likelihood does not decrease with each iteration and converges to a local maximum.

13.2. EM Algorithm for Gaussian Mixtures. The Expectation Maximization (EM) Algorithm is used to estimate the parameters of models with hidden or latent variables1, like Gaussian Mixture Models.



(a) (b)

Fig. 7. *A mixture of 3 Gaussians in 2d. (a) We show the contours of constant probability for each component in the mixture. (b) A surface plot of the overall density. Adapted from Figure 2.23 of [Bis06]. Generated by [4]*



(a) (b)

Fig. 8. *Figure 3.12: (a) Some data in 2d. (b) A possible clustering using $K = 5$ clusters computed using a GMM. Generated by [5]*

The algorithm start off iteratively, alternating b/w two steps i.e

- Initialize the parameters, including the means (μ_1, μ_2), variances (σ_1^2, σ_2^2), and the responsibilities γ.
- E-step: Compute the responsibilities $p_j(x_i)$, which represent the probability that data point x_i belongs to component j.
- M-step: Update the parameters of the model (e.g., the means and variances) using weighted sums based on the computed responsibilities from the E-step.
- Repeat these steps until convergence, i.e., until the parameters no longer change significantly.

The EM algorithm iteratively maximizes the likelihood function by alternating b/w these steps, updating the estimates for the hidden variables and model parameters. We follow the presentation in [8, 9]:

EM Algorithm for Mixture of Two Gaussian Distributions
Input: n samples $x_1, \dots, x_n$.
Output: ML-estimate $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \gamma)$.

1. Start with arbitrary values for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$, and with $\gamma = \frac{1}{2}$.
2. For $j = 1, 2$, let
$$f(x, \mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp^{\left(-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right)}$$
3. Repeat until convergence
(or no significant improvement in $L(\bar{x}, \gamma, \mu_1, \mu_2, \sigma_1, \sigma_2)$):
   (a) For $i = 1$ to n:
   - $p_1(x_i) = \frac{\gamma f(x_i, \mu_1, \sigma_1)}{\gamma f(x_i, \mu_1, \sigma_1) + (1-\gamma) f(x_i, \mu_2, \sigma_2)}$
   - $p_2(x_i) = 1 - p_1(x_i)$

   (b) For $j = 1, 2$:
   - $\mu_j = \frac{\sum_{i=1}^n p_j(x_i) x_i}{\sum_{i=1}^n p_j(x_i)}$
   - $\sigma_j^2 = \frac{\sum_{i=1}^n p_j(x_i)(x_i - \mu_j)^2}{\sum_{i=1}^n p_j(x_i)}$

   (c) $\gamma = \frac{1}{n} \sum_{i=1}^n p_1(x_i)$

14. Conclusion.

The topics covered in this section cover the foundation for modeling using probability in ML. Covariance, multivariate Gaussian distributions, Principal Components Analysis (PCA), and understanding conditional distributions, all of us allow us to model complicated systems in which variables are interrelated. Multiple approaches like Expectation Maximization (EM) and linear transformations can help us in handling the incomplete data and lower dimensionality. It is important to understand these topics if we want to create and understand usable and valid ML models.

15. Exercise.

1. Complete the proof we started in class but didn't fully work out:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

What is the implication of this identity when X and Y are independent random variables?

Answer:

We can start off by saying that variance is:

$$\text{Var}(X + Y) = \mathbb{E}[(X + Y)^2] - (\mathbb{E}[X + Y])^2$$

By expanding the above equation:

$$= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2$$

Then we will be distributing the expectations i.e:

$$= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2$$

And by recognizing variance and covariance, we get:

$$= (\mathbb{E}[X^2] - \mathbb{E}[X]^2) + (\mathbb{E}[Y^2] - \mathbb{E}[Y]^2) + 2(\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y])$$

$$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$$

as we know that when X and Y are independent, Cov(X,Y) = 0, therefore:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

2. Recall Chebyshev's inequality:

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq \frac{\mathrm{Var}(X)}{t^2}$$

Answer:
Using Chebyshev's inequality, prove the law of large numbers. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ where X_i are i.i.d. with mean μ and variance σ^2. By calculating the expectation and variance of sample mean, we get:

$$\mathbb{E}[\bar{X}_n] = \mu$$

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

By applying Chebyshev's inequality to $\bar{X}_n$:

$$P(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Now we will take limit as n→ inf:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

This proves the convergence in probability.

3. Let X_1, X_2 be two independent Gaussian random variables with mean zero and variances 4 and 9, respectively. Define:

$$X := (X_1, X_2, X_1 + X_2, X_1 - X_2)$$

Compute the covariance matrix of X.

Answer:
Given that:

- $X_1 \sim N(0, 4)$
- $X_2 \sim N(0, 9)$
- Independent X_1, X_2

Calculating all pairwise covariances:

- Diagonal entries (variances):

$$\mathrm{Var}(X_1) = 4$$

$$\mathrm{Var}(X_2) = 9$$

$$\mathrm{Var}(X_1 + X_2) = 4 + 9 = 13$$

$$\mathrm{Var}(X_1 - X_2) = 4 + 9 = 13$$

- Off-diagonal entries:

$$\mathrm{Cov}(X_1, X_2) = 0$$

(independent)

$$\mathrm{Cov}(X_1, X_1 + X_2) = \mathrm{Var}(X_1) = 4$$

$$\mathrm{Cov}(X_1, X_1 - X_2) = \mathrm{Var}(X_1) = 4$$

$$\mathrm{Cov}(X_2, X_1 + X_2) = \mathrm{Var}(X_2) = 9$$

$$\mathrm{Cov}(X_2, X_1 - X_2) = -\mathrm{Var}(X_2) = -9$$

$$\mathrm{Cov}(X_1 + X_2, X_1 - X_2) = \mathrm{Var}(X_1) - \mathrm{Var}(X_2) = -5$$

Based on the above calcuklation, our final covariance matrix is:

$$\begin{pmatrix} 4 & 0 & 4 & 4 \\ 0 & 9 & 9 & -9 \\ 4 & 9 & 13 & -5 \\ 4 & -9 & -5 & 13 \end{pmatrix}$$

4. Let X, Y be two random variables, and let r be the correlation coefficient:

$$r = \frac{\operatorname{Cov}(X,Y)}{\sigma(X)\sigma(Y)}$$

Show that $-1 \leq r \leq 1$.

Answer:

Consider the variance of standardized variables which is: Let $U = \frac{X-\mu_X}{\sigma_X}$, $V = \frac{Y-\mu_Y}{\sigma_Y}$ By calculating the variance of U ± V:

$$\operatorname{Var}(U \pm V) = \operatorname{Var}(U) + \operatorname{Var}(V) \pm 2\operatorname{Cov}(U,V) = 2(1 \pm r)$$

Since variance must be non-negative therefore:

$$1 + r \geq 0 \Rightarrow r \geq -1$$

$$1 - r \geq 0 \Rightarrow r \leq 1$$

5. Let $X_1, X_2, \ldots, X_n$ be independent samples from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. The empirical mean estimate is:

$$\tilde{\mu} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

The empirical variance estimate is:

$$S_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \tilde{\mu})^2$$

Note that $\tilde{\mu}$ and S_n are the solutions to the Maximum Likelihood Estimation problem:

$$\prod_{i=1}^{n} p(x_i \mid \mathcal{N}(\tilde{\mu}, S_n)) = \max_{\theta_1,\theta_2} \prod_{i=1}^{n} p(x_i \mid \mathcal{N}(\theta_1, \theta_2))$$

Show that:

$$\mathbb{E}[S_n] = \frac{n-1}{n}\sigma^2$$

Answer:

$$\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$$

we will be taking the expectations which is:

$$\mathbb{E}[S_n] = \frac{1}{n}\mathbb{E}\left[\sum X_i^2 - n\bar{X}^2\right]$$

calculating those terms will give us:

$$\mathbb{E}[X_i^2] = \sigma^2 + \mu^2$$

$$\mathbb{E}[\bar{X}^2] = \operatorname{Var}(\bar{X}) + \mathbb{E}[\bar{X}]^2 = \frac{\sigma^2}{n} + \mu^2$$

Lastly substituting this:

$$= \frac{1}{n}\left[n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right]$$

$$= \frac{1}{n}\left(n\sigma^2 - \sigma^2\right) = \frac{n-1}{n}\sigma^2$$

This shows the sample variance is biased downward by a factor of $\frac{n-1}{n}$. The unbiased estimator uses denominator (n-1) instead of n.

REFERENCES

[1] Multivariate normal distribution
https://en.wikipedia.org/wiki/Multivariate_normal_distribution

[2] Pearson correlation coefficient
https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[3] Gaussian Plot 2-D
https://github.com/probml/pyprobml/blob/auto_notebooks_md/notebooks.md#gauss_plot_2d.ipynb

[4] GMM Plot Demo
https://github.com/probml/pyprobml/blob/auto_notebooks_md/notebooks.md#gmm_plot_demo.ipynb

[5] GMM 2D
https://github.com/probml/pyprobml/blob/auto_notebooks_md/notebooks.md#gmm_2d.ipynb

[6] Gauss Infer 2D
https://github.com/probml/pyprobml/blob/auto_notebooks_md/notebooks.md#gauss_infer_2d.ipynb

[7] Gauss Imputation known params demo
https://github.com/probml/pyprobml/blob/auto_notebooks_md/notebooks.md#gauss_imputation_known_params_demo.ipynb

[8] Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis, Mitzenmacher, Michael and Upfal, Eli, Cambridge university press

[9] Probabilistic machine learning: an introduction, Murphy, Kevin P, MIT press