Introduction

A major task of probability theory is applying known probability densities to derive number characteristics such as expectation values, variances etc. Statistical inference, however, focuses on estimating parameters of unknown probability densities first, and then apply the estimated distributions.
The following is a summary of some important probability distributions. The relations among these distributions are also studied in order to highlight the specific conditions for each specific distribution. For a more complete list of probability distributions, see the following link: https://en.wikipedia.org/wiki/List_of_probability_distributions (https://en.wikipedia.org/wiki/List_of_probability_distributions) For their number characteristics, click each individual link.

Probability distribution for discrete time and space

The starting point: Bernoulli distribution

The PMF is:

$$f(x,p) = p^x(1-p)^{1-x} \quad x \in \{0,1\}$$

Bernoulli distribution is also called 0-1 distribution. Its expectation and variance are p and $p(1-p)$ respectively.

Probability of x occurrences after n Bernoulli trials: Binomial distribution

$$f(x,n,p) = C_n^x p^x (1-p)^{n-x} \quad x = 0,1,2,\dots n$$

Flipping n times and assume x up and $n-x$ down. There are C_n^x such cases, and therefore the above formula.

Conditions are required for the Binomial distribution:

- Two possible outcomes for each experiment. The probability for each outcome is constant.
- Each experiment is **independent** and identical.

These conditions are also required in other five distributions below that are strongly related binomial distribution. The expectation and variance are np and $np(1-p)$, which are n times of those of Bernoulli distribution.

Probability of first occurrence after n Bernoulli trials: Geometric distribution

$$f(n,p) = p(1-p)^{n-1}$$

Note we cannot obtain this by setting $x = 1$ in the binomial distribution. This is because we only want the $x = 1$ case for the first time, but not all possible cases where we flip n times and have only one occurrence.

Probability of first xth occurrences after n Bernoulli trials: Negative Binomial distribution

Also called Pascal distribution. If the xth occurrence occurs on the nth trial, that implies $(x-1)$th occurrences must have happened in the previous $(n-1)$ trials. Because these $(x-1)$ occurrences can occur in any order, the probability of this happening is just Binomial PMF:

$$P(X = x-1; n-1) = C_{n-1}^{x-1} p^{x-1} (1-p)^{n-k}$$

Furthermore, the probability that we will have an occurrence on the nth trial is p, therefore

$$f(x;p) \equiv P(X = x) = C_{n-1}^{x-1} p^{x} (1-p)^{n-x}$$

This is true for $n = x, x+1, \ldots$. For $n < x$, the probability is zero.
There are alternative formula in the following link https://en.wikipedia.org/wiki/Negative_binomial_distribution (https://en.wikipedia.org/wiki/Negative_binomial_distribution).

Probability distribution for continuous time and space

Sometimes we want to know the probability of an event occurring within a continuous time period or space. If we separate the continuous time or space into many tiny time or space intervals, then it is possible to analyze these distributions with the same methods introduced in previous chapter.

The value of PDF at value x is not the probability of $P(X = x)$. So PDF $f(x)$ can take on values larger than one (but the integral of $f(x)$ over any subset of R will be at most one.

From binomial to Poisson distribution

We now calculate the probability that there would be x events in t time, given some average number of occurrences ($\lambda = vt$) in time t. This can be handled with Bernoulli trials.

- Divide time t period into n equal intervals.
- Probability of occurrence in any interval is $p = \lambda/n$.

If the conditions for binomial distribution (see previous chapter) are satisfied, then the probability that the event will occur x times within n trials is

$$P(N=x,n)=C_n^x\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x}=\frac{\lambda^x}{x!}e^{-\lambda}\quad(n\to\infty)=\frac{(vt)^x}{x!}e^{-vt}$$

Conditions required to apply Poisson distribution

- From its relation to binomial distribution, we know Poisson distribution must also follow the conditions similar to binomial: (1) Each event is independent and identical. (2) The mean occurrence rate vt or λ (corresponding to np in binomial) is constant. **Note definition of λ**. There might be different definitions for different distributions. (3). The probability of two or more occurrences at same time or location is negligible.
- From the derivation above, the formula is true only for very big n. So Poisson distribution is only true for rare event, or small $p=\frac{\lambda}{n}$.

From these conditions we know that when $n\to\infty$ and when p is small, then binomial distribution can be replaced by Poisson distribution, which is easier to calculate. However, we must also be careful of other conditions such as iid, constant rate to apply Poisson formula. For example, the number of student who arrive at the student union per minute is not a constant (low rate during class time for example), and the arrivals of individual students are not independent (they tends to come in groups).

From geometrical to exponential distribution

When doing experiments in each tiny time interval during a long time period. What is the probability it takes time duration of t to have the first occurrence? This is similar to the geometric distribution for e.g. binomial coin flipping. What is the probability of first occurrence after n trials? Here taking time duration t is similar to taking how n trials.

Assuming T_1 is the time for first event. Then the probability that first event has not occurred is

$$P(T_1>t)\equiv P(X=0;t,v)=\frac{(vt)^0}{0!}e^{-vt}=e^{-vt}$$

The CDF for the occurrence of first event is therefore $P(T\le t)=1-e^{-vt}$, and PDF is derivative $P(T_1=t)=ve^{-vt}$

From negative binomial to Gamma distribution

Similar to geometric distribution, we can use Poisson distribution to derive the Gamma distribution.
Assuming T_k is the time until kth event. Then the probability of k or more occurrences in time t is

$$P(T_k\le t)=\sum_{x=k}^{\infty}P(X_t=x)=1-\sum_{x=0}^{k-1}\frac{(vt)^x}{x!}e^{-vt}$$

Thus the PDF is:

$$P(T_k = t) = \frac{d}{dt} P(T_k \leq t) = \frac{v(vt)^{k-1}}{(k-1)!} e^{-vt}$$

This special form of Gamma distribution give the probability of kth event after a time period of t. This is similar to the kth events after, e.g., flipping n times of coins, as described by negative binomial distribution. **Note the v here has the same definition of λ used in several places discussed later. But this definition is different from that in the Poisson distribution before**. General Gamma is given below.

Gamma distribution and its derivative distributions

General

https://en.wikipedia.org/wiki/Gamma_distribution (https://en.wikipedia.org/wiki/Gamma_distribution)
A shape parameter is a kind of numerical parameter of a parametric family of probability distributions.Such a parameter must affect the shape of a distribution rather than simply shifting it (as a location parameter does) or stretching/shrinking it (as a scale parameter does).

The gamma distribution is a two-parameter family of continuous probability distributions. It is a generalization of the exponential distribution. The exponential distribution, Erlang distribution, and chi-squared distribution are special cases of the gamma distribution. There are **three different parametrizations** (all positive real parameters) in common use:

- With a shape parameter κ and a scale parameter θ.
- With a shape parameter $\alpha = \kappa$ and an inverse scale parameter $\beta = \frac{1}{\theta}$, called a rate parameter.
- With a shape parameter κ and a mean parameter $\mu = \kappa\theta = \frac{\alpha}{\beta}$

The parameterization with κ and θ appears to be more common in econometrics and certain other applied fields, where for example the gamma distribution is frequently used to model waiting times. The parameterization with α and β is more common in Bayesian statistics.

Normally κ or α can be understood as the number of events k in the discrete case. When κ or α is an integer, $\Gamma(\kappa + 1) = \kappa!$. The other scale parameter θ or inverse scale parameter β is related to the rate of the events. In fact the β here is just the rate parameter λ, v, or r in many other distributions discussed later. However, note the λ in Poisson distribution is often defined as vt, so the λ in Poisson case is normally different from other distributions.

Gamma function and gamma distribution

Gamma function is defined as $\Gamma(\kappa) = \int_0^\infty x^{\kappa-1} e^{-x} dx \quad \kappa > 0$. Making the substitution $y = x/\lambda$ where λ is a positive real constant, the gamma function becomes $\Gamma(\kappa) = \int_0^\infty (\lambda y)^{\kappa-1} e^{-\lambda y} \lambda dy \quad \kappa > 0$. Divide both sides by $\Gamma(\kappa)$ yielding

$$\int_0^\infty \frac{\lambda^\kappa y^{\kappa-1} e^{-\lambda y}}{\Gamma(\kappa)} dy = 1,$$

which is suitable for a PDF of a random variable.

A continuous random variable X with pdf $f(x) = \frac{\lambda^\kappa x^{\kappa-1} e^{-\lambda x}}{\Gamma(\kappa)} \quad x > 0$ for some real constant $\lambda > 0$ and $\kappa > 0$ is a gamma(λ, κ) random variable. Check the $f(x)$ for $\kappa > 1, \kappa = 1, \kappa < 1$.

Properties of gamma function:

$\Gamma(n+1) = n! \quad n = 0, 1, \ldots$

$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \quad \alpha > 0$

$\Gamma(\frac{1}{2}) = \sqrt{\pi}$

Special cases of gamma distribution

Exponential distribution

When $\kappa = 1$

$$f(x) = \lambda e^{-\lambda x} \quad x > 0$$

Comparing to $P(T_1 = t) = ve^{-vt}$ obtained earlier, the λ here is just the rate v, or sometimes denoted as r. Both v and r are rate and have a dimension of 1 over time. However, the λ in the Poisson distribution before is defined as $\lambda = vt = rt$. So the λ there is different from the λ here. **Be careful of this point**. In many places, λ, v, r etc., have the same meaning for rate. However, in Poisson case, it is often defined as vt or rt.

Erlang distribution

When κ is a positive integer k

$$f(x) = \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} \quad x > 0$$

If $X_1, X_2, \ldots X_k$ are iid exponential (λ) random variable, then $X_1 + X_2 + \ldots + X_k$ ~ Erlang(λ, k). **Note the definition of λ here is same as an example later, but different from that of Poisson**.

χ^2 distribution

When κ is $k/2$, where k is a parameter known as the degrees of freedom, and $\lambda = 1/2$

$$f(x) = \frac{x^{k/2-1}e^{-x/2}}{\Gamma(k/2)2^{k/2}} \quad x > 0$$

The χ^2 distribution is related to the standard normal distribution. If a random variable Z has the standard normal distribution, then Z^2 has the χ^2 distribution with one degree of freedom. If $Z_1, Z_2, \ldots Z_k$ are independent standard normal variables, then $Z_1^2 + Z_2^2 + \ldots + Z_k^2$ has a χ^2 distribution with k degrees of freedom, as described by the PDF above.

Gamma-distribution related application

If the events are occurring according to the Poisson distribution the time till the occurrence of the first event is described by the exponential distribution. The time till the occurrence of the second event is described by the Gamma distribution with $k = 2$ (summation of the time intervals between two consecutive events). Comparing this to the discrete counterparts of geometric and negative binomial distributions discussed earlier.
For example it is known that under free flow conditions vehicular arrival pattern follows Poisson distribution. If the mean arrival rate is 5 vehicles/minute, the headways between consecutive vehicular arrivals follow exponential distribution with the following PDF $f(t) = \frac{1}{5}e^{-5t}$. The time gap between every two vehicles follows the Gamma distribution with PDF $f(t) = \frac{1}{5}(\frac{1}{5}t)e^{-5t}$. The time at which kth vehicle arrives at the measurement location follows the PDF $f(t) = \frac{1}{5}\frac{(\frac{1}{5}t)^{k-1}}{(k-1)!}e^{-\frac{1}{5}t}$. Generally this can be written as

$$f(t) = \lambda\frac{(\lambda t)^{k-1}}{\Gamma(k)}e^{-\lambda t}$$

Note the definition of λ is different from that in Poisson distribution. From above, we can understand that gamma distribution is a generalization of exponential distribution. To have an intuition on gamma distribution, imagine the random variable as the waiting time of some event. Or in the discrete case, the number of total flips to have the kth event.

Conditional probability and Bayes rule

Conditional probability

see machine learning notes

Bayes rule

see machine learning notes

Likelihood function vs probability

https://en.wikipedia.org/wiki/Likelihood_function (https://en.wikipedia.org/wiki/Likelihood_function)

- In statistics, a likelihood function (or just the likelihood) is **a particular function of the parameter** of a statistical model given data. In informal contexts, "likelihood" is often used as a synonym for "probability". In statistics, the two terms have different meanings. **Probability is function of x given parameter θ, while is likelihood is function of θ given x. where x is the outcome of random variable X.** Likelihood is used with each of the four main foundations of statistics: frequentism, Bayesianism, likelihoodism, and AIC-based.
- Let X be a discrete random variable (discrete) with probability mass function p depending on a parameter θ, then the function $L(\theta \mid x) = p_\theta(x) = P_\theta(X = x)$ considered as a function of θ, is the likelihood function (of θ, given the outcome x of the random variable X. For continuous random variable, we have similar definitions.
- An intuitive understanding: Because likelihood is function of parameters for a given x, we can think whether x is LIKELY (hence likelihood) to be in the probability distribution of given by θ_1, or $theta_2$ ....A more specific example about biased coin tossing in explaining expectation maximization algorithm (by Do and Batzoglou, 2008). After a tossing experiment, i.e. the x, we may calculate whether this experiment is LIKELY from coin A or coin B, which are modeled by different parameters.
- **Confusion of different notations**
  - When judging whether it is a likelihood or a probability, we need make sure which is the function argument, but not the specific written form. The $p_\theta(x)$ above is more like a function of x given θ. But it describes a function of θ given x. So it is a likelihood function.
  - When we describe the probability (not likelihood) of "the value x of X given the parameter value θ, we often write it as $P(X = x \mid \theta$. Although formally it is like a conditional probability, it is not. So instead, we write it as $P(X = x; \theta)$ to emphasize that it is not a conditional probability.
  - The likelihood is sometimes written as $L(\theta \mid x)$ and sometimes as $L(x \mid \theta)$. In other words, the order of the appearance of θ does not matter. The key is whether θ is taken as a function argument (variable). Anyway, the $\mid$ sign does not indicate conditional probability.
- In the Maximum Likelihood Estimation (MLE), we are just maximizing the likelihood. In deriving many supervised learning algorithms such as those in generalized linear models, we usually maximize the likelihood with the form $P(y \mid x; \theta$. Here y is conditioned on x but parameterized on θ. However, because the main purpose is treating $P(y \mid x; \theta$ as a function of θ and find its optimal value, we are maximizing likelihood, rather conditional probability.
- In deriving unsupervised learning algorithms with expectation maximization (EM), we are still maximizing likelihood but with an iterative approach. The likelihood function takes a form of joint probability, but essentially we should take it a likelihood as it is a function of parameters.

Expectation

Mean/average (or just expectation), variance, covariance, correlation etc., are obtained by expectation operator acting on the corresponding quantity. In other words, they are all 'expectation values'. However, the 'expectation value' is often used to refer to only the mean or average of a random variable. Other related concepts include sample mean, sample variance, sample covariance, covariance matrix, correlation matrix.

Expectation value / mean / average

- Finite case: $E(X) = \sum_{i=1}^{k} x_i p_i$. The expression of $E(x) = \frac{1}{k} \sum_{i=1} x_i$ is a special case of $p_i = \frac{1}{k}$.
- Countably infinite case $E(X) = \sum_{i=1}^{\infty} x_i p_i$, where $\sum_{i=1}^{\infty} |x_i| p_i$ converges.
- Absolutely continuous case $E(X) = \int_{\mathbb{R}} x f(x) dx$.

Variance

The variance of a random variable is the **expectation value** of the squared deviation from the mean.

- General definition $Var(X) = E((x - \mu)^2)$, or $Var(X) = E(X^2) - (E(X))^2$.
- Discrete random variable $Var(X) = \sum_{i=1}^{n} p_i (x_i - \mu)^2$, or $Var(X) = \sum_{i=1}^{n} p_i x_i^2 - \mu^2$. For a set of n equally likely values, p_i can be replaced by $\frac{1}{n}$. In this case, it can also be written as other forms without referring to mean(see Wikipedia).
- Continuous random variable $Var(X) = \int (x - \mu)^2 f(x) dx = \int x^2 f(x) dx - \mu^2$.

Covariance

The covariance is the expectation value of quantity $(X - E[X])(Y - E[Y])$. Variance is the special case of covariance where X and Y are same. So variance is also called auto-variance.

- General definition $cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$.
- Discrete variables If the random variable pair (X, Y) can take on values (x_i, y_i) for $i = 1, 2, \ldots n$ with equal probabilities $\frac{1}{n}$, the $cov(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i - E(X))(y_i - E(Y))$. Like variance, it can also expressed without directly referring to the means. If each pair of value is not with equal probability, then $\frac{1}{n}$ need be replaced with probability p_i for each pair of data.

Correlation

It is just the scaled form of covariance. $corrX, Y = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$.

Covariance matrix

$\mathbf{X} = (X_1, X_2, \ldots X_n)$is a random vector and X_i is scaler random variables. The the covariance matrix Σ is the matrix whole (i, j) element is $\mathbf{\Sigma}_{ij} = \operatorname{cov}(X_i, X_j)$. This is equivalent to $\mathbf{\Sigma} = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T]$. Recall that this is similar to the scalar-valued random variable covariance matrix except the transpose sign.
For the reason above, some people call the above $Var(\mathbf{X}) = \mathbf{\Sigma}$ as variance because it is the natural generalization to higher dimensions of the 1-dimensional variance. But anyway it is commonly called covariance matrix. However, the notation for the cross-covariance is defined as $\operatorname{cov}(\mathbf{X, Y}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{Y} - E[\mathbf{Y}])^T]$. In this regard, the above covariance matrix is also called **variance-covariance** matrix, as the diagonal of the matrix is variance.

Correlation (covariance) and dependence

- Non-zero off-diagonal elements of the covariance matrix $X^T X$ or XX^T (data centered) indicate that the relevant variables must be linearly dependent.
- The non-zero covariance/correlation elements, however, give no clue about the nonlinear dependence among variables. The nonlinearity requires defining nonlinear correlation / covariance, which should include more than two factors in the product of normal covariance definition.

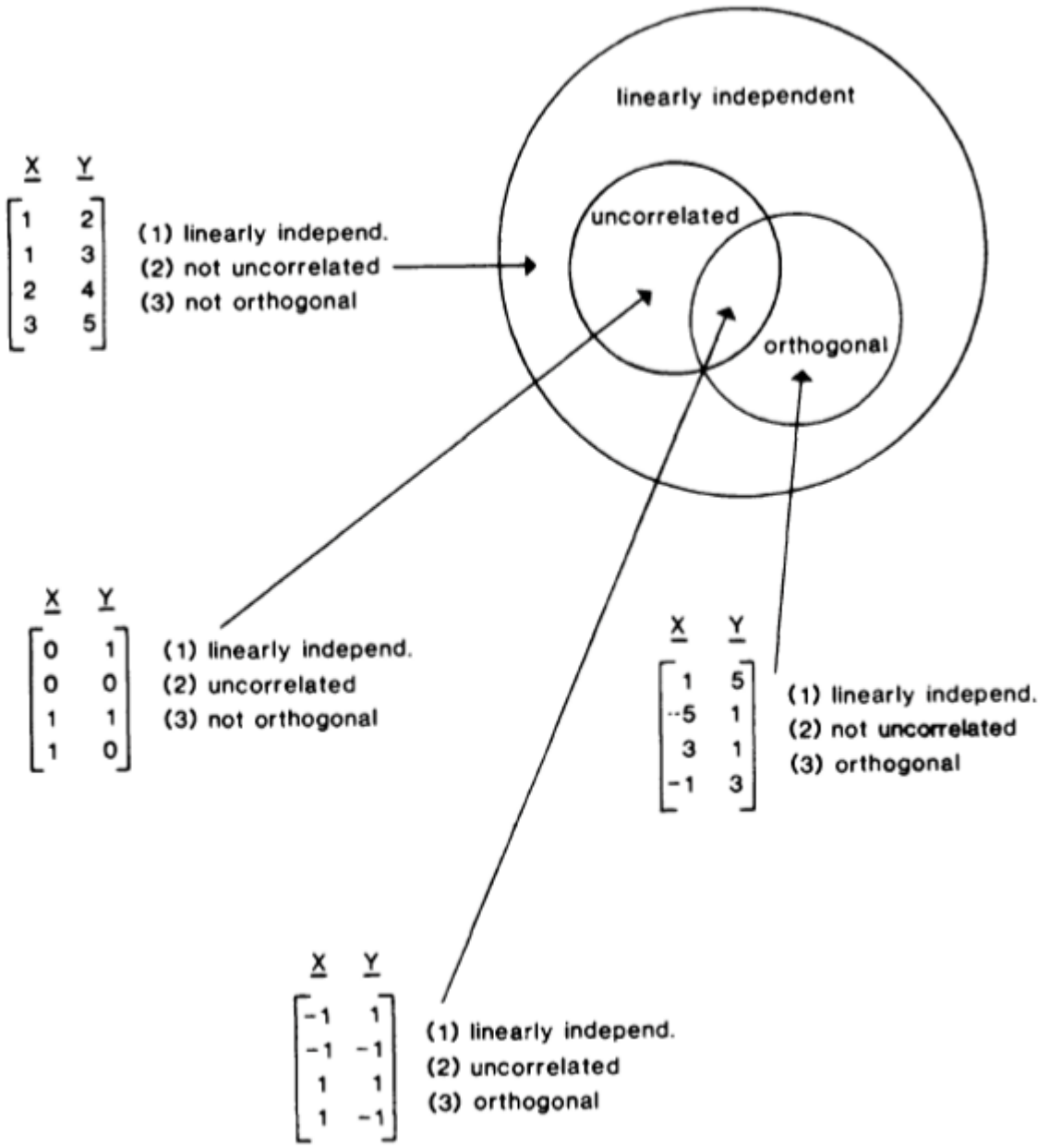
Correlation and basis change

- Two correlated variables will become linearly uncorrelated in the eigen-basis of covariance matrix. This does not mean that the correlation between the two variables are gone, but just means that two new variables, which are linear combination of the original variables, are no longer linear correlated in the new basis.

Uncorrelated, orthogonal and independent

- If we have only diagonal nonzero elements in covariance matrix, then it is safe to say there is no linear correlation, or linear dependence among relevant variables. However, it is in general not true to say that the variables are independent. They are linearly independent, but not necessarily nonlinearly independent.
- Assuming X is uniformly distributed in (-1,1), then it can be shown $cov(X, X^2) = 0$. The relationship between X and X^2 is non-linear, while correlation and covariance are measures of linear dependence between two variables. This shows two uncorrelated variables **does not in general imply that they are independent**.

- Independence implies uncorrelation but uncorrelation DOES NOT imply independence. However, when two variables are Gaussian, then uncorrelation and independence are equivalent.
- Relation of LINEAR independence, orthogonality and uncorrelation
  - Linearly independent vectors are those vectors that do not fall along the same line; that is, there is no multiplicative constant that will expand, contract, or reflect one vector onto the other!
  - Orthogonal vectors are a special case of linearly independent variables!
  - Uncorrelated vectors imply that once each variable is centered then the vectors are perpendicular.



Sample mean, variance, covariance matrix

The definitions of sample mean and sample variance are similar to those of mean and variance, except people often use the factor of $\frac{1}{n}$ for sample variance in order to achieve unbiased estimation. The next is about sample covariance matrix.
https://en.wikipedia.org/wiki/Sample_mean_and_covariance (https://en.wikipedia.org/wiki/Sample_mean_and_covariance) Assume X is a $N \times K$ matrix describing N observations with K features. The sample mean vector is $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x_i}$, where $\mathbf{x_i}$ is **the row vector** of the matrix X but not the column vector. We obtain K component means for the sample mean vector. The mean is calculated along the column vector. The sample covariance matrix is $Q = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})^T$.
Important notes This is similar to In the derivation of singular value decomposition (SVD), we calculate the eigenvalue/eigenvectors of matrix $A^T A$ and AA^T. If A is a $N \times K$ matrix, then $A^T A$ is a $K \times K$ matrix. Also we have $A^T A = \sum_{i=1}^{N} a_i a_i^T$. This is outer-product multiplication of matrices, where a_i is the column vector of A^T and thus its shape is $K \times 1$. The process here is exactly same as the covariance matrix definition except we subtracted mean there. So in PCA, calculate the eigenvectors of covariance matrix is same as the calculate the eigenvectors of $A^T A$, which is the key step in SVD. Check more detailed comparison between PCA and SVD in other notes.