Parametric models and exponential families

Contents

1 Motivation and formalization

In order to obtain an estimate of an unknown quantity, say μ_0, say, for example, a speed, it is common to take n measurements $x_1, \ldots, x_n$ and calculate their mean:

$$\bar{x}_n = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

But why should the observations be combined in this way ?

$\rightarrow$ the empirical mean is a relevant measure of the center of the observations, since

$$\bar{x}_n = \arg\min_a \sum_{i=1}^{n} (x_i - a)^2.$$

But (at this level) we can't justify why $\bar{x}_n$ is a good estimate (approximation) of the true value μ_0 since no explicit assumption has been made to connect the data $(x_1, \ldots, x_n)$ and μ_0.

To establish such a connection, we can, for example, presume that:

(i) each x_i is an observed value of a rv X_i, and
(ii) X_i, $i = 1, \ldots, n$, have a common mean μ_0.

Even more specifically, we can, for example, assume the following ***additive error model***

$$X_i = \mu_0 + \epsilon_i,\ \epsilon_i \sim N(0, \sigma_0^2),$$

or equivalently $X_i \sim N(\mu_0, \sigma_0^2)$. In this way, the problem we face is the estimation of $\mu_0 = E(X_i)$ from the sample $(X_1, \ldots, X_n)$.

1.1 General formalization

The data $\boldsymbol{x}^t = (x_1, \ldots, x_n)$ that we observe are believed to be generated by a rve $\boldsymbol{X}^t = (X_1, \ldots, X_n)$, which represents our random sample. We assume that $\boldsymbol{X}$ follows some joint distribution which is (partly) unknown. The set of assumptions made about this underlying joint distribution is what we call a ***statistical model***.

$X_1, \ldots, X_n$ are typically assumed to be ***iid*** copies of some population rv, which we denote hereafter by X. In this case the statistical model reduces to the set of assumptions about the distribution of X. To discribe this latter, we usually use a pd (i.e., probability density or mass function) f or a cdf (i.e., cumulative distribution function) F. Under the iid

assumption, the joint pd and the joint cdf of $\boldsymbol{X}$, denoted by f_n and F_n, respectively, are given by

$$f_n(\boldsymbol{x}) = f_n(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) \text{ and } F_n(\boldsymbol{x}) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} F(x_i).$$

The IID hypothesis plays a crucial role

Without the iid assumption the statistical analysis of the data becomes much more complicated. For example, as a statistical model, we could assume that

$$\boldsymbol{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with an unknown $\boldsymbol{\theta}^t = (\mu_1, \ldots, \mu_n, \sigma_{11}, \ldots, \sigma_{nn}) \in \mathbb{R}^{n+\frac{n\times(n+1)}{2}}$, where $\mu_i = E(X_i)$ and $\sigma_{ij} = Cov(X_i, X_j)$.

Now, under the iid assumption, $\boldsymbol{\theta}$ reduces to $(\mu_1, \sigma_{11}) \in \mathbb{R}^2$. □

1.2 Parametric models

A parametric model or parametric family is a set of distributions indexed by a ***finite*** dimensional parameter $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_d)$, $d \geq 1$. That is to say that the pd of X, *the rv that generated the data*, is known up to the parameter $\boldsymbol{\theta}$,

which is unknown. In which case, we denote the pd of X by $f(x;\boldsymbol{\theta})$ and its cdf by $F(x;\boldsymbol{\theta})$, and write $X \sim f(x;\boldsymbol{\theta})$ or $X \sim F(x;\boldsymbol{\theta})$.

The set of possible values for the parameter $\boldsymbol{\theta}$, that we denote by $\boldsymbol{\Theta}$, is called ***the parameter space***. Among all the elements of $\boldsymbol{\Theta}$, the parameter value $\boldsymbol{\theta}_0$ that actually generates the data is referred to as the ***true value***. Now, to keep things simple, when this does not affect understanding, we leave off the subscript 0 in $\boldsymbol{\theta}_0$ and simply use $\boldsymbol{\theta}$ to refer to both the unknown generic parameter of interest and its true value.

Example 1.1.

- The Bernoulli model:
$$f(x;\theta) = \theta^x(1-\theta)^{(1-x)}I(x \in \{0,1\}),\ \theta \in (0,1).$$
Thus, $X \sim Ber(\theta_0)$, for some particular but unknown $\theta_0 \in [0,1] = \Theta$.

- The Exponential model:
$$f(x;\theta) = \theta^{-1}e^{-x/\theta}I(x > 0),\ \theta > 0.$$
Thus, $X \sim Exp(\theta_0)$, for some unknown $\theta_0 \in (0,\infty) = \Theta$.

- The Normal model:
$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} I(x \in \mathbb{R}),\ \boldsymbol{\theta}^t = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty).$$
Thus, $X \sim N(\mu_0, \sigma_0^2)$, for some unknown $(\mu_0, \sigma_0^2) \in \mathbb{R} \times (0, \infty) = \boldsymbol{\Theta}$. □

If the distribution of X is not completely determined by a finite number of parameters, then the model is *nonparametric* or *semiparametric* (i.e., a mix of finite- and infinite-dimensional parameters).

Example 1.2.

- X has a pd f, with $\int f''(x)dx < \infty$ and/or $\int x^2 f(x)dx < \infty$.
- X has a symmetric distribution about 0, i.e. it has a pd satisfying $f(-x) = f(x)$, $\forall x$.
- X has a pd f satisfying $f(x;\theta) = f_0(x - \theta)$, where θ is an unknown parameter and f_0 an unknown pd symmetric about 0.

These models/distributions cannot be indexed by a finite dimensional parameter. The first two cases are examples of (fully) nonparametric models, while the last case is an example of a semiparametric model. □

Parametric or nonparametric inference ?

Let us assume that our objective is to estimate $\theta := F(s) = P(X \leq s) \in [0, 1]$ for a given s.

- Without making any assumption about the distribution of X, a reasonable estimator is the empirical cdf $\hat{F}(s) = n^{-1} \sum_i I(X_i \leq s)$.
- If we assume that $X \sim N(\mu, 1)$, then it is more reasonable to estimate θ by $\Phi(s - \hat{\mu})$, where Φ is the cdf of a $N(0,1)$ and $\hat{\mu}$ is any given estimator of μ.

  $\rightarrow$ Incorrect assumptions and restrictions on the underlying distribution of X will lead to wrong or ***biased*** conclusions. But correct assumptions lead to better and more ***efficient*** estimation. $\square$

If there is no $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that $X \sim f(x; \boldsymbol{\theta})$, then the model $\{f(x; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is said to be ***misspecified***. An example of misspecification is when a normal distribution is used for exponential data. A conclusion drawn from a statistical model is ***valid only if the chosen model is correctly specified***. In reality, no model is 100% correct, but some models are more useful than others in approximating the true underlying distribution of the data. Here, unless explicitly stated, we always assume that any model we fit to the data is correctly specified.

1.3 Identifiability

For a given statistical model, a given parameter $\boldsymbol{\theta}$ corresponds to a single distribution. However, this does not rule out the possibility that there may exist $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ such that $f(x; \boldsymbol{\theta}_1) = f(x; \boldsymbol{\theta}_2)$, $\forall x$. In this case, we cannot distinguish

between these two parameter values, even if we are given an infinite sample, since both will yield data which are distributed identically. We refer to such situation as an ***identifiability problem***.

Identifiability is an important property of a statistical model, which determines whether the parameter of the model can be recovered (estimated) from the observed data, which is only possible if different values of $\boldsymbol{\theta}$ lead to different distributed samples. Mathematically, this can be formulated by saying that, in a given model $\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, the parameter $\boldsymbol{\theta}$ (or the model) is identifiable, if, $\forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,

$$f(x;\boldsymbol{\theta}_1) = f(x;\boldsymbol{\theta}_2),\ \forall x \Rightarrow \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2.$$

Example 1.3.

- The Bernoulli, the Exponential, and the Normal models, as defined above, are identifiable.
- Let $X = \mu_1 + \epsilon$, where $\epsilon \sim N(\mu_2, 1)$ and μ_1 and μ_2 are unknown. Suppose that we observe X (and not ϵ), then $\theta = \mu_1 + \mu_2$ is identifiable but $\boldsymbol{\theta} = (\mu_1, \mu_2)$ is not.
- Let $X = |Y|$, where $Y \sim N(\mu, 1)$ and μ is unknown. Suppose that we observe X, then μ is not identifiable.

Let's verify the identifiability of the last Normal model. For that, observe that $f(x, \boldsymbol{\theta}_1) = f(x, \boldsymbol{\theta}_2)$, $\forall x$, is equivalent to

$$\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{(x-\mu_2)^2}{\sigma_2^2} = 2\log\frac{\sigma_2}{\sigma_1},\ \forall x$$

Since a parabolic function $ax^2 + bx + c$ becomes null if and only if $a = b = c = 0$, and since in our case $a = \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}$, we have that $\sigma_1 = \sigma_2$. As consequence, $f(x; \boldsymbol{\theta}_1) = f(x; \boldsymbol{\theta}_2)$, $\forall x$, is equivalent to

$$(x-\mu_1)^2 - (x-\mu_2)^2 = 0,\ \forall x \iff \mu_1 = \mu_2.$$

To verify the last example above (with $X = |Y|$), observe that $P_\mu(X \leq x) = \Phi(x-\mu) + \Phi(x+\mu) - 1$, where Φ is the cdf of $N(0,1)$. It follows that, $P_1(X \leq x) = P_{-1}(X \leq x)$, $\forall x$, i.e., $\mu = 1$ and $\mu = -1$ lead to the same distribution for X. This demonstrates that μ is not identifiable. □

If a model is not identifiable, it is common to introduce additional constraints/assumptions on it in order to make it identifiable. In that case, the set of these requirements is called the ***identifiability conditions.*** For instance, in our example above with $X = |Y|$, if we assume that $\mu > 0$, i.e., $\Theta = (0, \infty)$, then μ becomes identifiable; can you prove this ?

1.4 Purpose of inferential statistics

Statistical inference is the process of learning about a given probability model using observed data. To be more precise, suppose we are given a data set $\boldsymbol{x}$ which we assume to be generated from the model $\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. The aim of parametric statistical inference is to gain knowledge about the unknown parameter $\boldsymbol{\theta}$ from $\boldsymbol{x}$.

There are three major parametric statistical inference procedures:

1. **Point estimation**: A single value is computed from the data $\boldsymbol{x}$ and used as an estimate (approximation) of the the true parameter value $\boldsymbol{\theta}_0$.
2. **Hypothesis testing**: Sets up some specific hypotheses regarding $\boldsymbol{\theta}_0$ and assesses whether or not the data $\boldsymbol{x}$ support these hypothesis.
3. **Confidence set estimation**: Use the observed data $\boldsymbol{x}$ to construct a set of possible values for $\boldsymbol{\theta}_0$. The resulting set must have a high (predetermined) probability of including the true value.

Other well-known topics in statistical inference include *model selection*, *model validation* and *prediction*.

2 Exponential family

One important class of statistical models is exponential family models. These models are widely used in statistics and machine learning. They are characterized by a simple and elegant mathematical structure, which makes them analytically tractable and computationally efficient. Exponential family contains most of the standard discrete and continuous distributions that are used for modeling, such as (multivariate) normal, poisson, binomial, multinomial, exponential, and gamma.

The reason for the special status of the exponential family is that a number of important and useful results in inference can be unified within it. This family also forms the basis for an important class of regression models, known as generalized linear models.

2.1 One-parameter exponential family

A family of probability distributions that depend on a single (scalar) parameter θ is a one-parameter exponential family if it can be expressed as

$$f(x;\theta) = h(x)\exp(g(\theta)T(x) - B(\theta)), \ \forall x.$$

Here x can be a scalar or vector, $h(x) \geq 0$ and $T(x)$ are a functions of x only (*cannot depend on* θ), and $g(\theta)$ and $B(\theta)$ are functions of θ only (*cannot depend on* x). $B(\theta)$ is a normalizing constant, ensuring that $f(x;\theta)$ sums or integrates to 1. The set $\Theta = \left\{\theta : \int h(x) \exp\left(g(\theta)T(x)\right)dx < \infty\right\}$ is the parameter space of the family. $T(X)$ is referred to as ***natural sufficient*** statistic or simply natural statistic.

Note that the above parameterization is not unique since, for example, g could be multiplied by a nonzero constant a if T is divided by a. Also, in many cases $T(x) = x$.

An exponential family can be reparameterized as

$$h(x) \exp(\eta T(x) - A(\eta)).$$

This expression is called the ***canonical (or natural) representation***, and $\eta = g(\theta)$ is the ***canonical parameter***. Here $h(x)$ and $T(x)$ are the same as in the original parameterization, and $A(\eta) = B\left(g^{-1}(\eta)\right)$ is the new normalizing constant (assuming g is invertible). Notice that this parametrization is not unique either.

It's analytically convenient and easier to work with an exponential family in its canonical form. Once a result has been derived for the canonical form, we can rewrite it in terms of the original parameter θ if desired.

To verify that a family of pd's is an exponential family, we must identify all the functions h, T, g, and B (or A). The next example illustrates this.

Example 2.1.

- Poisson:

$$\begin{aligned}\frac{\theta^x e^{-\theta}}{x!} &= \frac{1}{x!}\exp(x\log(\theta)-\theta),\ x=0,1,\ldots, \text{ and } \theta>0\\ &\equiv \frac{1}{x!}\exp(x\eta-e^{\eta}),\ \eta\in(-\infty,\infty).\end{aligned}$$

- Binomial:

$$\begin{aligned}C_n^x\theta^x(1-\theta)^{n-x} &= C_n^x\exp\left(x\log\left(\frac{\theta}{1-\theta}\right)+n\log(1-\theta)\right),\ x=0,1,\ldots,n \text{ and } \theta\in(0,1)\\ &\equiv C_n^x\exp(x\eta-n\log(1+e^{\eta})),\ \eta\in(-\infty,\infty).\end{aligned}$$

- Normal with a *known* σ ($\theta=\mu$):

$$\begin{aligned}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-x^2}{2\sigma^2}\right)\exp\left(\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}\right),\ x\in(-\infty,\infty) \text{ and } \mu\in(-\infty,\infty)\\ &\equiv \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-x^2}{2\sigma^2}\right)\exp\left(\eta x-\frac{\eta^2\sigma^2}{2}\right),\ \eta\in(-\infty,\infty).\ \square\end{aligned}$$

Example 2.2 (Counter-example)**.** The following density

$$\begin{aligned} f(x;\theta) &= \exp(-(x-\theta)),\ \ x \geq \theta \text{ and } \theta \in (-\infty, \infty) \\ &= \exp(-(x-\theta))I(x \geq \theta) \end{aligned}$$

is not an exponential family because of $I(x \geq \theta)$ that cannot be factored into "x" part and an "θ" part. □

In general, for a pd $f(x;\theta)$ to be an exponential family, its *support* $S = \{x :\ \ f(x;\theta) > 0\}$ *must be free from* θ.

2.2 Properties of exponential family

A special fact about the exponential family is that the integral $\int h(x)\exp(g(\theta)T(x) - B(\theta))dx$, in the continuous case, or the sum $\sum_x h(x)\exp(g(\theta)T(x) - B(\theta))$, in the discrete case, can be differentiated, with respect to θ (or with respect η for the canonical representation), any number of times under the integral/sum sign (i.e. the derivatives can always be moved inside the integral/sum). This is the source of many interesting results about the exponential family. One of these results is given in the following.

Proposition 2.1. *With the canonical parameterization, the mean and variance of* $T \equiv T(X)$ *are given by*

$$E_\eta(T) = A^{'}(\eta) \text{ and } Var_\eta(T) = A^{''}(\eta). \tag{1}$$

For example, for $N(\mu, \sigma^2)$, given the results above (see Example 2.1), we see that

$$E(X) = \partial_\eta \left(\frac{\eta^2 \sigma^2}{2} \right) = \eta\sigma^2 = \mu \quad \text{and} \quad Var(X) = \partial_\eta(\eta\sigma^2) = \sigma^2$$

To show the first equality in (1), we can differentiate $\eta \mapsto \int h(x) \exp(\eta T(x) - A(\eta))dx = 1$, with respect to η, and then use the fact that the derivative can be moved inside the integral, which gives

$$\begin{aligned} 0 &= \int \partial_\eta h(x) \exp(\eta T(x) - A(\eta))dx \\ &= \int h(x)(T(x) - A^{'}(\eta)) \exp(\eta T(x) - A(\eta))dx \\ &= E_\eta(T) - A^{'}(\eta). \end{aligned}$$

As for the second equality, we can differentiate two times to get

$$\begin{aligned} 0 &= \int \partial_\eta h(x)(T(x) - A^{'}(\eta)) \exp(\eta T(x) - A(\eta))dx \\ &= \int h(x)(-A^{''}(\eta)) \exp(\eta T(x) - A(\eta))dx + \int h(x)(T(x) - A^{'}(\eta))^2 \exp(\eta T(x) - A(\eta))dx \\ &= -A^{''}(\eta) + E_\eta(T - A^{'}(\eta))^2. \end{aligned}$$

In terms of the original parameterization, with θ, we can write

$$E_\theta(T) = \frac{B^{'}(\theta)}{g^{'}(\theta)} \text{ and } Var_\theta(T) = \frac{\partial_\theta E_\theta(T)}{g^{'}(\theta)}.$$

These can be proven directly from the definition of $f(x;\theta)$ by following the same derivation as we did above for the canonical parameterization. Another way to obtain these results is to apply the chain rule to (1). In fact, the definitions given above yield to $A'(\eta) = \frac{B'}{g'}(g^{-1}(\eta))$.

Attention. In the formulas given above, we used the subscripts θ and η to indicate the parameterization used when calculating the expect values and the variances. This notation will be used wherever relevant hereafter.

Another interesting fact about the exponential family is that its structure is preserved under iid sampling. This is better explained in the following.

Proposition 2.2. *If $X_1, \ldots, X_n$ are iid rv from the exponential family, as defined above, with a natural statistic T then the the joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$:*

$$\left[\prod_{i=1}^{n} h(x_i)\right] \exp\left(\eta \sum_{i=1}^{n} T(x_i) - nA(\eta)\right)$$

is also an exponential family with natural statistic $\sum_{i=1}^{n} T(X_i)$.

2.3 Multiparameter exponential family

The multiparameter version of the exponential family is given by

$$\begin{aligned} f(x;\boldsymbol{\theta}) &= h(x) \exp\left(\sum_{j=1}^{J} g_j(\boldsymbol{\theta}) T_j(x) - B(\boldsymbol{\theta})\right) \\ &\equiv h(x) \exp\big(\boldsymbol{\eta}^t \boldsymbol{T}(x) - A(\boldsymbol{\eta})\big), \qquad \text{(Canonical parametrization)} \end{aligned}$$

where $\boldsymbol{\theta}^t = (\theta_1, \ldots, \theta_J)$, $\boldsymbol{\eta}^t = (\eta_1, \ldots, \eta_J)$, with $\eta_j = g_j(\boldsymbol{\theta})$, $\boldsymbol{T}^t(x) = (T_1(x), \ldots, T_J(x))$, and $g_j, A, B : \mathbb{R}^J \to \mathbb{R}$.

The properties that we have seen above for the one-parameter case also apply to the multi-parameter case. For example, it can be shown that

$$E(T_j(X)) = \partial_{\eta_j} A(\boldsymbol{\eta}), \text{ and } Cov(T_j(X), T_k(X)) = \partial_{\eta_j \eta_k} A(\boldsymbol{\eta}).$$

In matrix form, we can write $E(\boldsymbol{T}) = \boldsymbol{\nabla} A(\boldsymbol{\eta})$ and $Var(\boldsymbol{T}) = \boldsymbol{\nabla}^2 A(\boldsymbol{\eta})$.

Example 2.3. Normal distribution with an unknown μ and σ ($\boldsymbol{\theta} = (\mu, \sigma^2)$):

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \left(\frac{\mu^2}{2\sigma^2} + \log(\sigma)\right)\right) \\ &\equiv \frac{1}{\sqrt{2\pi}} \exp\left(\eta_1 x + \eta_2 x^2 - A(\boldsymbol{\eta})\right), \end{aligned}$$

where $A(\boldsymbol{\eta}) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$ Here $T_1(x) = x$ and $T_2(x) = x^2$.

It follows that,

$$\partial_1 A = -\frac{\eta_1}{2\eta_2} = \mu, \quad \partial_2 A = \frac{\eta_1^2 - 2\eta_2}{4\eta_2^2} = \mu^2 + \sigma^2 = E(X^2)$$
$$\partial_1^2 A = -\frac{1}{2\eta_2} = \sigma^2, \quad \partial_2^2 A = \frac{\eta_2 - \eta_1^2}{2\eta_2^3} = 2\sigma^2(\sigma^2 + 2\mu^2) = Var(X^2),$$

and $\partial_{1,2} A(\boldsymbol{\eta}) = \frac{\eta_1}{2\eta_2^2} = 2\mu\sigma^2 = Cov(X, X^2)$. □

3 Some useful tools

Law of total expectation/variance

For any two rv X and Y,

$$E(Y) = E\left(E[Y|X]\right),$$
$$Var(Y) = E\left(Var[Y|X]\right) + Var\left(E[Y|X]\right),$$

where $Var[Y|X] = E\big[(Y - E[Y|X])^2|X\big] = E[Y^2|X] - (E[Y|X])^2$.

Expected value of a non-negative rv

If X be a non-negative rv, then $E(X) \geq 0$. Moreover, $E(X) = 0$ if and only if $X = 0$ (with probability 1).

Markov-Chebyshev's inequality

If X is a non-negative rv, then $E(X) \geq kP(X \geq k),\ \forall k \in \mathbb{R}$. As a consequence, for any rv X and any constant $k > 0$, $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$, where $\mu = E(X)$ and $\sigma = \sqrt{Var(X)}$.

To see the first inequality, observe that, $\forall X, \forall k$, $X = XI(X \geq k) + XI(X < k)$. Which, by the fact that $X \geq 0$, implies that, $X \geq kI(X \geq k)$, and hence $E(X) \geq kP(X \geq k)$. Applying this last equality to $(X - \mu)^2$, instead of X, we obtain the second inequality.

Jensen's inequality

If g is a convex function (Reminder: $g''(x) \geq 0, \forall x \in I \Rightarrow g$ is convex in I), then

$$E(g(X)) \geq g(E(X))$$

Moreover, if g is strictly convex ($g'' > 0$), then this inequality is strict unless X is constant.
The opposite hold for concave function (Reminder: g is concave if and only if $-g$ is convex).

Example. Let X be a non-constant rv. Since $x \mapsto |x|^a$ is (strictly) convex for $a = (>)1$, we have that, for example, $E|X| \geq |E(X)|$ and $E(X^2) > (E(X))^2$. And since $x \mapsto \sqrt{x}$ and $x \mapsto \log(x)$ are strictly concave in $(0, \infty)$, we have that $E\left(\sqrt{X}\right) < \sqrt{E(X)}$ and $E(\log X) < \log(EX)$, provided that $X > 0$. □

Cauchy-Schwarz's inequality

For any two rv X and Y,

$$(E(XY))^2 \leq E(X^2)E(Y^2)$$

As a consequence, we get the inequality

$$(Cov(X,Y))^2 \leq Var(X)Var(Y).$$

Moreover, if X and Y are not constant, then the last inequality becomes an equality if and only if X and Y are linearly dependent, i.e. if and only if there exists numbers $a \neq 0$ and b such that $Y = aX + b$, with probability one.

The composite function rule (chain rule)

- Basic version : If $h(x) = f(g(x))$, then
$$h'(x) = f'(g(x))g'(x)$$

  By putting $z = f(y)$ and $y = g(x)$, the above formula can be expressed as $\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$.

This formula can be generalized in several ways.

- Case of multiple compositions: If $z = f(y)$, $y = g(x)$, and $x = h(t)$, then
$$\frac{dz}{dt} = \frac{dz}{dy} \cdot \frac{dy}{dx} \cdot \frac{dx}{dt}$$

- Case of a function of tow variables: If $z = f(x, y)$, $x = g(t)$, and $y = h(t)$, then

$$\frac{dz}{dt} = \frac{\partial z}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial z}{\partial y} \cdot \frac{dy}{dt}$$

Taylor's theorem

Suppose f is a function such that $f^{(n+1)}$ $(n \geq 0)$ is continuous on some interval I. Then, for any $x, a \in I$, there exists a $\theta \in [0, 1]$ such that

$$f(x) = \sum_{i=0}^{n} \frac{(x-a)^i}{i!} f^{(i)}(a) + \frac{(x-a)^{(n+1)}}{(n+1)!} f^{(n+1)}(a + \theta(x-a))$$

We can use this result to approximate the function f, and write that, in a sufficiently small neighbourhood of a,

$$f(x) \approx \sum_{i=0}^{n} \frac{(x-a)^i}{i!} f^{(i)}(a) = f(a) + (x-a)f(a) + \frac{(x-a)^2}{2} f^{'}(a) + \ldots + \frac{(x-a)^n}{n!} f^{(n)}(a).$$

This is called the the nth order Taylor polynomial approximation of f around a.
A similar result holds for functions of several variables. For example, the second-order Taylor polynomial approxima-

tion of $f : \mathbb{R}^d \mapsto \mathbb{R}$ around a point $\boldsymbol{a}$ is

$$f(\boldsymbol{x}) \approx f(\boldsymbol{a}) + \boldsymbol{\nabla}^t f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{a})^t \boldsymbol{\nabla}^2 f(\boldsymbol{a})(\boldsymbol{x} - \boldsymbol{a}).$$

Definite matrix

Definte matrices play a very important role in statistics and optimization ....

Let $\boldsymbol{A}$ be a *symmetric* matrix ($\boldsymbol{A}^t = \boldsymbol{A}$). $\boldsymbol{A}$ is said to be

- *positive definite*, if $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} > 0, \forall \boldsymbol{x} \neq \boldsymbol{0}$.
- *positive semidefinite*, if $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} \geq 0$, $\forall \boldsymbol{x}$.

If the inequalities are reversed, then $\boldsymbol{A}$ is *negative definite* or *negative semidefinite*, respectively.

Here are some examples. The matrix $\boldsymbol{A} = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$ is positive definite, since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = 3x_1^2 + 2x_2^2 > 0$, $\forall \boldsymbol{x} \neq \boldsymbol{0}$. The matrix $\boldsymbol{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is positive semidefinite since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = (x_1 + x_2)^2 \geq 0$, $\forall \boldsymbol{x}$. The matrix $\boldsymbol{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$ is positive

definte since $\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} = 2(x_1^2 - x_1x_2 + x_2^2) > 0$, $\forall \boldsymbol{x} \neq \boldsymbol{0}$. The matrix $\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ is *indefinite* since $\boldsymbol{x}^t\boldsymbol{A}\boldsymbol{x} = x_1^2 + 4x_1x_2 + x_2^2$ can be positive or negative.

Different methods exist to check if a matrix is positive definite, such as the Cholesky decomposition, the eigenvalues, or the principal minors. In this course we will not go into the details of these methods. It is sufficient to know that a 2×2 matrix $\boldsymbol{A} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ is positive definite if and only if $a > 0$ and $det(\boldsymbol{A}) := ac - b^2 > 0$.

There also are many interesting properties of positive (semi)definite matrices, such as the fact that :

- A positive definite matrix is invertible and its inverse is also positive definite.
- For a positive (semi)definite matrix $\boldsymbol{A}$, there exists a unique positive (semi)definite matrix $\boldsymbol{B}$ such that $\boldsymbol{B} \times \boldsymbol{B} = \boldsymbol{A}$. This $\boldsymbol{B}$ is called the (natural) square-root of $\boldsymbol{A}$ and is denoted by $\boldsymbol{A}^{1/2}$.

Some interesting properties of the Variance-Covariance matrix

Let $\boldsymbol{X}$ be a rve in $\mathbb{R}^d$ and $\boldsymbol{\Sigma} = Var(\boldsymbol{X}) = \big(Cov(X_j, X_k)\big)_{1 \leq j,k \leq d}$. $\boldsymbol{\Sigma}$ is symmetric ($\boldsymbol{\Sigma}^t = \boldsymbol{\Sigma}$) and positive semidefinite. This last property follows directly from the fact that $\boldsymbol{a}^t\boldsymbol{\Sigma}\boldsymbol{a} = Var(\boldsymbol{a}^t\boldsymbol{X})$.

$\boldsymbol{\Sigma}$ is positive definite if and only if the components of $\boldsymbol{X}$ are linearly independent (with probability 1); i.e. it is not possible to express any component of $\boldsymbol{X}$ (i.e. X_1 or X_2, ...) as a linear combination of the others.

Some properties of the multivariate normal

- If $\boldsymbol{B}$ is a $p \times d$ matrix, $\boldsymbol{a}$ is a $p-$dimensional vector, and $\boldsymbol{b}$ a $d-$dimensional vector, then

$$\boldsymbol{b}^t \times N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N\left(\boldsymbol{b}^t\boldsymbol{\mu}, \boldsymbol{b}^t\boldsymbol{\Sigma}\boldsymbol{b}\right), \text{ and } \boldsymbol{a} + \boldsymbol{B} \times N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N_p\left(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^t\right).$$

- $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Rightarrow \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_d(\boldsymbol{0}, \mathbb{1})$, where $\boldsymbol{\Sigma}^{-1/2}$ is the square-root of a of $\boldsymbol{\Sigma}^{-1}$, $\mathbb{1}$ is the identity matrix, and $\boldsymbol{0}$ the $d-$dimensional vector of zeros.
- $\boldsymbol{X} \sim N_d(\boldsymbol{0}, \mathbb{1}) \Rightarrow \boldsymbol{X}^t\boldsymbol{X} \sim \chi^2_d$, where χ^2_d the chi-squared distribution with d degrees of freedom.