Parametric models and exponential families

Contents

1 Motivation and formalization **3**

1.1 General formalization 4

1.2 Parametric models 5

1.3 Identifiability 9

1.4 Purpose of inferential statistics 12

2 Exponential family **13**

2.1 One-parameter exponential family 13

2.2 Properties of the exponential family 17

2.3 Multiparameter exponential family 21

3 Some useful tools **23**

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 $\boldsymbol{x} = (x_1, \ldots, x_n)$ that we observe are believed to be generated by a rve $\boldsymbol{X} = (X_1, \ldots, X_n)$, 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 describe this latter, we usually use a pd f or a cdf 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 $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_n, \sigma_{11}, \ldots, \sigma_{nn}) \in \mathbb{R}^{n+\frac{n\times(n+1)}{2}}$, 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 $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $d \geq 1$. That is to say that the pd of X – the random variable that generated the observed data – is known up to the unknown parameter $\boldsymbol{\theta}$. In which case, we denote the pd of X by $f(x; \boldsymbol{\theta})$ and its cdf by $F(x; \boldsymbol{\theta})$.

We may write $X \sim f(x;\boldsymbol{\theta})$ or, less commonly, $X \sim F(x;\boldsymbol{\theta})$. When the distribution has a well-known name (e.g., Normal, Poisson, Binomial), the notation $\sim$ is usually followed by the distribution name together with its parameters, as in $X \sim N(\mu, \sigma^2)$ or $X \sim \text{Poisson}(\lambda)$.

The set of possible values for the parameter $\boldsymbol{\theta}$, that we denote by $\boldsymbol{\Theta}$, is called ***the parameter space***. Within $\boldsymbol{\Theta}$, the particular value $\boldsymbol{\theta}_0$ that actually generated the observed data is referred to as the true parameter value (or simply the ***true value***).

Example 1.1.

- The Bernoulli model: $X \sim f(x;\theta)$, with

$$f(x;\theta) \equiv P_\theta(X = x) = \theta^x(1-\theta)^{(1-x)} I(x \in \{0,1\}),\ \theta \in (0,1).$$

  $\iff X \sim Ber(\theta_0)$, for some specific but unknown $\theta_0 \in (0,1) = \Theta$.

- The Exponential model: $X \sim f(x;\theta)$, with

$$f(x;\theta) = \theta^{-1} e^{-x/\theta} I(x \geq 0),\ \theta > 0.$$

  $\iff X \sim Exp(\theta_0)$, for some unknown $\theta_0 \in (0,\infty) = \Theta$.

- The Normal model: $X \sim f(x;\boldsymbol{\theta})$, with

$$f(x;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ \boldsymbol{\theta}^t = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty).$$

$\iff X \sim N(\mu_0, \sigma_0^2)$, for some unknown $(\mu_0, \sigma_0^2) \in \mathbb{R} \times (0, \infty) = \boldsymbol{\Theta}$. □

Attention. Hereafter, for simplicity and when no ambiguity arises, we drop the subscript 0 in $\boldsymbol{\theta}_0$ and use $\boldsymbol{\theta}$ to denote both the generic parameter and the unknown true parameter value.

Misspecified model

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.

In what follows, unless stated otherwise, we assume that the statistical models under consideration are correctly specified. □

Parametric vs nonparametric models ?

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 $f(x;\mu,\sigma) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right)$, where μ and $\sigma > 0$ are unknown parameters and f_0 an unknown pd symmetric about 0.

These models 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. □

The question of whether to use a parametric or a nonparametric model depends primarily on prior knowledge about the data-generating process and on the associated risk–benefit trade-off: efficiency/interpretability of parametric models against the flexibility/robustness of nonparametric ones.

For example, suppose we want to estimate $\theta := F(s) = P(X \leq s) \in [0,1]$ for a given s.

- If we assume that $X \sim N(\mu, 1)$, then it is reasonable to estimate θ by $\hat{\theta}_{\text{para}} = \Phi(s - \hat{\mu})$, where Φ is the cdf of a $N(0,1)$ and $\hat{\mu}$ is any given estimator of μ, e.g., the sample mean $\overline{X}_n$.
  - Advantage: Efficient (makes optimal use of avliable informations $\rightarrow$ small variance), if normality holds.
  - Risk: If the distribution is not normal, $\hat{\theta}_{\text{para}}$ can be severely biased.
- Without making any assumption about the distribution of X, a reasonable estimator is the empirical cdf $\hat{F}(s) = n^{-1} \sum_{i=1}^{n} I(X_i \leq s)$.
  - Advantage: Consistent for any distribution.
  - Risk: Larger variance than the parametric estimator when normality is true.

  $\rightarrow$ Incorrect assumptions about the underlying distribution of X produce biased conclusions, whereas correct assumptions yield more efficient estimation. $\square$

1.3 Identifiability

For a given parametric model, each parameter value $\boldsymbol{\theta}$ determines exactly the distribution of X. However, this does not exclude the possibility that two distinct parameter values $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ may generate exactly the same distribution, that is, $f(x; \boldsymbol{\theta}_1) = f(x; \boldsymbol{\theta}_2)$, $\forall x$.

In such a situation, the two parameter values are indistinguishable from the data, even with an infinite sample, since both produce identical distributions. This phenomenon is known as an ***identifiability problem***.

Identifiability is an important property of a statistical model, which determines whether the parameter of interest 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, for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ 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. Why ?
- 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 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 $ax^2 + bx + c$ vanishes (i.e., $ax^2 + bx + c = 0, \forall x$) 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 $\boldsymbol{x} = (x_1, \ldots, x_n)$ 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 true parameter value $\boldsymbol{\theta}$.

2. **Hypothesis testing**: Sets up some specific hypotheses regarding $\boldsymbol{\theta}$ and evaluate the degree to which the data $\boldsymbol{x}$ support these hypotheses.

3. **Confidence set estimation**: Use the observed data $\boldsymbol{x}$ to construct a set of possible values for $\boldsymbol{\theta}$. 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 the 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 $\Theta = \{\theta : \int h(x) \exp(g(\theta)T(x))dx < \infty\}$, in the continuous case, and $\Theta = \{\theta : \sum_x h(x) \exp(g(\theta)T(x))dx < \infty\}$, in the discrete case, is the parameter space of the family. $T(X)$ is referred to as ***natural sufficient*** statistic or simply the sufficient statistic. In many cases $T(x) = x$.

The exponential-family representation above is not unique. In particular, the functions $g(\theta)$ and $T(x)$ are only defined up to nonzero linear rescaling: one may multiply g by a nonzero constant a and divide T by the same constant without changing the resulting distribution. This ambiguity is purely algebraic and has no impact on the underlying probability distributions or any statistical conclusions drawn from them.

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, while the normalizing function becomes $A(\eta) = B\big(g^{-1}(\eta)\big)$, provided that g is invertible. The set $\Lambda = \{\eta : \int h(x) \exp(\eta T(x))dx < \infty\}$ is called the natural parameter space, with the convention that the integral is replaced by a sum in the discrete case.

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 given pd is a member of the exponential family, we must identify all the functions h, T, g, and B (or equivalently A). The key step is to rewrite the density so that: (1) all terms depending only on x are absorbed into $h(x)$; (2) all terms depending only on θ are collected into $B(\theta)$; and (3) the remaining mixed term factors as $g(\theta)\,T(x)$ inside the exponential. The next example illustrates this procedure.

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:

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

- Normal with a *known* $\sigma > 0$ ($\theta = \mu$):

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

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

$$f(x;\theta) = \exp(-(x-\theta)) I(x \geq \theta) = \begin{cases} \exp\big(-(x-\theta)\big), & x \geq \theta, \\ 0, & x < \theta. \end{cases}$$

is not an exponential family. The obstruction comes from the indicator term $I(x \geq \theta)$ which cannot be factored into a

function of x times a function of θ, as required for the exponential-family form. □

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

2.2 Properties of the exponential family

A useful fact about exponential families 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 ***by moving the derivative inside the integral or sum***. This property is the source of many important results about exponential families. One such result is given next.

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

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

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 use (1) and apply the chain rule to $\eta \mapsto A(\eta) = B\big(g^{-1}(\eta)\big)$, which yields

$$A'(\eta) = \frac{B'}{g'}\big(g^{-1}(\eta)\big) \equiv \frac{B'}{g'}(\theta)$$
$$A''(\eta) = \frac{1}{g'(g^{-1}(\eta))}\left(\frac{B'}{g'}\right)'\big(g^{-1}(\eta)\big) \equiv \frac{1}{g'(\theta)}\left(\frac{B'}{g'}\right)'(\theta),$$

with $\theta = g^{-1}(\eta)$.

Attention. In the formulas of expectations and variances above, the subscripts θ and η indicate the parameterization with respect to which the calculations are carried out. We will use this notation whenever necessary; otherwise, the subscripts will be omitted to lighten the notation.

Example 2.3. For $N(\mu, \sigma^2)$, see Example 2.1, we have that

- From the Canonical form :

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

- From the original form :

$$E(X) = \frac{\partial_\mu \frac{\mu^2}{2\sigma^2}}{\partial_\mu \frac{\mu}{\sigma^2}} = \frac{\frac{\mu}{\sigma^2}}{\frac{1}{\sigma^2}} = \mu \quad \text{and} \quad Var(X) = \frac{\partial_\mu \mu}{1/\sigma^2} = \sigma^2.$$

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 sufficient statistic T then the joint distribution of $\boldsymbol{X} = (X_1, \ldots, X_n)$:*

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

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

2.3 Multiparameter exponential family

A distribution is said to belong to the J-parameter exponential family ($J \geq 1$) if its density or probability mass function can be represented in the form

$$f(x;\boldsymbol{\theta}) = h(x)\exp\big(\boldsymbol{g}^t(\boldsymbol{\theta})\boldsymbol{T}(x) - B(\boldsymbol{\theta})\big) = h(x)\exp\left(\sum_{j=1}^{J} g_j(\boldsymbol{\theta})T_j(x) - B(\boldsymbol{\theta})\right),$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ is the parameter vector, $\boldsymbol{g}^t(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_J(\boldsymbol{\theta}))$, with each $g_j : \mathbb{R}^J \mapsto \mathbb{R}$, $B : \mathbb{R}^J \to \mathbb{R}$, and the vector of statistics $\boldsymbol{T}^t(x) = (T_1(x), \ldots, T_J(x))$ is called the sufficient statistic vector.

If we reparameterize by setting $\eta_j := g_j(\boldsymbol{\theta})$, $j = 1, \ldots, J$, the family is called the J-parameter canonical exponential family, and we obtain

$$f^*(x;\boldsymbol{\eta}) = h(x)\exp\big(\boldsymbol{\eta}^t\boldsymbol{T}(x) - A(\boldsymbol{\eta})\big),$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)$ is called the canonical (or natural) parameter vector, and $A : \mathbb{R}^J \to \mathbb{R}$.

The properties established in the one-parameter case extend directly to the multi-parameter setting. Using the

shorthand $T_j \equiv T_j(X)$, one obtains

$$E(T_j) = \partial_{\eta_j} A(\boldsymbol{\eta}), \text{ and } Cov(T_j, T_k) = \partial_{\eta_j \eta_k} A(\boldsymbol{\eta}),\ j,k = 1,\ldots,J.$$

In matrix notation, these identities become $E(\boldsymbol{T}) = \boldsymbol{\nabla} A(\boldsymbol{\eta})$ and $Var(\boldsymbol{T}) = \boldsymbol{\nabla}^2 A(\boldsymbol{\eta})$.

Example 2.4. 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 $(\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right) \iff (\mu, \sigma^2) = \left(-\frac{\eta_1}{2\eta_2}, -\frac{1}{2\eta_2}\right)$, and $A(\eta_1, \eta_2) = -\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\log(-2\eta_2)$.
Here $T_1(x) = x$ and $T_2(x) = x^2$. And we have that,

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

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

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

3 Some useful tools

In this section, we will go through some important mathematical and statistical properties that will be used later in the course.

Law of total expectation/variance

For any two rv X and Y,

$$\begin{aligned} E(Y) &= E\left(E[Y|X]\right), \\ Var(Y) &= E\left(Var[Y|X]\right) + Var\left(E[Y|X]\right), \end{aligned}$$

where $Var[Y|X] := E\left[(Y - E[Y|X])^2|X\right] = E[Y^2|X] - \left(E[Y|X]\right)^2$, and $E(Y|X = x) = \int y f_{Y|X}(y|x)dy$.

Expected value of a non-negative rv

If X is a non-negative rv, then $E(X) \geq 0$. Moreover, $E(X) = 0$ if and only if $X = 0$ almost surely (i.e., $P(X = 0) = 1$).

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

$$f(E(X)) \leq E(f(X)).$$

Moreover, if f is strictly convex ($f'' > 0$), then this inequality is strict unless X is almost surely constant (i.e., X takes the same value all the time; thus $\exists c$ such that $P(X = c) = 1$).
The opposite holds for concave functions (Reminder: f is concave if and only if $-f$ is convex).

Example. Let X be a non-constant rv. Since $x \mapsto |x|^a$ is convex for $a \geq 1$, and strictly convex for $a > 1$, we have that, for example, $|E(X)| \leq E|X|$ 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 $\sqrt{E(X)} > E\left(\sqrt{X}\right)$ and $\log(EX) > E(\log X)$, provided that $X > 0$ almost surely. □

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 random variables, the last inequality becomes strict unless X and Y are linearly dependent; that is, equality holds if and only if there exist constants $a \neq 0$ and b such that $Y = aX + b$ almost surely.

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 $\dfrac{dz}{dx} = \dfrac{dz}{dy} \cdot \dfrac{dy}{dx}$.

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 two 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 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 approximation 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

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

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

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

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 definite since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = 2(x_1^2 - x_1 x_2 + x_2^2) > 0$, $\forall \boldsymbol{x} \neq \boldsymbol{0}$. The matrix $\boldsymbol{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ is *indefinite* since $\boldsymbol{x}^t \boldsymbol{A} \boldsymbol{x} = x_1^2 + 4x_1 x_2 + x_2^2$ 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 matrix is positive definite if and only if it is positive semidefinite and invertible; its inverse is then also positive definite.
- For every positive (semi)definite matrix $\boldsymbol{A}$, there exists a unique positive (semi)definite matrix $\boldsymbol{B}$ such that $\boldsymbol{B}^2 := \boldsymbol{B} \times \boldsymbol{B} = \boldsymbol{A}$. This $\boldsymbol{B}$ is called the (natural or principal) 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 (almost surely); i.e. $\nexists \boldsymbol{a} \neq \boldsymbol{0}$ such that $\boldsymbol{a}^t \boldsymbol{X} =$ constant.

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 $\boldsymbol{\Sigma}^{-1}$.
- $\boldsymbol{X} \sim N_d(\boldsymbol{0}, \mathbb{1}) \Rightarrow \boldsymbol{X}^t\boldsymbol{X} \sim \chi^2_d$, where χ^2_d the chi-squared distribution with d degrees of freedom.