Exponential families Part I - Statistics

Amine Hadji

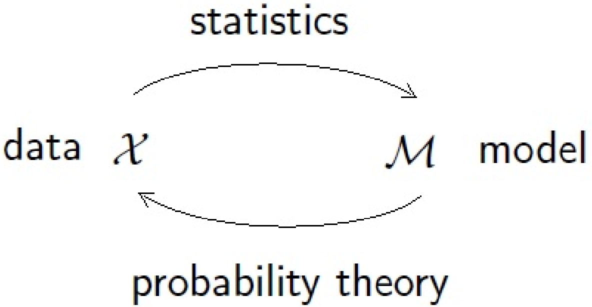
University of Leiden

m.a.hadji@math.leidenuniv.com

November 21, 2018

Overview

Probability and Statistics



Statistical model

"All models are wrong, but some are useful" - George E.P. Box

Definition (Statistical model)

A *statistical model* is a pair of sets $(\mathcal{X}, \mathcal{M})$ with $\mathcal{X}$ is the set of possible observations (i.e. the sample space), and $\mathcal{M}$ is a set of probability distributions on $\mathcal{X}$.

$\mathcal{X}$ is generally implicitly defined via $\mathcal{M}$

Definition (Parametric model)

A *parametric model* $\mathcal{M}_\Theta = \{P_\theta : \theta \in \Theta\}$ is a set of probability distributions P_θ described by a parameter θ, such that the dimension of Θ is finite.

$\mathcal{M}_\Theta$ can also be seen as the image of Θ for the map $\theta \mapsto P_\theta$.

Statistical model

Example 1 - Binomial random variable
Let X be a binomial random variable $B(r, \theta)$. $\mathcal{X} = \{1, 2, ..., r\}$, $\Theta = [0, 1]$ and for all $\theta \in \Theta$, P_θ is defined as:

$$P_\theta(X = i) = \binom{r}{i} \theta^i (1 - \theta)^{r-i} \text{ for all } i = 1, 2, ..., r.$$

The set of P_θ for all $\theta \in \Theta$ is a binomial random variable model, which is a subset of the probability simplex Δ_r.

Statistical model

Example 2 - Multivariate Normal random variable

Let $X \in \mathcal{X} = \mathbb{R}^m$ be a multivariate normal random variable $N(\mu, \Sigma)$. $\Theta = \mathbb{R}^m \times PD_m$, where PD_m is the cone of symmetric $m \times m$ positive definite matrices. For all $\theta = (\mu, \Sigma) \in \Theta$, the density f_θ of X is defined as:

$$f_\theta(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\Big)$$

The set of probability distributions P_θ generated by the densities f_θ for all $\theta \in \Theta$ is a model of non-degenerate multivariate normal random variable.

Statistical model

Example 3 - A non-parametric model
Let $\mathcal{M}_\Theta = \{P_\theta : \theta \in \Theta\}$ be a model, such that P_θ is generated by the density θ and.

$$\Theta = \{\theta \in C^0(\mathbb{R}) : \theta \text{ is a density}, \int_{\mathbb{R}} x\theta(x)dx = 0\}$$

$\mathcal{M}_\Theta$ is a non-parametric model, since the dimension of the parameter space Θ is infinite.

Identifiability

Definition (Identifiability)

A parametric model is said to be *identifiable* if the map $\theta \mapsto P_\theta$ which defines it is one-to-one (i.e. bijective).

Examples 1 and 2 are both identifiable.

Identifiability

Example - A non-identifiable model
Let $X \in \mathbb{R}^m$ be a multivariate normal random variable $N(\mu_1 + \mu_2, \Sigma)$. $\Theta = \mathbb{R}^m \times \mathbb{R}^m \times PD_m$. For all $\theta = (\mu_1, \mu_2, \Sigma) \in \Theta$, the density f_θ of X is defined as:

$$f_\theta(x) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-(\mu_1+\mu_2))'\Sigma^{-1}(x-(\mu_1+\mu_2))\right)$$

The model defined by these densities is not identifiable.

Data

In statistics, we assume that *data* x is a realization of a random variable X with probability distribution P. In the rest of the talk, we'll refer to both the realization and the random variable as data, and we'll note it X (i.e. $P(X) = P(X = x)$).

Ideally, $P \in \mathcal{M}$ meaning that our data is generated by a distribution in our model, but that's not always the case.

Data

The most common way to generate data is the case of independent and identically distributed data.

Definition (Independent and identically distributed)

Let $X = (X_1, X_2, ..., X_n)$ such that $X_i \in S$ for all $i = 1, 2, ..., n$, $\mathcal{X} = S^n$. The data X is independent and identically distributed (iid) if $X_1, X_2, ..., X_n$ are mutually independent and all have the same distribution P, which we call marginal distribution.

Then the distribution of the data is defined by the marginal distribution.

$$P(X_1, X_2, ..., X_n) = \Pi_{i=1}^{n} P(X_i)$$

Sufficience

Definition (Statistic)

A *statistic* is a function T from $\mathcal{X}$ to another space S.

Definition (Sufficient statistic)

A statistic is *sufficient* for θ if the data $X \in \mathcal{X}$ and the parameter θ are conditionally independent given $T(X)$.

In a discrete model, this can be rewritten as $P(X|T(X) = t, \theta) = P(X|T(X) = t)$, or even $P(X|\theta) = h(X)g(T(X), \theta)$.

Sufficience

Theorem (Fisher-Neyman factorization)

If the data X has a density f_θ, then T is sufficient for θ if and only if there exist two nonnegative functions g and h such that

$$f_\theta(x) = h(x)g(T(x), \theta).$$

Estimation - Maximum likelihood

Definition (Estimator)

An *estimator* of a parameter θ (resp. $g(\theta)$) is a statistic with codomain Θ (resp. $g(\Theta)$).

Definition (Likelihood)

Let X the data with distribution P_θ, which depends on $\theta \in \Theta$. The *likelihood function* is the function $\mathcal{L} : \Theta \to \mathbb{R}$ such that

- $\mathcal{L}(\theta|X) = P_\theta(X)$ if X is discrete.
- $\mathcal{L}(\theta|X) = f_\theta(X)$ if X is continuous with density $f_\theta(X)$.

The *log-likelihood function* is the logarithm of the likelihood function $\ell(\theta|X) = \log \mathcal{L}(\theta|X)$.

Estimation - Maximum likelihood

Definition (Maximum likelihood estimator)

The *maximum likelihood estimator* for θ is the maximizer of the likelihood ($\hat{\theta}_{MLE}(X) = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta|X)$).

Proposition

Let X and Y two independent data sets generated by the same distribution P_θ. Let T be a sufficient statistic for θ. If $T(X) = T(Y)$, then $\hat{\theta}_{MLE}(X) = \hat{\theta}_{MLE}(Y)$.

Estimation - Maximum likelihood

Theorem (Consistency)

If the model has nice properties (identifiability being the most important) and the data $X = (X_1, X_2, ..., X_n)$ are iid with distribution P_θ, then $\hat{\theta}_{MLE}(X)$ is consistent, meaning that it converges in probability to θ (i.e. $\lim_{n\to\infty} P(\|\hat{\theta}_{MLE}(X) - \theta\| > \varepsilon) \to 0$)

Definition

Consider a sample space $\mathcal{X}$ on which is defined a σ-finite measure ν, let $T : \mathcal{X} \to \mathbb{R}^k$ be a statistic and $h : \mathcal{X} \to \mathbb{R}_+$ a measurable function.

$$N_k(T, h) := \{\eta \in \mathbb{R}^k : \int_{\mathcal{X}} h(x) \exp(\eta' T(x)) d\nu(x) < \infty\}$$

is called a *natural parameter space*. For $\eta \in N_k(T, h)$, we can define a probability density p_η on $\mathcal{X}$ as

$$p_\eta(X) = h(X) \exp(\eta' T(X) - \phi(\eta)).$$

Definition

Definition (Exponential families 1)

Let k be a positive integer. The set of probability distributions $\{p_\eta : \eta \in N_k(T, h)\}$ forms a *regular exponential family* of order k if N is an open set in $\mathbb{R}^k$.
The statistic T is called a *canonical sufficient statistic* and η a *natural parameter*.

Definition (Exponential families 2)

Let $\mathcal{M}_\Theta = \{p_\theta : \theta \in \Theta\}$ be a parametric model. If all $p_\theta \in \mathcal{M}_\Theta$ can be written as

$$p_\theta(X) = h(X) \exp(\eta(\theta)' T(X) - \phi(\eta(\theta))),$$

then $\mathcal{M}_\Theta$ is an *exponential family*.

Examples

Example 1 - Binomial random variable
Let $\mathcal{M}_\Theta$ the binomial random variable model with r trials. If $\theta \in]0,1[$, all $p_\theta \in \mathcal{M}_\Theta$ can be written as

$$p_\theta(X) = \binom{r}{X} \theta^X (1-\theta)^{r-X}$$
$$p_\theta(X) = \binom{r}{X} \exp(X \log(\theta/(1-\theta)) + r \log(1-\theta)).$$

Then $\mathcal{M}_\Theta$ is an exponential family, $T(X) = X$ is a sufficient statistic and $\eta = \log(\theta/(1-\theta))$ is a natural parameter.

Examples

Example 2 - Multivariate Normal random variable
Let $\mathcal{M}_\Theta$ be the model of non-degenerate multivariate normal random variables in $\mathbb{R}^m$. All densities f_θ generating a distribution in $\mathcal{M}_\Theta$ can be written as:

$$f_\theta(X) = \frac{1}{(2\pi)^{m/2}|\Sigma|^{1/2}} \exp\Big(-\frac{1}{2}(X-\mu)'\Sigma^{-1}(X-\mu)\Big)$$
$$f_\theta(X) = \exp\Big(\eta(\theta)'T(X) - \phi(\eta(\theta))\Big)$$

with

$$T(X) = (X_1, \dots, X_m, -X_1^2/2, \dots, -X_m^2/2, -X_1X_2, \dots, -X_{m-1}X_m)$$
$$\eta(\theta) = (\sum_j \sigma_{1,j}\mu_j, \dots, \sum_j \sigma_{m,j}\mu_j, \sigma_{1,1}, \dots, \sigma_{m,m}, \sigma_{1,2}, \dots, \sigma_{m-1,m}).$$

where $\mu = (\mu_i)_{1\le i\le m}$ and $\Sigma^{-1} = (\sigma_{i,j})_{1\le i,j\le m}$

Examples

Example 3 - Uniform random variable
Let $\mathcal{M}_\Theta$ be the model of uniform random variables $U_{[0,\theta]}$ (i.e. with densities $f_\theta = \mathbf{1}_{[0,\theta]}/\theta$). All densities f_θ generating a distribution in $\mathcal{M}_\Theta$ are in the form:

$$f_\theta(X) = \frac{1}{\theta}\mathbf{1}_{0\leq X\leq\theta}.$$

Since there is no way of writing f_θ in an exponential form, $\mathcal{M}_\Theta$ is not an exponential family.

Setting

From now on, we'll assume that X is discrete with sample space $\mathcal{X} = \{1, ..., r\}$ and that its distribution p_η is in an exponential family of order k. These assumptions leads to a number of simplications of the representation of the model. Indeed $p_\eta(X) = h(X)\exp(\eta' T(X) - \phi(\eta))$ can be written as

$$p_\theta(X) = \frac{1}{Z(\theta)} h_X \prod_j \theta_j^{a_{jX}},$$

where $h_X = h(X)$, $\theta_j = e^{\eta_j}$, $a_{jX} = T_j(X)$ and $Z(\theta)$ is a normalizing parameter.

In particular, if all the a_{jX} are integers, then the exponential family is a parametrized family of probability distribution where the parametrizing functions are rational.

Log-linear model

An equivalent representation of the model is given by

$$\log p_\theta = \log(h) + \log(\theta)' A - \log Z(\theta)\mathbf{1},$$

where $\log p_\theta$ (resp. $\log(h)$) is the vector $(\log p_\theta(x))_{1\leq x\leq r}$ (resp. $(\log h(x))_{1\leq x\leq r}$) and A is the matrix $(a_{jx})_{1\leq j\leq k, 1\leq x\leq r}$.
If we make the assumption that the vector $\mathbf{1} = (1, 1, ..., 1)$ is in rowspan(A), then the representation is equivalent to say that $\log p_\theta$ belongs to the affine space $\log(h)+\text{rowspan}(A)$ for all θ.

Log-linear model

Definition

Let $A \in \mathbb{Z}^{k \times r}$ be a matrix of integers such that $\mathbf{1} \in \text{rowspan}(A)$ and $h \in (\mathbb{R}_+)^k$. The *log-affine model* associated to A and h is the set of probability distributions

$$\mathcal{M}_{A,h} := \{p \in \text{int}(\Delta_{r-1}) : \log p \in \log h + \text{rowspan}(A)\}.$$

If $h = \mathbf{1}$, then $\mathcal{M}_{A,h}$ is called a *log-linear model*.

Toric ideals

Definition

Let $A \in \mathbb{Z}^{k \times r}$ be a matrix of integers such that $\mathbf{1} \in \text{rowspan}(A)$ and $h \in (\mathbb{R}_+)^k$. The monomial map associated to A and h is the rational map
$\phi^{A,h} : \mathbb{R}^k \to \mathbb{R}^r$, where $\phi_j^{A,h}(\theta) = h_j \prod_{i=1}^k \theta_j^{a_{ij}}$.

Definition (Toric ideal)

Let $A \in \mathbb{Z}^{k \times r}$ be a matrix of integers such that $\mathbf{1} \in \text{rowspan}(A)$ and $h \in (\mathbb{R}_+)^k$. The ideal $I_{A,h} := I(\phi^{A,h}(\mathbb{R}^k))$, which is a subset of $\mathbb{R}[p_\theta]$, is called the *toric ideal* associated to A and h. When $h = \mathbf{1}$, we denote $I_A := I_{A,\mathbf{1}}$.

Toric ideals

Since generators for the ideal $I_{A,h}$ can be easily obtained from generators of I_A, we focus on the case of the toric ideal I_A.

Proposition

Let $A \in \mathbb{Z}^{k \times r}$ be a matrix of integer. Then the toric ideal I_A is a binomial ideal and

$$I_A = \langle p^u - p^v : u, v \in \mathbb{N}^r \text{ and } Au = Av \rangle.$$

If $\mathbf{1} \in \text{rowspan}(A)$, then I_A is homogeneous.

Examples

Example 1 - Twisted Cubic

Let $A = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{pmatrix}$. I_A is the vanishing ideal of the parametrization

$$p_1 = \theta_2^3, \ p_2 = \theta_1\theta_2^2, \ p_3 = \theta_1^2\theta_2, \ p_4 = \theta_1^3.$$

The toric ideal is generated by three quadratic binomials

$$I_A = \langle p_1p_3 - p_2^2, p_1p_4 - p_2p_3, p_2p_4 - p_3^2 \rangle.$$

The variety $V(I_A)$ is the twisted cubic curve. Note that if $h = (1, 3, 3, 1)$, then I_A is the vanishing ideal of the binomial random variable model with 3 trials.

Examples

Example 2 - Discrete independent random variables
Let X_1, X_2 two independent variables respectively in $\{1, ..., r_1\}$ and $\{1, ..., r_2\}$ such that

$$P(X_1 = i, X_2 = j) = p_{ij} = \alpha_i \beta_j \; i \in \{1, ..., r_1\} \text{and } j \in \{1, ..., r_2\},$$

where α and β are independent parameters. Since the distribution is a rational function of the parameters, it's possible to find A a matrix $(r_1 + r_2) \times (r_1 r_2)$ representing the toric ideal of this model.

Examples

For example with $r_1 = 2$ and $r_2 = 3$, we get

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

By computing a Gröbner basis for the toric ideal I_A, we can see that

$$I_A = \langle p_{i1j1}p_{i2j2} - p_{i1j2}p_{i2j1} : i_1, i_2 \in \{1, ..., r_1\}, j_1, j_2 \in \{1, ..., r_2\} \rangle.$$