Distances and Divergences in Robust, Semi and Non-Parametric Statistics and IA (Neural Networks)

Biomedical Applications

Catherine Huber

Université Paris-Cité (France), Appl. Math. Lab. (MAP5).

EDMSA, Université de Caen, May 14th, 2024

catherine.huber@parisdescartes.fr
https://www.biomedicale.univ-paris5.fr/catherine.huber/
https://map5.mi.parisdescartes.fr/

OUTLINE

1. **Parametric models : robustness needed**
   — **Optimality is lost if the model is not strictly respected** by the data, which is unavoidable. It leads to :
   — **Minimize the maximum loss on a neighborhood** of the model (**minimax** procedures), involving a distance on probability spaces.

2. **Non parametric models : function estimation**
   — Optimal function estimation : **best speed** of convergence.
   — Examples : probability density, spectral density, hazard function.

3. **Biomedical applications : diagnosis and survival data**
   — **Diagnosis** on sparse contingency tables :
   hierarchical log-linear models.
   — **Censoring and truncation** of survival data.
   Cox, Frailty and FHT semi-parametric models.

4. **No model : neural networks (NN)**
   — **Prediction** performance and **explainability**.

FOREWORD : distances and divergences

Distances and divergences on probability spaces $(E, \mathcal{B}, \mathcal{P})$ and the relationships between them and information theory play a major role in statistics. Among them I shall cite

1. **Prohorov** distance[1] $\pi(P,Q)$ particularly **useful for robustness as it takes into account rounding and gross errors** :

$$\begin{array}{lcll} \pi(P,Q) & = & \inf(\varepsilon > 0 \quad : & Q(B) \leq P(B^{\varepsilon}) + \varepsilon) \quad \forall B \in \mathcal{B} \quad , E \text{ metric(d)} \\ \pi(P,Q) & \in & [0\,,1] & B^{\varepsilon} = \{z \in E : \exists x \in E, d(z,x) \leq \varepsilon\}. \end{array}$$

2. **Total variation distance** $TV(P,Q)$, more tractable than Prohorov :

$$TV(P,Q) = \sup_{B \in \mathcal{B}} |Q(B) - P(B)| \in [0\,,1] \tag{1}$$

3. **Kullback-Leibler** $KL(P,Q)$, a divergence (a distance when symmetrized), strongly related to information

$$KL(P,Q) = \int \log(\frac{dP}{dQ})\, dP \in [0\,\infty[\tag{2}$$

1. Bretagnolle, Jean et Huber, Catherine. “Lois empiriques et distance de Prohorov”. Séminaire de probabilités de Strasbourg, vol. 12, p. 332-341 (1978).

4. **Shannon** [2] [3] (or mutual) **information of X and Y** :

$$I(X,Y) = KL(\mathcal{L}(X,Y), \mathcal{L}(X) \otimes \mathcal{L}(Y)) \tag{3}$$

5. **Hellinger** distance, $h(P,Q)$, also :

$$h^2(P,Q) = \frac{1}{2} \int (\sqrt{dP} - \sqrt{dQ})^2 \in [0\,1] \tag{4}$$

Depending on the objective, one or the other is used :

1. In robustness : to define the expected neighborhood of the assumed parametric model.
2. More generally : to define the risk of a procedure and its speed of convergence as a function of the size of the data set.

---

2. Bretagnolle, Jean, and Catherine Huber."Estimation des densités : risque minimax." Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47 :119-137, (1979).

3. Russac, Yoan, Claire Vernade, and Olivier Cappé. "Weighted linear bandits for non-stationary environments." Advances in Neural Information Processing Systems 32 (2019).

Relationships between TV and KL [4] :

— Pinsker inequality :

$$TV(P,Q) \leq \sqrt{\frac{1}{2} KL(P,Q)} \tag{5}$$

— Tsybakov version of Pinsker inequality :

$$TV(P,Q) \leq 1 - \frac{1}{2} \exp(-KL(P,Q)) \tag{6}$$

— Bretagnolle-Huber inequality (BH) :

$$TV(P,Q) \leq \sqrt{1 - \exp(-KL(P,Q))} \tag{7}$$

KL additivity for product distributions allows to define the complexity of a statistical problem :

$$KL(P^{\otimes n}, Q^{\otimes n}) = \mathbf{n}\ KL(P,Q)$$

4. Wikipedia : Bretagnolle-Huber Inequality, see also Canonne, Clément L. "A short note on an inequality between KL and TV." arXiv preprint arXiv :2202.07198 (2022)

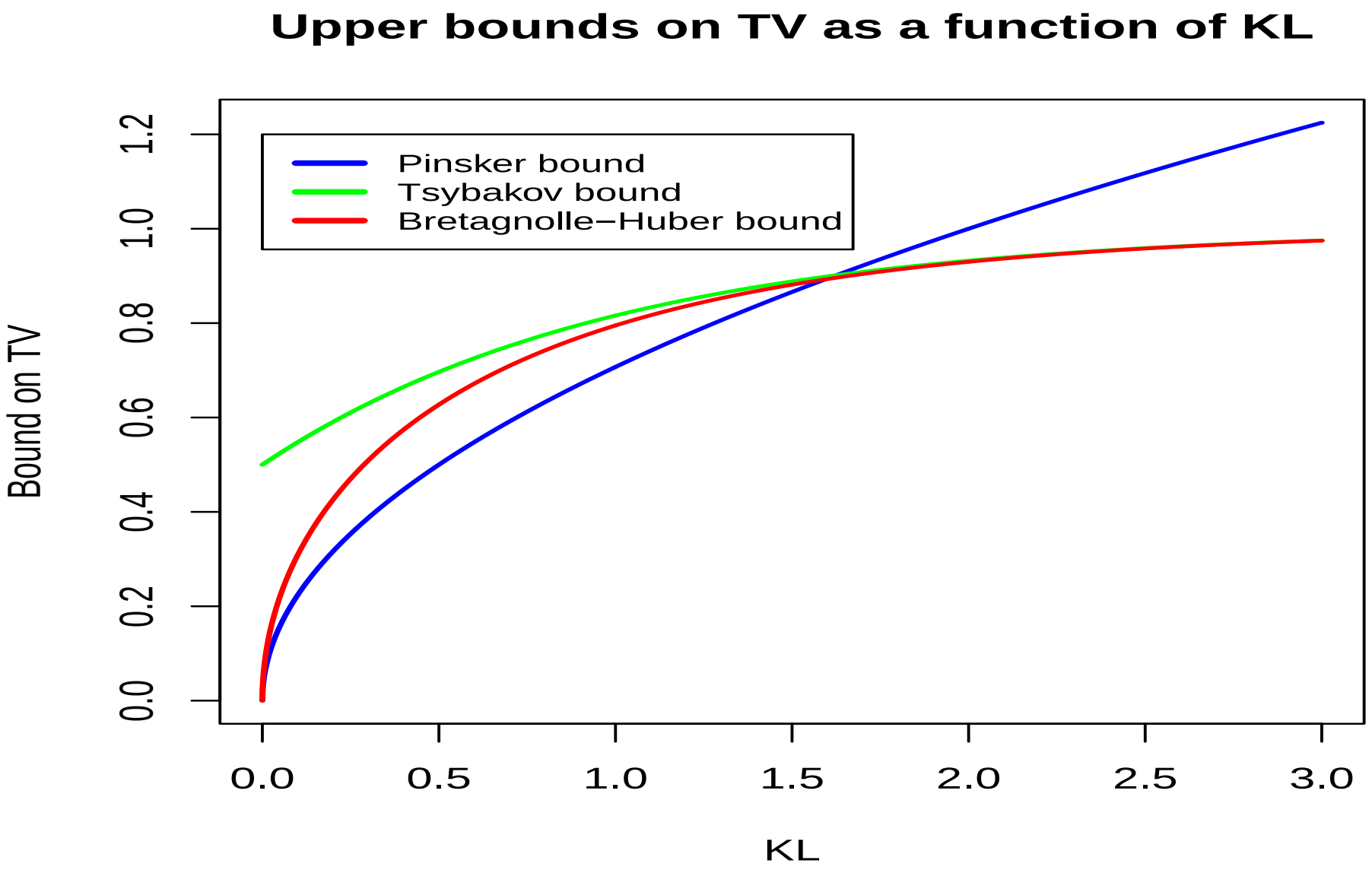


FIGURE 1 – Three bounds of TV distance with respect to Kullback distance

Some other ways to define discrepancy between two probabilities

1. **The p-Wasserstein distances**
   $\Gamma(P,Q)$: the set of probabilities on $E \times E$ having marginals P and Q.

$$W_p(P,Q) := \inf_{\gamma \in \Gamma(P,Q)} \{ \int_{E\times E} \|x-y\|_2^p \, d\gamma(x,y) \} \tag{8}$$

   Properties of W_p
   a. Characteristic : it incorporates the geometry of the domain.
   b. Associated with an optimal coupling of P, Q related to optimal tranport (Monge-Kantorovitch).
   c. Upper bounds easy : $W_p \leq \int_{E\times E} \|x-y\|_2^p \, d\gamma(x,y) \forall \gamma \in \Gamma(P,Q)$.
   d. W_2^2 easy for product measures : $W_2^2(\otimes_{i=1}^n (P_i,Q_i)) = \sum_{i=1}^n W_2^2(P_i,Q_i)$
   e. Useful for WGAN Neural Networks [5] :
   A Generative Adversarial Network (GAN) simultaneously trains two models, a generator and a discriminator :

5. Martin Arjovsky, Soumith Chintala, Leon Bottou, Wasserstein Generative Adversarial Networks, 2017

— the generator learns to output fake samples from an unknown distribution

— the discriminator learns to distinguish fake from real samples.

2. **The f divergences** [6] : $D_f(P,Q) := \int_E f(dP/dQ)\, dQ$

$$
\begin{aligned}
f(t) &= t\,\log(t) && \Rightarrow \text{ Kullback-Leibler, KL} \\
&= \tfrac{1}{2}(\sqrt{t}-1)^2 && \Rightarrow \text{ Hellinger } h^2 \\
&= |t-1| && \Rightarrow \text{ Total Variation, TV} \\
&= (t-1)^2 && \Rightarrow \text{ Pearson } \chi^2 \\
&= \frac{2(1-t^{(1-\alpha)/2})(1-t^{(1-\beta)/2})}{(1-\alpha)(1-\beta)} \rightharpoonup && \Rightarrow \text{ AB divergence}
\end{aligned}
$$

6. Cai, Yuhang and Lim, Lek-Heng, (2022), "Distances between probability distributions of different dimensions". IEEE Transactions on Information Theory, 68 :6, 4020-4031.

I. PARAMETRIC MODELS : ROBUSTNESS NEEDED

Motivation :

1. A random phenomenon is **known to obey a parametric model** : its probability is known up to a finite number of real numbers.
2. A **discrepancy between the probability of the phenomenon under study and the observations is unavoidable** due to gross errors and rounding errors. It can be represented by a distance and a corresponding neighborhood of the model.
3. **J.W. Tukey** showed that **optimal procedures for the strict model loose rapidly their good properties** even for an undetectable deviation.

Solution [7] :
Optimize the worst performance on a neighborhood of the model : find a minimax procedure. This can be done with a small loss for the strict model.

7. Peter Jost Huber, Robust Statistics, Wiley (1981).

Example 1 [8] (instability of optimal parametric estimators)

The mean $\overline{X} = (X_1 + \cdots + X_n)/n$ of n observations of $X \sim (1-\varepsilon)N(\theta, \sigma^2) + \varepsilon(N(\theta, 9\sigma^2))$ is an efficient ML estimator of θ for $\varepsilon = 0$ (unbiased, minimum variance). But its efficiency decreases down to 0.7 when ε increases from 0 to 0.10.

Contamination fraction ε	0.00	0.02	0.05	0.10
$\overline{X}_n$ efficiency	1.00	0.90	0.80	0.70

Any optimal estimator for **any** $\varepsilon \in [0.01; 0.10]$ has **efficiency** > 0.96.

8. Tukey, John Wilder. "A survey of sampling from contaminated distributions." Contributions to probability and statistics : 448-485, (1960).

Example 2 [9] (robustify a simple test $H_0 : P = P_0$ against $H_1 : P = P_1$.)

To **minimize the maximum loss over two neighborhoods $\mathcal{H}_0$ of P_0 and $\mathcal{H}_1$ of P_1, find a least favorable pair** [10] $(q_0, q_1) \in \mathcal{H}_0 \times \mathcal{H}_1$ i.e. such that

$$p_0(\frac{q_1}{q_0} > k) \leq q_0(\frac{q_1}{q_0} > k) \leq q_1(\frac{q_1}{q_0} > k) \leq p_1(\frac{q_1}{q_0} > k) \ \ \forall (p_0, p_1) \in \mathcal{H}_0 \times \mathcal{H}_1 \quad (9)$$

The optimal test of q_0 against q_1, based on the ratio q_1/q_0, **is minimax** as its performance for testing any $p_0 \in \mathcal{H}_0$ against any $p_1 \in \mathcal{H}_1$ is better than for testing q_0 against q_1.

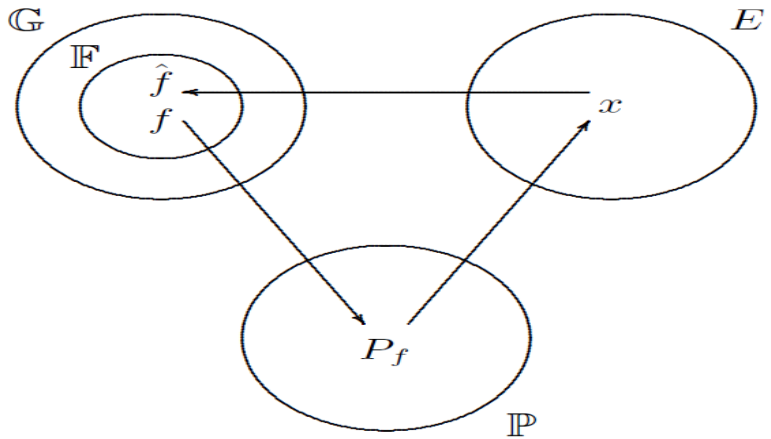
9. Huber-Carol, C. "Asymptotics of robust tests.", Thèse de doctorat, ETH Zurich, (1970)
10. Huber-Carol, C. Lecture Notes in Maths, 1215, "Robustness Theory", p.1-128, Springer Verlag, (1986)

II. NON PARAMETRICS : FUNCTION ESTIMATION

Framework :

1. f, unknown function, $f \in \mathbb{F}$, $\mathbb{F}$ a set of "smooth functions".
2. $\mathbb{P} = \{P_f : f \in \mathbb{F}\}$
3. $X \sim (P_f)^{\otimes n}$ has its values in a measurable space $(E, \mathcal{B})^{\otimes n}$.

f is to be estimated based on observation x of X.



f can be a probability density, the spectral density of a Gaussian process, the intensity of a Poisson process, the hazard rate of a positive random variable.

1. Best achievable rate of convergence [11]

It is obtained via the relationship between the distance D on $\mathbb{G} \supset \mathbb{F}$, and the corresponding distance on $\mathbb{P}$, and the **construction inside $\mathbb{F}$ of a finite set $\mathbb{F}_0$** (Assouad hypercube or Fano Pyramid) **to be discriminated**, shown to be **as difficult as the initial infinite dimensional problem** : [12]

Discrimination of two points distant $\Delta \in \mathbb{F}$

If $D(f_1, f_2) \geq \Delta$ and $U = D(\hat{f}, f_1)$ *and* $V = D(\hat{f}, f_2)$, then $U + V \geq \Delta$ (triangular inequality) leads to two inequalities :

$$\begin{aligned} E_P(U) + E_Q(V) &\geq \frac{\Delta}{2} \exp(-4h^2(P,Q)) \\ E_P(U) + E_Q(V) &\geq \frac{\Delta}{2} \exp(-4KL(P,Q)) \end{aligned}$$

leads to a lower bound for the risk of discrimination of the k equidistant points of $\mathbb{F}_0$, whose maximum risk is greater than the uniform bayesian risk.

11. Bretagnolle, Jean, and Catherine Huber. "Estimation des densités : risque minimax." Séminaire de probabilités de Strasbourg 12. 342-363 (1978)

12. Huber-Carol, Catherine. "A Cramer-Rao type inequality for estimating a hazard with censoring." 2017 Conference Lifetime Data Science on Precision Medicine and Risk Analysis with Lifetime Data. (2017)

2.Robust divergence BHHJ for function estimation :

BHHJ density power divergence [13] , is indexed by a positive parameter a :

$$D_a(P,Q) = \int \{dP^{1+a}(x) - (1+\frac{1}{a})dQ(x)dP^a(x) + \frac{1}{a}dQ^{1+a}(x)\}dx, \quad a \in (0,1)$$

a controls the trade-off between robustness and efficiency

$$BHHJ \xrightarrow[a\to 0]{} KL \Rightarrow \text{Maximum Likelihood,efficient}$$
$$BHHJ \xrightarrow[a\to 1]{} L^2 \Rightarrow \text{Mean square error, robust but not efficient}$$

The **small contribution of outliers to L^2 distance** based on histograms or kernel density estimates **makes this robustness intuitively apparent**.

13. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C., 1998. Robust and efficient estimation by minimising a density power divergence. Biometrika 85, 549–559.

III. BIOMEDICAL APPLICATIONS : 1. DIAGNOSIS

hierarchical log-linear models

Diagnosis on a sparse contingency table (most cells empty) [14] :

n	$=$	1000 patients		
$\boldsymbol{X}$	$p=9$	symptoms :	$\in \{0,1\}^p$	$\Rightarrow 2^9 = 512$ symptom profiles
M	$m=2$	diseases :	$\in \{0,1\}$	$\Rightarrow$ 1024 cells, most of them empty

$$A_{2\times 512} = \left[\underbrace{\begin{array}{cccc} n_{11} & n_{12} & \cdots\cdots\cdots & n_{1p} \\ n_{21} & n_{22} & \cdots\cdots\cdots & n_{2p} \end{array}}_{N=2^p=512}\right] \Bigg\} m = 2$$

$$\begin{array}{rcll} \log(P(\boldsymbol{X}=\mathbf{x}|M)) & = & C+ & \sum_{j=1}^{p} g_j(x_j) + \sum_{j\neq j'} g_{j,j'}(x_j, x_{j'}) \\ & + & \sum_{j\neq j'\neq k} & g_{j,j',k}(x_j, x_{j'}, x_k) + \cdots + g_{1,2,\cdots,p}(x_1, x_2, \cdots, x_p) \end{array}$$

where all expectations of g functions on any argument are 0. Keep interactions up to order k : cut off all functions of more than k arguments.

$k=1 \Rightarrow$ **independence of symptoms : easy but irrealistic**

$k=2 \Rightarrow$ **order 2 dependence only**

$k=3 \Rightarrow$ **influence of a third factor on the way two factors interact**

14. Huber, Catherine, and Joseph Lellouch. "Estimation in Sparse Contingency Tables." International Statistical Review,193-203, (1974)

Illustration on an artificial example, 2 diseases, 3 symptoms :

— Every symptom is present with probability $1/2$ in M_1 and in M_2 $\Rightarrow$ none of them is able alone to discriminate M_1 and M_2.

— Every pair $(Z_j, Z_{j'})$ is uniform on the 4 values for M_1 and M_2 $\Rightarrow$ none of the 3 pairs $(Z_j, Z_{j'})$ can discriminate M_1 and M_2.

— But **the three of them altogether lead to a perfect diagnosis.** :

$$M = M_1 \quad \Leftrightarrow \quad (Z_1, Z_2, Z_3) \quad \in \quad A := \{(0,0,0), (0,1,1), (1,0,1), (1,1,0)\}$$

$$M = M_2 \quad \Leftrightarrow \quad (Z_1, Z_2, Z_3) \quad \in \quad A^c := \{(0,0,1), (0,1,0), (1,0,0), (1,1,1)\}$$

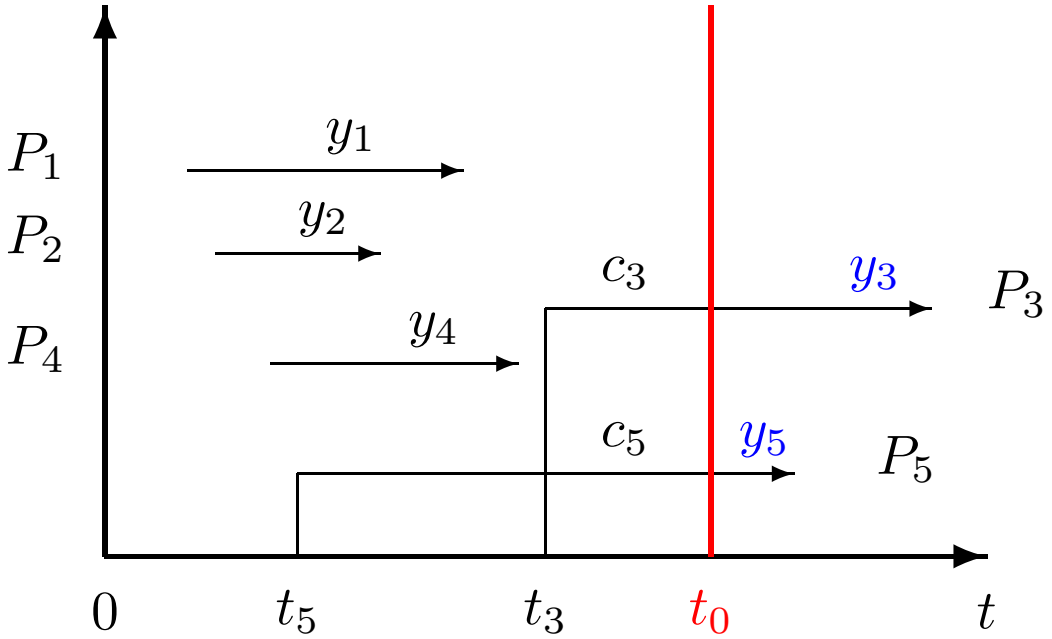
This will show again when dealing with the **explainability of neural networks**, (cf Shapley values) [15].

15. Owen, Art B., and Clémentine Prieur. "On Shapley value for measuring importance of dependent inputs." SIAM/ASA Journal on Uncertainty Quantification : 986-1002, 5.1 (2017).

2. SURVIVAL DATA ANALYSIS

Specificity of survival data : censoring and truncation

A simple example : survival times of 5 patients, end of the study at time t_0 : survival times y_1, y_2, y_4 of patients P_1, P_2, P_4 are observed :



P_3 and P_5 are still alive when the study stops at t_0 :
y_3 and y_5 are not observed, they are **right censored**. **Ignore them ?**
No, provide the information : $y_3 \geq c_3 := t_0 - t_3, \quad y_5 \geq c_5 := t_0 - t_5$

General Censoring and Truncation
A non parametric approach

Truncation of Y by the set B :

B truncates Y if Y is observed only if $Y \in B$.

Censoring of Y by the set A :

Y, not observed, is known to be in A.

Survival data imply three probabilities :

1. Censoring law : P_c
2. Truncation law : P_t
3. Survival law : P_s

Objective :
Estimate P_s in spite of the presence of **two nuisance infinite dimensional parameters** P_c and P_t.

Consistency and speed of convergence are obtained, under regularity assumptions, for the Non Parametric Maximum Likelihood Estimator (NPMLE) [16] [17] **of the density of** P_s, based on the **Hellinger bracketing entropy** :

$$H(\varepsilon, \mathcal{F}, h(\mu)) = \log(N_{[\,]})$$

where $\mathcal{F}$ is a set of densities on $(E, \mathcal{B}, \mu)$, $V(g^L, g^R) = \left\{g : \ g^L \leq g \leq g^R\right\}$ is bracketted by (g_l, g_R), and $N_{[\,]}(\varepsilon, \mathcal{F}, h(\mu))$ **is the smallest value of** m such that

$$\mathcal{F} \subset \bigcup_{j=1}^{m} V(g_j^L, g_j^R), \text{ where } h(g_j^L, g_j^R) \leq \varepsilon, \ j = 1, \ldots, m.$$

Analogous quantities for other distances, like L_2 for example, are defined :

$$H(\varepsilon, \mathcal{F}, L_2(\mu)) = \ln N_{[\,]}(\varepsilon, \mathcal{F}, L_2(\mu)).$$

16. Huber, Catherine, Valentin Solev, and Filia Vonta. "Interval censored and truncated data : Rate of convergence of NPMLE of the density." Journal of Statistical Planning and Inference 139.5 : 1734-1749, (2009).

17. Vonta, F., and C. Huber. "On the estimation of structural parameters in frailty models for interval censored and truncated data." Volume 14 No 4 14.4 : 40-49, (2010).

SEMI-PARAMETRIC SURVIVAL MODELS

Most usual models are based on **hazard rate h, the probability that the event takes place at time t, knowing that it did not take place before**

$$\mathbf{h}(t) = \frac{f(t)}{S(t)} \quad \text{where} \quad \begin{array}{lll} S(t) & = & P(Y \geq t) \quad \text{survival function} \\ f(t) & = & -S'(t) \quad \text{density function} \end{array}$$

1. **COX MODEL** [18] The hazard rate h is assumed to be equal to a baseline hazard $h_0(t)$ modified by p covariates $\boldsymbol{X} = (X_1, \cdots, X_p)$ whose weights are the **parameters** $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ to be estimated as well as h_0 [19] :

$$\boxed{h(t|\boldsymbol{X}) = h_0(t)\, \mathbf{e}^{\boldsymbol{\beta}^T \boldsymbol{X}}}$$

Baseline hazard h_0 : any function

18. Cox, David Roxbee, and David Oakes. "Analysis of survival data." Vol. 21. CRC press, 8th edition (1998)

19. Bretagnolle, Jean, and Catherine Huber-Carol."Effects of omitting covariates in Cox's model for survival data." Scandinavian journal of statistics : 125-138,(1988).

2. FIRST HITTING TIME model (FHT) or THRESHOLD REGRESSION model (TR) [20]

Threshold regression model : three different ways of acting on the time to onset of the disease for the potentially influential factors :

(a) **Initial covariates** : they act on the **"initial amount of health"** : gender, past family disease history, genetic factors,...

(b) **Lifetime covariates** : they act on (or testify for) the **evolution of the initial amount of health** : smoking habits, biological features, environment,...

(c) **Occupational exposure** : it may **accelerate the time to onset of the considered disease**

20. Lee, Mei-Ling Ting. "A survey of threshold regression for time-to-event analysis and applications." Taiwanese Journal of Mathematics 23.2, 293-305 (2019).

The model

The amount of health relative to the disease is a stochastic process $H(t)$:

$$H(t|h,\mu) = h + \mu t + B(t) \tag{10}$$

1. $h > 0$ the initial amount of health
   function of intial covariates.
2. $\mu < 0$ the slope of the process
   function of lifelength covariates
3. $B(t)$ a Brownian motion
   error term
4. $R(t)$ a non decreasing continuous function on $\mathbb{R}^+$
   measuring the acceleration due to occupational exposure
   (to asbestos in our case).

The time T to onset of the disease, is defined as the first time $H(R(t))$ hits 0 :

$$T(h,\mu,R) = \inf\{t \geq 0 : H(R(t)|h,\mu) \leq 0\} \tag{11}$$

Motivating example[21] :
Expected years of life free of lung cancer lost due to occupational exposure to asbestos on a French case-control study.
The data set
Between 1999 and 2002 in 4 Parisian hospitals, 860 cases, 901 controls, matched on gender and hspital.

1. Basic information : hospital, gender, past family disease history, tobacco, age at interview (calendar time), age at incidence of lung cancer,
2. Asbestos exposure : The occupational history up to age X is measured on each of the successive employments by duration, and probability/frequency/intensity of exposure, each with 3 levels.
3. Matching between diseased and controls was done on hospital, gender, age at interview.

21. Chambaz, Antoine, Dominique Choudat, and Catherine Huber-Carol. "Acceleration, due to occupational exposure, of time to onset of a disease." Theory and Practice of Risk Assessment, Springer International Publishing, 2015.

A partial result

gender	age	asbestos	family	tobacco	years lost
Male	65	228	0	1	3.1
Male	57	125	0	1	2.7
Male	60	25	0	1	1.6
Female	41	36	0	1	3.0
Male	66	24	1	1	1.4
Female	61	78	0	0	3.2

TABLE 1 – Expected number of years free of lung cancer lost due to occupational asbestos exposure.

IV NEURAL NETWORKS

A. SIMPLE NEURAL NETWORK

It has **a single neurons layer** and is a parametric version of a statistical semi-parametric process called PPRD (Projection Pursuit Regression and Discrimination).

PROJECTION PURSUIT (PPRD)

a. Regression

The target $Y \in \mathbb{R}$ is the response variable to $\boldsymbol{X} = (X_1, \cdots, X_d) \in \mathbb{R}^d$. The PPR $\widehat{Y}$ of Y is defined as :

$$\widehat{Y} = \widehat{f}(\boldsymbol{X}) := \sum_{m=1}^{M} \widehat{g_m}(\widehat{\mathbf{w}_m^T \boldsymbol{X}}) := \sum_{m=1}^{M} \widehat{g_m}(V_m) \tag{12}$$

$\mathbf{w}_m, m = 1, \cdots, M$ are unitary d-dimensional vectors and $g_m : \mathbb{R} \to \mathbb{R}$ ridge functions. Estimations based on an observed training set :$(\mathbf{x}_i, y_i), i = 1, \cdots, n$.

For M big enough, any function can be approximated by (12).

b. Discrimination : K categories

The response Y is one of K categories and the prediction $\widehat{f_k}(\mathbf{x}_i)$ is the probability of category k when $\mathbf{x} = \mathbf{x}_i$.

c. Error measurement : KL (Kullback-Leibler) for discrimination

$$\begin{array}{lll} R(\boldsymbol{\theta}) & := & \sum_{k=1}^{K} \sum_{i=1}^{n} (y_{ik} - \widehat{f_k(x_i)})^2 \quad \text{quadratic error} \\ R_{KL}(\boldsymbol{\theta}) & := - & \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log(\widehat{f_k(x_i)}) \quad \text{crossed entropy} \end{array}$$

d. Interpretation in terms of the initial inputs is difficult as each feature X_j is scattered into every linear combination of $\boldsymbol{X}$.

NEURAL NETWORK as a SPECIAL CASE of PPRD

Our framework is a discrimination problem : the target $\boldsymbol{Y} = (Y_1, \cdots, Y_K)$ is a category, each Y_k being a (0,1 variable) to be predicted by $\boldsymbol{X} = (X_1, \cdots, X_d)$
A. A layer of M neurons with entries $\boldsymbol{X}$ produces a prediction $\widehat{\boldsymbol{Y}}$ of $\boldsymbol{Y}$ using $(d+1) \times M$ coefficients α and $(M+1) \times K$ coefficients β :

$$\begin{array}{lcll} V_m & := & \alpha_0 + \boldsymbol{\alpha}_m^T \boldsymbol{X} & m = 1, 2, \cdots, M \\ Z_m & = & \sigma(V_m) & \sigma \text{ is the activation function} \\ T_k & = & \beta_{0k} + \boldsymbol{\beta}_k^T \boldsymbol{Z} & k = 1, 2, \cdots, K \\ f_k(\boldsymbol{X}) & = & g_k(\boldsymbol{T}), & k = 1, 2, \cdots, K \end{array}$$

where $g_k(\boldsymbol{T}) = \frac{e^{T_k}}{\sum_{i=1}^{K} e^{T_i}} \Rightarrow$ all $g_k(\boldsymbol{T})$ are positive and add to 1.

$$\boxed{\widehat{Y_k} := \widehat{f_k}(\boldsymbol{X})}$$

is the estimated probability of category k.
B. Minimize the error $R(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})$ by an optimal choice of the parameters $\mathbf{w} = (\boldsymbol{\alpha}, \boldsymbol{\beta})$, obtained by gradient descent of R with respect to $\mathbf{w}$.

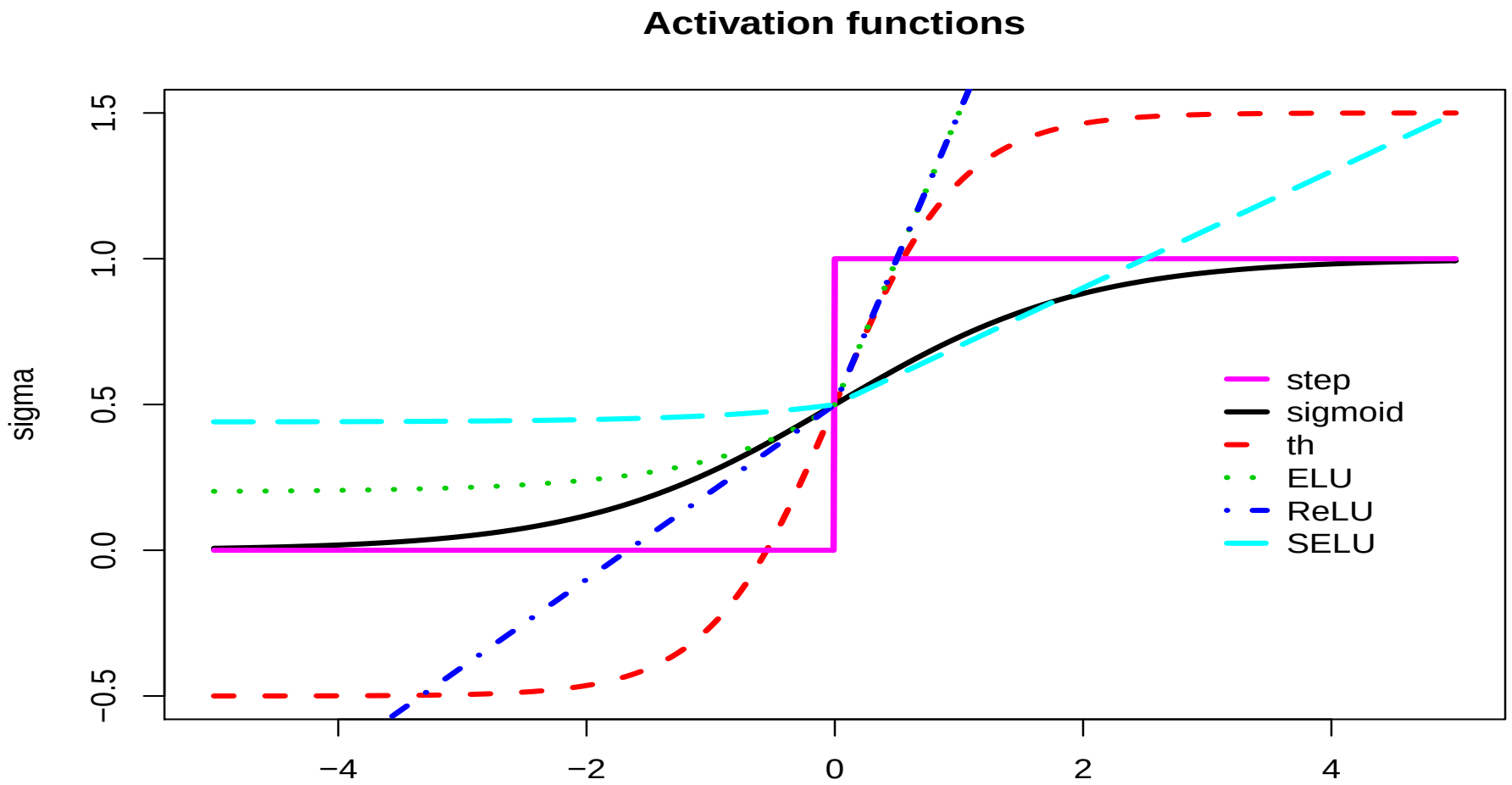


FIGURE 2 – Several activation functions

Possible choices for the activation function σ, smoothed versions of the step function $s(u) = 1\{u \geq 0\}$:

$$\sigma(u) = \frac{1}{1+e^{-u}} \quad \text{the sigmoïd, the most usual one}$$

$$\sigma(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}} \quad \text{hyperbolic tangent (th(u))}$$

$$\sigma(a,u) = \begin{cases} a(e^u - 1) & \text{for } u < 0 \\ u & \text{for } u \geq 0 \end{cases} \quad \text{Exponential Linear Unit (ELU)}$$

$$\sigma(a,u) = \begin{cases} au & \text{for } u < 0 \\ u & \text{for } u \geq 0 \end{cases} \quad \text{Rectified Linear Unit (ReLU)}$$

$$\sigma(a,b,u) = b\begin{cases} a(e^u - 1) & \text{for } u < 0 \\ u & \text{for } u \geq 0 \end{cases} \quad \text{Scaled Exponential Linear Unit (SELU)}$$

IV COMPARING PREDICTION and INTERPRETATION for GLM and NN

A. Simulation of a logistic model

1. The simulation :
   - 6 simulated risk factors
     - 3 **relevant** risk factors are $\boldsymbol{X} = (X_1, X_2, X_3)$
       - X_1, binomial(p=0.3,size=3), coefficient $a_1 = 1$,
       - X_2, exponential(1), coefficient $a_2 = 2$,
       - X_3, Poisson($\lambda = 3$), coefficient $a_3 = -1$.
     - 3 **irrelevant** risk factors are $\boldsymbol{Z} = (Z_1, Z_2, Z_3)$ independent of Y
       - Z_1, binomial(p=0.5, size=2), coefficient $b_1 = 0$,
       - Z_2, normal($\mu = 3, sd = 1$), coefficient $b_2 = 0$,
       - Z_3, Poisson($\lambda = 5$), coefficient $b_3 = 0$.
   - The model (including a normal error $\varepsilon \sim \mathcal{N}(0, 0.1)$) :

$$\ln\left(\frac{P(Y=1|\boldsymbol{X}=\mathbf{x},\boldsymbol{Z}=\mathbf{z})}{P(Y=0|\boldsymbol{X}=\mathbf{x},\boldsymbol{Z}=\mathbf{z}))}\right) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \varepsilon \quad (13)$$

2. **Prediction performances of GLM (the true model) and NN** :

   — Size of training set : $(2/3)\ n$, leaving $(1/3)\ n$ for the test set

   — Respective correct prediction probabilities for diseased (p_d), non diseased (p_{nd}) and global (p_g) on the test set :

Method	p_d	p_{nd}	p_g	$CI_{95\%}(p_d)$	$CI_{95\%}(p_{nd})$
GLM	0.833	0.752	0.788	0.827 0.838	0.746 0.758
NN	0.857	0.752	0.808	0.849 0.864	0.742 0.762

TABLE 2 – Probability of correct predictions due to GLM and NN for diseased (p_d), for non diseased (p_{nd}), global p_g and 95% confidence intervals

Conclusion :

Probabilities of correct prediction are similar for GLM and NN.

3. Interpretation of risk factors impact by GLM

— **GLM estimates the weight of every risk factor** in **x** and **z** :

Risk factor	True coeff	coeff by GLM	p-value
x_1	1	1.06	10^{-10}
x_2	2	2.04	$5 * 10^{-27}$
x_3	-1	-1.03	$5 * 10^{-28}$
z_1	0	-0.30	0.23
z_2	0	0.09	0.40
z_3	0	0.10	0.050

TABLE 3 – Respective weights of risk factors **x** (relevant) and **z** (irrelevant) with corresponding p-values

The weight (the relative importance) of x_2 is the highest.

— **Interpretation of risk factors impact by NN**

Before permuting each factor in turn, the mean probability to predict correctly D is

$$p_d = 0.857 \qquad 95\%\ CI = [0.849, 0.864]$$

After permutation of each factor in turn the mean probability of correct prediction **decreases for relevant factors**, and **is stable for irrelevant ones** :

m.x_1	m.x_2	m.x_3	**relevant** factors
0.842	0.762	0.787	< 0.849
m.z_1	m.z_2	m.z_3	**irrelevant** factors
0.857	0.856	0.855	≈ 0.857

TABLE 4 – Mean correct probability of prediction of occurrence of the disease p_d when doing N=100 permutations of each risk factor $x_1, x_2, x_3, z_1, z_2, z_3$.

Conclusion :

Relevant factors are identified by NN as well as by GLM. Moreover, **the most influent factor is again x_2 : its permutation leads to the highest decrease** of the probability of correct prediction.

B. Alzheimer data :

1. **Description of the data set** (from Pitie Salpetriere Hospital, Paris)
   - $n = 4356$ patients, $n_1 = 142$ developed an Alzheimer within 4 years.
   - 13 risk factors : age at inclusion, gender, education, cardiac disease, depress, incapacity, high blood pressure, birth date, three genetic factors, psychological disease,
   - Objective : how to predict who will develop an Alzheimer ?
   - **Compare neural network (NN) with classical logistic model (GLM)**

$$P(Y = 1|\boldsymbol{X} = \mathbf{x}) = \frac{\exp(\mathbf{w}^T\mathbf{x})}{1 + \exp(\mathbf{w}^T\mathbf{x})}$$

   - The very unbalanced counts for diseased (142) and controls (4214) creates difficulties for prediction which can be overcome by duplication of the diseased [22].

22. Yann Le Cun, personal communication, 2019

2. **Prediction performances of GLM and NN for Alzheimer :**

Method	p_d	p_{nd}	p_g	$CI_{95\%}(p_d)$	$CI_{95\%}(p_{nd}))$
GLM	0.73	0.73	0.73	0.71 0.76	0.71 0.75
NN	0.75	0.77	0.76	0.74 0.76	0.76 0.78

TABLE 5 – Correct predictions due to GLM and NN for dements (p_d), for non dements (p_{nd}), global p_g, and 95% confidence intervals after duplication.

3. **Interpretation for GLM and NN**
   — GLM gives an estimation of the weight of every risk factor : age is compared to age < 70

Age $\in [70\ 80]$	:	risk multiplied by	3	$(CI_{95\%} = [1.6\ 5.9]$
Age > 80	:		8	$(CI_{95\%} = [4.3\ 16]$
Cardiac disease	:		2	$(CI_{95\%} = [1.2\ 2.9]$
Depress	:		2.5	$(CI_{95\%} = [1.5\ 3.3]$
Incapacity	:		3.5	$(CI_{95\%} = [2.2\ 5.1]$
APOE4	:		2	$(CI_{95\%} = [1.3\ 2.8]$

— NN : Risk factors impact for Neural Networks

Permutation	p_d	p_{nd}	p_g	$CI_{95\%}(p_d)$	$CI_{95\%}(p_{nd}))$	
none	0.7553	0.7739	0.7650	0.7412 0.7634	0.7662 0.7798	
gene AA	0.7419	0.7724	0.7581	0.7395 0.7442	0.7699 0.7749	$\approx$
gene AG	0.7457	0.7751	0.7613	0.7418 0.7495	0.7717 0.7786	$\approx$
age	0.7098	0.7410	0.7264	0.7057 0.7139	0.7338 0.7481	$\downarrow$
APOE4	0.7341	0.7629	0.7494	0.7289 0.7393	0.7594 0.7665	$\downarrow$
cardiac disease	0.7446	0.7748	0.7606	0.7401 0.7491	0.7721 0.7775	$\approx$
gene CC	0.7473	0.7779	0.7635	0.7428 0.7518	0.7747 0.7811	
depress	0.7381	0.7671	0.7535	0.7343 0.7420	0.7636 0.7706	$\downarrow$
education	0.7473	0.7772	0.7632	0.7444 0.7503	0.7748 0.7797	$\approx$
gender	0.7447	0.7758	0.7612	0.7403 0.7490	0.7725 0.7792	$\approx$
Hypertension	0.7510	0.7808	0.7668	0.7457 0.7564	0.7765 0.7852	$\approx$
Incapacity	0.7282	0.7609	0.7455	0.7243 0.7320	0.7584 0.7634	$\downarrow$
psychotropes	0.7419	0.7724	0.7581	0.7395 0.7442	0.7699 0.7749	$\approx$
gene TC	0.7465	0.7773	0.7628	0.7450 0.7480	0.7748 0.7799	$\approx$

TABLE 6 – Effect, on prediction ability, of permutation of each risk factor.

CONCLUSIONS and PERSPECTIVES

1. **Prediction and interpretation**

   (a) **Prediction performances** :
   similar in our case of moderate size data.

   (b) **Interpretation** :
   — Easy for **linear** models in statistics (GLM) :
   influence of each factor measured by its estimated coefficient.
   But **it fails in our artificial diagnosis example** while NN succeeds.

   — Uneasy for **non linear** approaches :
   NN (in AI), a parametric version of PPRD
   Semi-parametric PPRD model (in statistics) :
   The model changes when the vicinity of the explanatory variables (the entries) changes. This leads to have **global and local explanations**.

2. **Two important remarks**

(a) **NN may be implemented to solve statistical models**

An example is **Cox model revisited by Neural Networks** [23]

A NN is used to minimize a function analog to $-\mathcal{L}_c$ but where the linear function $\mathbf{w}^T\mathbf{x}$ is replaced by a nonlinear one $h_\theta(\mathbf{x})$:

$$\mathcal{L}_c(\mathbf{w}) = \prod_i \delta_i \frac{\mathbf{e}^{\mathbf{w}^T\mathbf{x}_i}}{\sum_{j:t_j \geq t_i} \mathbf{e}^{\mathbf{w}^T\mathbf{x}_j}}$$

$$\mathcal{L}_{NN}(\theta) = -\prod_i \delta_i \frac{\mathbf{e}^{h_\theta(\mathbf{x}_i)}}{\sum_{j:t_j \geq t_i} \mathbf{e}^{h_\theta(\mathbf{x}_j)}}$$

The loss function minimized by the NN with parameters θ is $-\mathcal{L}_{NN}(\theta)$.

23. Katzman, Jared L., et al. "DeepSurv : personalized treatment recommender system using a Cox proportional hazards deep neural network." BMC medical research methodology 18.1 : 1-12, (2018).

(b) **NN take care of big data and overparameterization**

i. **Classical statistics need to reduce the dimension of big data**
Numerous devices (**most are linear**) :
PCA (Principal Component Analysis), **SVD** (Singular Value Decomposition), **MDS** (MultiDimensional Scaling).

ii. **Classical statistics need to penalize overparameterization**
In classical parametric statistics, the model $\mathcal{P} := \mathcal{P}_\Theta$ is defined up to a set of parameters $\theta \in \Theta$; increasing the number p of parameters may lead to a **perfect fit** to the training set which may **decrease the predictive ability** on a new sample : a penalization is applied, **Lasso (L^1 norm)** or **ridge (L^2 norm) penalizations**.

iii. **Overparameterization seems to cause no major problem to NN**

It has been observed that, in deep learning, one can simultaneously
— fit perfectly the training set (empirical risk equals 0),
— have an efficient predictive ability on a new sample.
In a recent paper [24], the authors have a theoretical proof of this surprising phenomenon in a special case (p. 36-40, a two layers network) under certain conditions.

3. **Importance of the nonlinearity**
   — **Role of the activation function**.
   The nonlinearity of the NN approach is due to the activation function σ.
   — **Nonlinear reduction method : Isomap**
   In a statistical setting, among the numerous devices whose purpose is to reduce the dimension (PCA, SVD , MDS) **most of them are linear**.

24. P.L. Bartlett, A. Montanari, A. Rakhlin, "Deep Learning : a statistical viewpoint", arXiv, 89 pages, March 16, (2021).

However, **based on the K nearest neighbours** $(j_1, j_2, \cdots, j_K)$ of every point i in the input space $\mathcal{X}$, assumed to be a metric space, ($\mathbb{R}^d$ in general), **a weighted graph is built**, the weight of each edge (i, j_k) being equal to $d(i, j_k)$, **and a geodesic distance** :

the **geodesic distance** of any pair of points (i, j) in the graph being the length of the **minimum path between them**.

This leads to discover the structure of the data, which may be a manifold rather than a linear subspace as is the case in PCA, SVD and also MDS, which constitutes my present research (SLALOM : Statistical Learning and Low Order Manifolds).

References

1. Bartlett, P.L., Montanari, A., Rakhlin A. (2021)*Deep learning : a statistical viewpoint*, arXiv preprint, arXiv :2103.09177.
2. Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C., (1998). *Robust and efficient estimation by minimising a density power divergence*. Biometrika 85, 549–559.
3. Cai, Yuhang and Lim, Lek-Heng, (2022), *Distances between probability distributions of different dimensions*. IEEE Transactions on Information Theory, 68 :6, 4020-4031.
4. Cox, D.R., Oakes, D. (1998), 8th edition. *Analysis of Survival Data*, Monographs on Statistics and Applied Probability 21, London : Chapman & Hall/CRC.
5. Donoho, D.L. (2018) *Data Science : The end of theory ?*, Vienna Conference.
6. Garson, G. David (1991) *A comparison of neural network and expert systems algorithms with common multivariate procedures for analysis of social science data*, Social Science Computer Review, **9(3)**,399–434.

7. Giudici, Paolo, Raffinetti, Emanuela (2021) *Shapley-Lorenz explainable artificial intelligence, Expert Systems with Applications*, 167 :114104.
8. Hastie, T., Tibshirani, R., Friedman, J. (2017), 2nd edition. *The Elements of Statistical Learning (Data mining, Inference, Prediction)*, Springer Series in Statistics.
9. Huber, C., Gross, S., Vonta, F. (2019) *Risk analysis : Survival data analysis vs Machine Learning. Application to Alzheimer prediction*, CRAS, CR Mécanique 347, 817-830.
10. Huber, C., Nikulin, M., Edts., Stochastic models in Survival Analysis and Reliability set, Wiley :
    — 2016 *Reliability of Engineering Systems and Technological Risks*, Vladimir Rykov.
    — 2017 *Stochastic Risk Analysis and Management*, Boris Harlamov.
    — 2017 *Chi-squared Goodness-of-fit Tests for Censored Data*, Mikhail Nikulin and Ekaterina Chimitova.
11. Katzman J.L. et al(2018) *DeepSurv : personalized treatment recommender system using a Cox proportional hazards deep neural network*, BMC Medical Research Methodology.

12. Lee M-LT, Gail M, Pfeiffer R, Satten G, Cai T, Gandy A, editors,(2013) *Risk Assessment and Evaluation of Predictions*, Springer.

13. Mattheou, K., Karagrigoriou, A. et al, (2008), *A model selection criterion based on the BHHJ measure of divergence.* J. Statist. Plann. Inference.

14. Owen, Art B., Prieur, Clemenine (2017). *On Shapley value for measuring importance of dependent inputs*, arXiv.

15. Panaretos, Victor M., Zemel, Yoav. (2019). *statistical, aspects of Wasserstein distances*, Annual review of statistics and its application, vol 6 :405-431.

16. Shapley, (1953). *A value for n-person games*, Princeton University Press.

17. Vonta, Filia and Karagrigoriu, Alex. (2010), *Generalized measures of divergence in survival analysis and reliability*, Journal of applied probability, vol 47 :1, 216–234, Cambridge University Press.

18. Li, Xuhong, et al. (2023)*G-LIME : Statistical learning for local interpretations of deep neural networks using global priors.* Artificial Intelligence 314 : 103823.

19. Zhang, Zhongheng, et al. (2018) *Opening the black box of neural networks : methods for interpreting neural network models in clinical applications.* Annals of translational medicine 6.11.

MERCI !