

Risk analysis: Survival data analysis vs. machine learning. Application to Alzheimer prediction

Catherine Huber-Carol, Shulamith T. Gross, Filia Vonta

▶ **To cite this version:**

Catherine Huber-Carol, Shulamith T. Gross, Filia Vonta. Risk analysis: Survival data analysis vs. machine learning. Application to Alzheimer prediction. Comptes Rendus Mécanique, 2019. hal-02446952

HAL Id: hal-02446952

https://u-paris.hal.science/hal-02446952

Submitted on 21 Jul 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S1631072119301780
Manuscript_ea844c7fb3ed89f35eb1309766edf739

Risk analysis: Survival data analysis vs Machine Learning. Application to Alzheimer prediction.

Catherine Huber-Carol [a] Shulamith Gross [b] Filia Vonta [c]

[a] *MAP5 CNRS 8145 @ Université de Paris*
45 rue des Saints-Pères, F-75270 Paris Cedex 06, France.
[b] *Lab VC-170 @ Baruch College of CUNY*
One Baruch way, NY, NY 10010, USA.
[c] *Department of Mathematics @ National Technical University of Athens*
9 Iroon Polytechneiou Str. 15780, Athens, Greece.

Abstract

We present here the statistical models that are most in use in survival data analysis. The parametric ones are based on explicit distributions, depending only on real unknown real parameters, while the preferred models are semi-parametric, like Cox model, which imply unknown functions to be estimated. Now, as big data sets are available, two types of methods are needed to deal with the resulting curse of dimensionality including non informative factors which spoil the informative part relative to the target: on one hand, methods that reduce the dimension while maximizing the information left in the reduced data, and then applying classical stochastic models; on the other hand algorithms that apply directly to big data, i.e. artificial intelligence (AI or machine learning). Actually, those algorithms have a probabilistic interpretation. We present here several of the former methods. As for the latter methods, which comprise neural networks, support vector machines, random forests and more (see second edition, January 2017 of Hastie, Tibshirani et al [18]), we present the neural networks approach. Neural networks are known to be efficient for prediction on big data. As we analyzed, using a classical stochastic model, risk factors for Alzheimer on a data set of around 5000 patients and $p = 17$ factors, we were interested in comparing its prediction performance with the one of a neural network on this relatively small sample size data.

Key words: Survival data analysis, Stochastic models, Machine learning, Neural networks, Nonlinear modeling, Alzheimer disease.

Email addresses: Catherine.huber@parisdescartes.fr (Catherine Huber-Carol), Shulamith.Gross@baruch.cuny.edu (Shulamith Gross), vonta@math.ntua.gr (Filia Vonta).

© 2019 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

1. Introduction

When analyzing the risk of an event E to occur, such as a degradation, a failure, a disease or even death, one may consider how the waiting time Y of onset of such a nocuous event is influenced by intrinsic and environmental factors $\boldsymbol{X} := (X_1, .., X_p)$:

$$Y = f(\boldsymbol{X}) \tag{1}$$

The function f is not deterministic. In classical survival data analysis, a stochastic model for f is chosen among several families of models, fully parametric, nonparametric or semi-parametric [31,1]. Then, f is to be estimated, based on the chosen model and from the observation of a training set, i.e. a sample of n observations, $(\boldsymbol{X}, Y)_i$ $i = 1, \cdots, n$, of the pair $(\boldsymbol{X}, Y)$. Finally, the stochastic model initially chosen has to be tested for fit; see E. J. Lehmann [37], and [27]. The classical versions of these models are available in R software. To adapt the analysis to specific situations, researchers have to elaborate extensions of these models and work them out using R, which is both a software and a programming language; see [43,45]. As a counterpart, the machine learning approach of this same problem consists in using an algorithm which has the potential risk factors $\boldsymbol{X}$ as entries and Y as output. Several families of algorithms are available: neural networks, random forests or support vector machines [18,33]. This second approach leads to the so-called "data driven models". In that respect, it seems to be more satisfactory than the subjective choice of a stochastic model that appears in the first approach. However, machine learning is often viewed as a "black box" as the algorithm goes back and forth until convergence is achieved, and it scatters thus the initial potential risk factors in such way that interpretation becomes difficult. Also, every machine learning method, even though it seems to be purely algorithmic, has a probabilistic interpretation. We shall see this feature in particular for neural networks, which are a parametric version of a stochastic model: the projection pursuit regression and discrimination model. Several recurrent problems occur when dealing with duration data:

(i) Incomplete data occur very frequently. For some subjects or items, the event does not occur before the end of the study, leading to missing data that are called "censored" data. More precisely, they are called "right censored data" as other phenomenons may occur like left censoring or interval censoring [25]. Right censoring means that the true value of Y is unknown but known to be "to the right" of the observed duration. Those incomplete data are not to be thrown away. They are taken into account using special devices.

(ii) Specially in medicine, it may happen that the training sample is small. If the sample size is small, it does not help in drawing conclusions. This happens with prospective studies. The simplest example of this case is a clinical trial: only one risk factor, the treatment. It compares a new treatment $X = 1$ to the usual one $X = 0$ to increase the life length of patients suffering from a specific disease. For ethical reasons, the sample size n is rather small and the patients are carefully chosen as the characteristics of the two groups should be comparable except for the treatment. This is achieved either by randomization (a double blind study, an experimental procedure in which neither the subjects nor the experimenters know which subjects are in the test and control groups), or pairing on possibly relevant factors [12] in order to ensure the comparability of the two sub-samples except for the treatment. For ethical reasons, the size of such samples is necessarily rather small. We should notice that a perfect comparability is achieved in the randomization case only asymptotically, and in the pairing case only if the set of matched factors is correctly chosen. In order to work with an increased sample size, one may use bootstrap techniques [14], which consist in drawing at random new samples out of the unique sample we have. Several theorems are needed to justify, in each specific case, the use of such duplications of a unique sample; see P. Hall [16].

(iii) But when the training set is not carefully collected on purpose like in the preceding example, one has to deal frequently with immense data bases, even in the medical field. Sample size n and/or

number of factors p may be very big. It happens when the data are obtained retrospectively, for example from an internet national or international data base. In France, big data sets of patients, with extensive information, are available from the national health insurance. Depending on the goal aimed at, there is a need to extract the useful information, reducing thus the high dimensionality of the data.

The next section, section 2, deals with several types of stochastic survival models. Section 3 gives an illustration of the difficulty of model selection. In section 4, we present some data reduction methods for the high dimensionality of data sets and in section 5 we define neural networks as a parametric version of a stochastic model: projection pursuit regression and discrimination. Finally, in section 6, we give an example of comparison of a stochastic model and a neural network to predict Alzheimer occurrence among patients at Pitié-Salpêtrière Hospital in Paris (France).

2. Stochastic survival models

The probability distribution of the waiting time Y may be defined by anyone of five different equivalent functions: its survival function $S(t) := P(Y \geq t)$, its distribution function $F(t) := P(Y < t)$, its density $f(t) := F'(t) := -S'(t)$ (whenever it is assumed to exist, which is mostly the case), its hazard function $h(t) := f(t)/S(t)$ and, finally, its integrated hazard $H(t) := \int_0^t h(s)ds$.
The relationships between them are

$$\begin{cases} S(t) = & 1 - F(t) = \exp(-H(t)) \\ h(t) = & H'(t) \ ; \ H(t) = -\ln(S(t)) \end{cases} \tag{2}$$

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

$$\begin{cases} h(t) = & \dfrac{f(t)}{S(t)} \\ f(t) = & -S'(t) \end{cases} \tag{3}$$

2.1. *Parametric survival models*

(i) Weibull and generalized Weibull [40]

(a) Weibull model
One of the most usual simple parametric model is due to Weibull. It has two positive real parameters λ and α, to be estimated from a training set. Weibull model $W(\lambda, \alpha)$ is defined as

$$h(t|\lambda, \alpha) = \alpha\lambda^\alpha t^{\alpha-1} \ ; \ (\lambda, \alpha > 0) \ ; \ t \geq 0 \tag{4}$$

$\alpha = 1$ h is constant $h = \lambda$ (no ageing, $\equiv$ exponential model $\mathcal{E}(\lambda)$ as $S(t) = exp(-\lambda t)$)

$0 < \alpha < 1$ h is decreasing $\infty \downarrow 0$ (see Figure 1 left)

$\alpha > 1$ h is increasing $0 \uparrow \infty$ (see Figure 1 right).

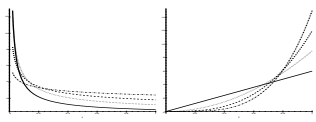


Figure 1. Weibull hazards for diverse values of α (shape), λ (scale)

As will be seen below, h is assumed to be a function of p real risk factors X_j named covariates in medicine and stresses in industrial setting.

(b) Generalized Weibull $GW(\lambda, \alpha, \gamma)$
It has a third parameter γ and reads

$$h(t|\lambda, \alpha, \gamma) = \frac{\lambda\alpha}{\gamma}(\lambda t)^{\alpha-1}\{1 + (\lambda t)^{\alpha}\}^{1/\gamma-1}; \quad (\lambda, \alpha, \gamma > 0); \quad t \geq 0 \tag{5}$$

GW allows multiple hazard shapes; see Figure 2

$$\begin{aligned}
&GW(\lambda, 1, 1) && = \mathcal{E}(\lambda), && \text{Exponential}\\
&GW(\lambda, \alpha, 1) && = W(\lambda, \alpha), && \text{Weibull}\\
&\text{If } \alpha > 1, \alpha > \gamma, && h \ : \ 0 \uparrow \infty\\
&\text{If } \alpha = 1, \gamma < 1, && h \ : \ \tfrac{\lambda}{\gamma} \uparrow \infty\\
&\text{If } 0 < \alpha < 1, \alpha < \gamma, && h \ : \ \infty \downarrow 0\\
&\text{If } 0 < \alpha < 1, \alpha = \gamma, && h \ : \ \infty \downarrow \lambda
\end{aligned}$$

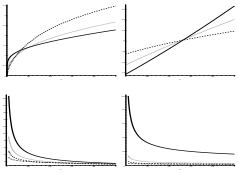


Figure 2. Generalized Weibull hazards for diverse values of the parameters

(c) Exponentiated Weibull $EW(\lambda, \alpha, \gamma)$
Exponentiated Weibull hazard [11,38] reads

$$h(t|\lambda, \alpha, \gamma) = \frac{\lambda\alpha\{1 - \exp[-(\lambda t)^{\alpha}]\}^{(1-\gamma)/\gamma}\exp[-(\lambda t)^{\alpha}](\lambda t)^{\alpha-1}}{\gamma\{1 - (1 - \exp[-(\lambda t)^{\alpha}])\}^{1/\gamma}}, \ (\lambda, \alpha, \gamma > 0), \ t \geq 0 \tag{6}$$

All moments of EW are finite. Sub-models are $EW(\lambda, \alpha, 1) = W(\lambda, \alpha)$ and $EW(\lambda, 1, 1) = \mathcal{E}(\lambda)$.
For $\alpha > 1$, $\alpha \geq \gamma$, the hazard h increases from 0 to ∞.
For $\alpha = 1$, $\gamma \leq 1$, the hazard h increases from $(\frac{\lambda}{\gamma})$ to ∞.
For $0 < \alpha < 1$, $\alpha < \gamma$, the hazard h decreases from ∞ to 0.
For $0 < \alpha < 1$, $\alpha = \gamma$, the hazard h decreases from λ to 0.

(ii) General parametric models
The waiting time Y is defined through a link function g and a random variable Z:

$$g(Y) = \boldsymbol{\beta}^T \mathbf{x} + \sigma Z \tag{7}$$

where $\mathbf{x}$ are the risk factors, $\boldsymbol{\beta}$ the parameters to be estimated, g identity or log, and density f of Z may be

$$\begin{cases} \text{logistic} : & f(z) = & \dfrac{e^{-z}}{(1+e^{-z})^2} \\ \text{normal} : & f(z) = & \dfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \\ \text{extreme} : & f(z) = & e^{z-e^z} \end{cases} \tag{8}$$

As an example, let us see how Weibull is represented in this setting: if Y is Weibull $W(\lambda, \alpha)$, $g = \log$ and $Z \sim \mathcal{E}(1)$, exponential(1) so that

$$\log(Y) = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} + \sigma \log(Z) = -\log(\lambda) + \frac{1}{\alpha} \log(Z) \tag{9}$$

In terms of $W(\lambda, \alpha)$ (4): $\lambda = e^{-(\beta_0 + \boldsymbol{\beta}^T \mathbf{x})}$, $\alpha = 1/\sigma$.

(iii) Estimation method: Maximum Likelihood (ML)
What is known is the observed training set $(y_i, \mathbf{x}_i)_{i=1,\cdots,n}$, and a stochastic model $\mathcal{P} = \mathcal{L}(Y|\boldsymbol{X})$, the probability distribution of Y knowing $\boldsymbol{X}$, implying unknown parameters to be estimated from the training set.

The unknown parameters are

— p parameters: $\boldsymbol{\beta} = \beta_1, \cdots, \beta_p$ quantifying the weight of each risk factor $X_j, j = 1, \cdots, p$

— the baseline hazard function $h_0(t)$ the probability, having "survived" up to time t, to have the event at time t when no risk factors is present. If h_0 is assumed to be a given function of k parameters, like a Weibull depending on $k = 2$ parameters, we have, in all, $p + 2$ parameters to estimate.

The values chosen for the parameters are those that maximize the probability of the observed training set, called Maximum Likelihood (ML) estimators. It should be mentioned that one has to be careful when using maximum likelihood looking at the ratio of the parameters to estimate as compared to the size of the training sample which should not exceed some value [23]
As the likelihood of one observation, $\mathcal{L}(\alpha, \lambda, \boldsymbol{\beta}|y, x_1, x_2, \cdots, x_p)$ is equal to $f(y, \alpha, \lambda, \boldsymbol{\beta}, \mathbf{x})$, and all observations are assumed to be independent, the likelihood of the observed training set is equal to the product $\mathcal{L} = \prod_{i=1}^n \mathcal{L}(\alpha, \lambda, \boldsymbol{\beta}|y_i, \mathbf{x}_i) = \prod_{i=1}^n f(y_i, \alpha, \lambda, \boldsymbol{\beta}, \mathbf{x}_i)$.
Now, find the values of $\alpha, \lambda, \boldsymbol{\beta}$ that maximize the likelihood $\mathcal{L}$:

$$\hat{\alpha}, \hat{\lambda}, \hat{\boldsymbol{\beta}} = \underset{\alpha, \lambda, \boldsymbol{\beta}}{\operatorname{argmax}} \ \mathcal{L}(\alpha, \lambda, \boldsymbol{\beta}|\mathbf{y}, \mathbf{x}) \tag{10}$$

Actually as it is easier to find the derivative of a sum than of a product, in order to maximize an expression, one deals with the log of the likelihood (named log-likelihood) rather than the likelihood itself.

(iv) Now, how to deal with right censored data?
Due to the nature of the data, it may happen that Y is not observed when the experiment stops at time C before the event takes place. One can then provide the fact that what is known is that $Y > C$ and replace in the likelihood, the density f at the unknown time Y by the survival S at the known time C. Other types of *censoring* may occur besides right censoring: left censoring, when the duration is known to be bigger than some observed duration C, interval censoring, when the duration is known to lie between two observed values, C_1 and C_2. Moreover, what may also happen is *truncation*: some items or patients may be skipped from the observed sample due to the experimental scheme. In that case, some special procedures have to be applied [13,24,20].

2.2. ***Semi-parametric survival models***

Semi-parametric models have a non-parametric part (an "infinite dimensional" unknown parameter: e.g. one or several functions) and a parametric one (a finite dimensional unknown real parameter), both to be estimated through a training set, but focussing on the parametric part, for the sake of interpretation.

(i) Cox model [8,9,4]
The hazard rate h is assumed to be equal to a baseline hazard $h_0(t)$ (the nonparametric part of the model), modified by p covariates $\boldsymbol{X} = (X_1, \cdots, X_p)$ whose weights are the parameters $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$ (the parametric part of the model) to be estimated, as well as h_0 which may be any positive function:

$$h(t|\boldsymbol{X}) = h_0(t)\ \mathbf{e}^{\boldsymbol{\beta}^T \boldsymbol{X}} \tag{11}$$

Usually, one is only interested in $\boldsymbol{\beta}$ and the objective is to get rid of the nuisance infinite dimensional parameter h_0. This can be done using Cox's partial likelihood (see below). Once the respective weights of the risk factors have been estimated, one can derive an estimation of the function h_0.

(ii) Estimation method
In this semi-parametric case, the baseline h_0 is free to be any function. The likelihood is then replaced by Cox's partial likelihood: at each failure time y_i, consider the ratio of the item that fails to all items still present at risk. Those at risk are those who are neither failed nor censored at time y_i. This eliminates h_0, as it has the same value, $h_0(y_i)$ for everyone, and takes advantage of all items that will be censored later on, in the denominators. Then, once we have estimated the coefficients $\boldsymbol{\beta}$ of the risk factors, we can estimate h_0.

(iii) Frailty models
Up to now, the observed population is assumed to be homogeneous and all items (or patients) are independent. In order to model a possible inhomogeneity, one can introduce into the model a random effect η, called frailty [46,29,30], acting multiplicatively on the hazard rate $h(t|\mathbf{x})$ of an individual with covariate vector $\mathbf{x}$.

$$h(t|\mathbf{x}, \eta) = \eta \exp(\boldsymbol{\beta}^T \mathbf{x})\ h_0(t) \tag{12}$$

A frailty model may be considered as a Cox model with an unobserved covariate $\ln(\eta)$ whose coefficient is equal to 1 and whose distribution function is known, derived from F_η the distribution of η. Flexible distributions are usually chosen for F_η, the most common being gamma, but also inverse gaussian and all stable distributions, leading to the model

$$S(t|\mathbf{x}, \eta) = \exp(-\eta \exp(\boldsymbol{\beta}^T \mathbf{x}) H_0(t)) \tag{13}$$

where $H_0(t)$ is the baseline cumulative hazard. Thus, as η is not observed, what is available is the survival integrated with respect to η:

$$\begin{aligned} S(t|\mathbf{x}) = & \int_0^\infty \exp(-u \exp(\boldsymbol{\beta}^T \mathbf{x}) H_0(t))\, dF_\eta(u) \\ = & \exp(-G(\exp(\beta^T \mathbf{x}) H_0(t))) \end{aligned} \tag{14}$$

where G is -log of the Laplace transform of η distribution function

$$G(y) \ = \ -\ln(\int_0^\infty \exp(-uy) dF_\eta(u) \tag{15}$$

The simple Cox model is obtained when G is the identity.

2.3. ***Nonparametric survival models***

In industrial setting, accelerated failure time models (AFT) [2,3] are more popular than Cox model and covariates are called stresses which are often controlled while the covariates are simply observed: for G a survival function, i.e. a decreasing function from 1 to 0 on $\mathbb{R}^+$, and r a positive function on a p-dimensional process $\mathcal{X}$

$$S(t|X(s), 0 \leq s \leq t) = G(\int_0^t r(X(s))ds) \quad \forall X \in \mathcal{X} \tag{16}$$

We can see there a totally nonparametric model, involving two unknown functions, G and r. It has a semi-parametric version when one of the two unknown functions involved, G or r, is replaced by a function depending on a finite number of parameters; and a fully parametric version if both functions are assumed to be known except for one or more real parameters to be estimated. Those replacements should be done with flexible functions able to adopt different shapes. Other examples of nonparametric models may be found in [10,15].

Let us remark that parametric models may seem to be too coercive as compared to nonparametric ones. One way to remedy this possible defect is to replace the distribution function of a parametric model by a neighborhood defined through a distance on probability distributions like Hellinger or Prokhorov distance. This leads to a fourth type of model, called robust models [26,21].

2.4. ***Latent variable model: First Hitting Time***

Health $L(t)$ of a patient, or operational state $L(t)$ of a technological material is a latent (not observed) variable that decreases to 0 due to three types of risk factors. The event occurs when this process reaches a boundary, usually 0 [36,35]. In order to make clear the structure of this model, let us give a motivating medical example:

One has to estimate the expected years of life free of lung cancer lost due to occupational exposure to asbestos. Such a study was required based on a french case-control survey [5]. The model is as follows:

$$L(t|h, \mu) = \ell + \mu t + B(t) \tag{17}$$

where

(i) $\ell > 0$, the initial amount of health, is a function of the initial covariates $\boldsymbol{X}_I$: gender, past family disease history, genetic factors,...(Note that in our example, the amount of health is meant with respect to lung cancer occurrence, not with respect to death itself).

(ii) $\mu < 0$, the slope of the process, is a function of $\boldsymbol{X}_I$ and also of lifetime covariates $\boldsymbol{X}_L$: smoking and food habits, environment, biological measurements (glycemia, cholesterol, $\cdots$).

(iii) $B(t)$ is a Brownian motion, the random part of the model.

(iv) Finally, the focus is put on a special risk factor $\boldsymbol{X}_S$ (here occupational exposure to asbestos), which accelerates by a function $r(t)$ the time to onset of lung cancer.

The initial amount, ℓ, will depend on $\mathbf{x}_I$, the slope, μ, on $\mathbf{x}_I$ and $\mathbf{x}_L$. Usually this dependence is assumed to be linear, which results in a parametric model, with the weights of each risk factor to be determined for ℓ and for μ. Now, given the values of $\mathbf{x}_I$ and $\mathbf{x}_L$ of a patient, one can compute the years free of disease that he lost due to his exposure to asbestos by using the acceleration function r, estimated from the training set. In the absence of any asbestos exposure the waiting time to the event (lung cancer occurrence) is

$$T(\ell, \mu) = \inf\{t \geq 0 : L(t|\ell, \mu) \leq 0\}, \tag{18}$$

$T(\ell, \mu)$, ($< \infty$ as $\mu < 0$) is inverse Gaussian ($\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$ is the standard normal cdf (cumulative distribution function)):

$$F(t|\ell,\mu) = 1 + e^{-2\ell\mu}\Phi\left((\mu t - \ell)t^{-1/2}\right) - \Phi\left((\mu t + \ell)t^{-1/2}\right), \tag{19}$$

In the presence of asbestos exposure, the time to event is accelerated by a nondecreasing continuous function r on $\mathbb{R}^+$ such that $r(t) \geq t\ \forall t$:

$$T(\ell,\mu,r) = \inf\{t \geq 0 : L(r(t)|\ell,\mu) \leq 0\} \tag{20}$$

This parametric model has a nonparametric version [19].

3. Illustration of the difficulty of model selection

Model selection is a difficult problem. The following artificial example is an illustration of the difficulty of eliminating irrelevant predictors. It shows that removing potential risk factors which seem to be independent of the outcome may lead to serious errors.
Let us consider a very simple diagnosis problem, discrimination between 2 diseases $M = M_1$ or $M = M_2$ based on 3 symptoms $X_j,\ j = 1, 2, 3$, meeting the following distribution:

$M = M_1$

$X_1 = 0$

X_2 \ X_3	0	1
0	1/4	0
1	0	1/4

$X_1 = 1$

X_2 \ X_3	0	1
0	0	1/4
1	1/4	0

$M = M_2$

X_2 \ X_3	0	1
0	0	1/4
1	1/4	0

X_2 \ X_3	0	1
0	1/4	0
1	0	1/4

(i) All 3 symptoms are present with the same probability 1/2 in M_1 and in M_2. This implies that all 3 symptoms seem to be independent of the type of the disease, M_1 or M_2.

(ii) For any pair of symptoms x_j, x_k, $P(M_1|x_j x_k) = P(M_2|x_j x_k) = 1/4$, so that every pair $(X_j, X_{j'})$ is uniform on its 4 values in M_1 as well as in M_2. As a result, none of the 3 pairs $(X_j, X_{j'})$ can discriminate M_1 and M_2.

(iii) But for any triplet of symptoms (x_1, x_2, x_3), whenever $P(M_1|x_1, x_2, x_3) = 1)$, $P(M_2|x_1, x_2, x_3) = 0$ and the reverse is also true. This implies that (X_1, X_2, X_3) altogether discriminate perfectly M_1 and M_2 and lead to a perfect diagnosis of M.

This phenomenon can be extended to any k-uple of risk factors, so that one has to take into account the maximum possible size of sets of risk factors allowed by the number of observed items as compared to the number of risk factors.

4. Reduction methods

Among existing methods (POD, PGD and others [42]), let us cite PGD (Proper Generalized Decomposition method) of Chinesta, Ladevèze et al [6,7]:

$$f(x_1,..,x_p) \simeq \sum_{i=1}^{q}(\prod_{j=1}^{p} f_i^j(x_j)) \tag{21}$$

and the classical Singular Value Decomposition (SVD) of a matrix $A_{n\times p}$, which is useful when the data is a high dimensional (rectangular) $(n \times p)$ matrix A. It replaces A by a product of three matrices, the inner one being diagonal with dimension r, the rank of A, smaller than min(n,p). It can be seen as a Principal Component Analysis of the matrix $A_{n\times p}$, which replaces the explanatory variables $\boldsymbol{X}$ for the outcome Y by a few linear combinations of them, giving thus a probabilistic interpretation of SVD. Finally, I will cite a decomposition of a function $f(x_1,..,x_p)$ of several discrete variables, based on variance analysis, that is useful for sparse contingency tables. For an overview of existing methods, see [41].

4.1. *Singular Value Decomposition (SVD)*

$$\boldsymbol{A} : \mathbb{R}^p \to \mathbb{R}^n \qquad \text{rank}(A) = r \leq \min(n,p) :$$

$$A_{n\times p} = \underbrace{\left[\begin{array}{cccc} x_{11} & x_{12} & ..... & x_{1p} \\ x_{21} & x_{22} & ..... & x_{2p} \\ ... & ... & ..... & ... \\ x_{n1} & x_{n2} & ..... & x_{np} \end{array}\right]}_{p} \Bigg\} \; n \; : \; n \ll p, \text{ a centered matrix}$$

Find $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}$ such that:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \qquad \text{or} \qquad \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{V} = \boldsymbol{D} \tag{22}$$

where

$$\left\{\begin{array}{ll} \boldsymbol{U}_{n\times r} \text{ orthonormal basis of} & \mathcal{E}_r \subsetneqq \mathbb{R}^n \text{ column space of } A \\ \boldsymbol{V}_{p\times r} \text{ orthonormal basis of} & \mathcal{E}'_r \subsetneqq \mathbb{R}^p \text{ row space of } A \\ \boldsymbol{D}_{r\times r} \text{ diagonal matrix} & (d_1, \cdots, d_r) \;\; ; \;\; r = \text{rank}(A) \end{array}\right. \tag{23}$$

Let us remark that $\boldsymbol{U}$ and $\boldsymbol{V}$ have left inverses:

$$\boldsymbol{V}^T_{r\times p}\boldsymbol{V}_{p\times r} = \boldsymbol{U}^T_{r\times n}\boldsymbol{U}_{n\times r} = \boldsymbol{I}_{r\times r} \tag{24}$$

In order to obtain $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{D}$ verifying Eqs. (22), we proceed as follows:

As $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_{r\times r}$, $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^T$ and, as $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}_{r\times r}$, $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^T$. $\boldsymbol{A}^T\boldsymbol{A}$, the Gram matrix associated to $\boldsymbol{A}$, and $\boldsymbol{A}\boldsymbol{A}^T$ are semi-definite positive matrices, so that the columns of $\boldsymbol{V}$ are the real eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$ and the columns of $\boldsymbol{U}$ are the real eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$ and $d_1^2, \cdots, d_r^2$ are the common positive eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^T$. As a consequence, the singular value decomposition of A is $\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$.

4.2. ***Statistical approach to SVD: Principal Component Analysis (PCA)***

$$A_{n\times p} = \begin{array}{c} \begin{array}{cccccc} & X_1 & X_2 & \ldots\ldots & \ldots\ldots & X_p \end{array} \\ \left.\begin{bmatrix} x_{11} & x_{12} & \ldots\ldots\ldots\ldots & x_{1p} \\ x_{21} & x_{22} & \ldots\ldots\ldots\ldots & x_{2p} \\ \ldots & \ldots & \ldots\ldots\ldots\ldots & \ldots \\ x_{n1} & x_{n2} & \ldots\ldots\ldots\ldots & x_{np} \end{bmatrix}\right\} n \\ \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{p} \end{array} : n \ll p, \text{ a centered matrix}$$

A represents n observations (on n patients or n technological devices) of the p-dimensional variable (covariate or stress) $\boldsymbol{X} = (X_1, \cdots, X_p)$, and it is considered as a cloud of n points in $\mathbb{R}^p$ or p points in $\mathbb{R}^n$. The goal is to find the best representation of the cloud in a space of smaller dimension, while keeping most of the initial information. This suggests to maximize the variance of linear combinations of the vectors X_j, which is obtained by projection on the first k eigenvectors of the symmetric semi-definite positive matrix $A^T A$. This matrix is an estimator of Σ, the unknown covariance matrix of $\boldsymbol{X}$. This can be seen from the following equations:
Let W be any linear combination of the $X_j's$, $W = a_1 X_1 + \cdots + a_p X_p$ It has a representation in terms of the eigenvectors, $W = b_1 V_1 + \cdots + b_r V_r$, so that $\widehat{Var(W)} = \mathbf{a}^T A^T A \mathbf{a} = b_1^2 \lambda_1 + \cdots + b_r^2 \lambda_r$, where $\sum_{j=1}^{r} b_j^2 = 1$. We can see from those equations that the "most informative k-dimensional subspace" in $\mathbb{R}^p$ is the space spanned by $(V_1, \cdots, V_k)$, the first k eigenvectors of $A^T A$.

4.3. ***Decomposition of a function of several random variables***

Let f be a real function of p real random variables. We have the following decomposition lemma based on the classical variance analysis:
Every integrable function $f = f(x_1, \cdots, x_p)$ may be decomposed uniquely into a sum:

$$f = C + \sum_{j=1}^{p} g_j(x_j) + \sum_{j,k\in\{1,\cdots,p\}^2,\ j<k} g_{jk}(x_j, x_k) + .. + g_{1,..,p}(x_1, .., x_p) \tag{25}$$

where C is a constant and all expectations of g functions on any of their arguments are 0.

This result will be essentially applied in the following case: the variables X_j are discrete and take a finite number of values. The expectations are taken with respect to the joint probability of $\boldsymbol{X} = (X_1, \cdot, X_p)$.
Constructive decomposition

$$\begin{aligned} C &= E(f(X_1, .., X_p)) \\ g_j(x_j) &= E(f(X_1, .., X_p)|X_j = x_j) - C \\ g_{j,k}(x_j, x_k) &= E(f(X_1, .., X_p)|(X_j, X_k) = (x_j, x_k)) - g_j(x_j) - g_k(x_k) - C \\ etc &\ \ldots \end{aligned}$$

This approach is very useful for sparse contingency tables as can be seen from the following diagnosis example, where the data leave many symptom profiles empty. We have $p = 9$ binary syptoms X_j, $j = 1, \cdots, p$, $X_j = 1$ if the symptom is present, which gives $p' = 2^9 = 512$ symptom profiles, which are supposed to discriminate two different forms, M_1 and M_2, of the same disease M.

Observations on $n = 150$ patients give counts for 1024 cells, so that many cells are empty in the observed sample, though their probabilities may not be assumed to be equal to 0.

$$A_{2\times 512} = \left.\underbrace{\begin{bmatrix} n_{11} & n_{12} & \ldots\ldots\ldots\ldots & n_{1p'} \\ n_{21} & n_{22} & \ldots\ldots\ldots\ldots & n_{2p'} \end{bmatrix}}_{p'=512}\right\} q = 2$$

$$\begin{aligned}\log(P(\boldsymbol{X} = \mathbf{x}|M_1)) &= C + \textstyle\sum_{j=1}^{p} g_j(x_j) + \sum_{j\neq j',\ j<j'} g_{j,j'}(x_j, x_{j'}) \\ &+ \textstyle\sum_{j\neq j'\neq k,\ j<j'<k} g_{j,j',k}(x_j, x_{j'}, x_k) + \cdots + g_{1,2,\cdots,p}(x_1, x_2, \cdots, x_p)\end{aligned}$$

We have two such developments, one for disease M_1, and one for disease M_2, with sums of functions of an increasing number, k, of variables. If we can assume that the symptoms are independent, which is very simple but often unrealistic, we can stop the preceding developments at $k = 1$. If we stop at $k = 2$, it means that we assume a possible dependence of order 2, but with no influence of any third symptom on the link between any two given symptoms. Now, stopping at k, named assumption H_k, i.e. cutting off all functions of more than k arguments, means that we allow possible order k interactions, but no interaction greater than k. Under assumption H_k, the marginals of order k of the data are sufficient statistics [22].

5. Neural Networks

The most usual Neural Network, the single hidden layer back-propagation network (or single layer perceptron), is a particular case of a stochastic model, the Projection Pursuit Regression and Discrimination (PPRD) model.

5.1. *Projection Pursuit Regression and Discrimination model (PPRD)*

This semi-parametric model is fit for regression as well as discrimination problems.

(i) Regression

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

$$\widehat{Y} = \widehat{f}(\boldsymbol{X}) := \sum_{m=1}^{M} \widehat{g_m}(\widehat{\boldsymbol{\omega}_m^T}\boldsymbol{X}) := \sum_{m=1}^{M} \widehat{g_m}(V_m) \tag{26}$$

in which $\boldsymbol{\omega}_m, m = 1, \cdots, M$ are M unitary p-dimensional vectors and functions $g_m : \mathbb{R} \to \mathbb{R}$. Function $g_m(V_m)$ is called a ridge function. The estimators are based on the observed training sample: $(\mathbf{x}_i, y_i), i = 1, \cdots, n$. This is an additive model, but not with respect to the initial variables $\boldsymbol{X}$ but with respect to appropriate linear combinations of them: $V_m = \boldsymbol{\omega}_m^T \boldsymbol{X}$.

A remarkable property of this model is that if we choose M big enough, *any continuous function may be approximated arbitrarily well.* This is what made the success of neural networks which are a parametric version of this stochastic model. A simple example of the nonlinearity of the model may be given in the simplest case where $p = M = 2$, $\omega_1 = (1/\sqrt{2}, 1/\sqrt{2})$, $\omega_2 = (1/\sqrt{2}, -1/\sqrt{2})$, $g_1(t) = t^2/4$, $g_2(t) = -t^2/4$, so that the resulting value of f is $f(\boldsymbol{X}) = X_1 X_2$. However, a drawback of this model is the difficulty of interpretation of the results in terms of the initial risk factors as each feature X_j is scattered into every linear combination of $\boldsymbol{X}$. The quadratic measurement error is:

$$R(\boldsymbol{\theta}) := \sum_{i=1}^{n} [y_i - \sum_{m=1}^{M} g_m(\boldsymbol{\omega}_m^T \mathbf{x}_i)]^2 \tag{27}$$

where $\boldsymbol{\theta}$ is the set of parameters of the problem ($\boldsymbol{\theta} := (\boldsymbol{\omega}_m, g_m)$).

(ii) Discrimination into K categories

Up to now we considered the case of a regression problem, Y being a random variable in $\mathbb{R}$. If the problem is a discrimination one, the target $\boldsymbol{Y} = (Y_1, \cdots, Y_K)$ is one of K categories, each Y_k being coded as a (0,1) variable. In that case, two different error measurements are considered:

$$R_2(\boldsymbol{\theta}) \quad := \quad \sum_{k=1}^{K} \sum_{i=1}^{n} (y_{ik} - f_k(x_i))^2 \text{ quadratic error}$$

$$R_{KL}(\boldsymbol{\theta}) := - \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log(f_k(x_i)) \text{ crossed entropy}$$

The index KL for the crossed entropy makes reference to Kullback Leibler distance which is equal up to an additive constant to crossed entropy as defined up there. We recall that the Kullback-Leibler distance of two probabilities P and Q is defined as

$$KL(P, Q) = \int \log(\frac{dP}{dQ})\, dP$$

5.1.1. *Neural Network as a special case of PPRD*

Let the framework be a discrimination problem: the target $\boldsymbol{Y}$ is a category, each Y_k being coded as a (0,1) variable. Y_k is modeled as a function g_k of a linear combination of variables obtained by a linear combination of *activated* M linear combinations of the inputs.

We can see that linearity comes in twice in this definition, with $p \times M$ coefficients α and $p \times K$ coefficients β. To those parameters are added the activation function σ and the K functions g_k:

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

where $g_k(t) = \frac{e^{T_k}}{\sum_{i=1}^{K} e^{Ti}}$. This choice ensures that all $\widehat{Y_k} := f_k(\boldsymbol{X})$ are positive and add to 1.

There are several possible choices for the activation function σ, see Figure 3. All of them are smoothed versions of the step function $s(u) = 1\{u \geq 0\}$ (up to an additive constant). The non-linearity of the model is due to the activation function. If σ is the identity, the model becomes linear.

$$\begin{aligned}
\sigma(u) \quad &= \frac{1}{1 + e^{-u}} \quad \text{the sigmoïd, the most usual one} \\
\sigma(u) \quad &= \frac{e^u - e^{-u}}{e^u + e^{-u}} \quad \text{hyperbolic tangent (th(u))} \\
\sigma(a, u) \quad &= \begin{cases} a(e^u - 1) & \text{for } u < 0 \\ u & \text{for } u \geq 0 \end{cases} \text{ Exponential Linear Unit (ELU)} \\
\sigma(a, u) \quad &= \begin{cases} au & \text{for } u < 0 \\ u & \text{for } u \geq 0 \end{cases} \text{ Rectified Linear Unit (ReLU)} \\
\sigma(a, b, u) \; &= b \begin{cases} a(e^u - 1) & \text{for } u < 0 \\ u & \text{for } u \geq 0 \end{cases} \text{ Scaled Exponential Linear Unit (SELU)}
\end{aligned}$$

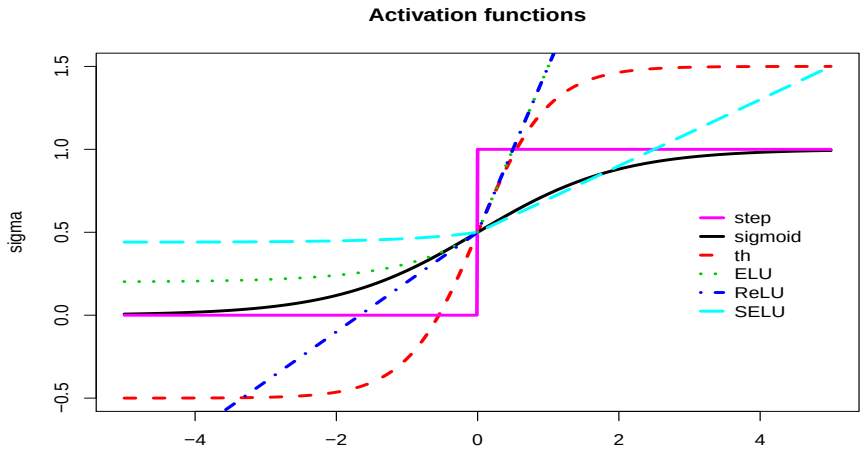


Figure 3. Several activation functions

6. Application to Alzheimer prediction

We analyze the comparison of prediction abilities of a stochastic model, the logistic regression (GLM), and a neural network (NN) approach, for the prediction for patient to develop Alzheimer in the next 4 years , based on the observation of 17 risk factors. The data set is issued from Pitié-Salpêtrière Hospital in Paris (France). It has $n = 5003$ patients each with $p = 17$ covariates, including 3 genetic factors, familial disease anteriority and personal factors (age, sex, education level, $\cdots$). 467 patients were excluded for missing values, 142 developed the disease and 4214 were free of disease, and thus called controls [28]. Let us notice that this is a very simple problem, which deals only with a discrimination problem between two outcomes (Alzheimer or not) and a data set which is not big at all as it involves only a (5003×17) matrix. Note also the very unbalanced counts for diseased (142) and controls (4214) which makes the discrimination uneasy for both methods. The logistic model reads

$$P(Y = 1|\boldsymbol{X} = \mathbf{x}) = \frac{\exp(\beta^T \mathbf{x})}{1 + \exp(\beta^T \mathbf{x})} \tag{28}$$

We proceed as follows

(i) Split at random 3/4 of the data set to be the training set. This is a rather huge training set due to the paucity of patients who developed the disease.

(ii) The remnant 1/4 will be the test set on which to predict who will be Alzheimer.

(iii) Use separately logistic model (GLM) and neural network (NN) on the training set to estimate the probabilities $R_{GLM}(\boldsymbol{X})$ and $R_{NN}(\boldsymbol{X})$ to develop Alzheimer based on the risk factors $\boldsymbol{X}$.

(iv) Predict, on the test set, who will be Alzheimer based on the estimations done with both methods. The result is four counts for each method: true positive, false positive, true negative, false negative. Or, equivalently, the probabilities of correct classifications of Alzheimer (p_d) and non Alzheimer (p_{nd})

(v) Repeat this process N times, for both methods, to obtain confidence intervals for the probabilities of correct prediction of Alzheimer, see Table 1:

Method	p_d	p_{nd}	p_g	$CI_{95\%}(p_d)$	$CI_{95\%}(p_{nd})$
GLM	0.72	0.73	0.73	0.55 0.85	0.70 0.76
NN(2)	0.68	0.73	0.73	0.50 0.85	0.65 0.77

Table 1: Comparison of prediction abilities of GLM and NN.

Average correct predictions due to GLM and NN are p_d for dementia cases, p_{nd} for people without dementia, p_g for global and $CI_{95\%}$ are the respective confidence intervals.

Hereafter is an illustration, see Figure 4, of the neural network we used, with 2 layers, the first one with 3 neurons and the second one with 2 neurons. In order to have a more readable picture, we choose to show as entries only the four relevant predictive factors: age, incapacity, depression and the gene APOE4, a reduced number (only 4) as compared to the initial 17 entries.

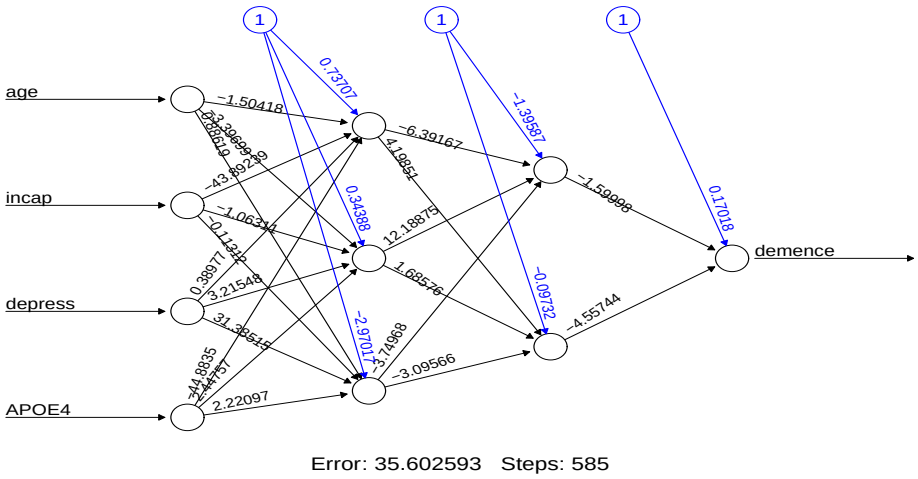


Figure 4. 2 layers, 3 and 2 neurons

Some comments

(i) It was very surprising that the education level did not play a role in the occurrence of the disease.

(ii) The fact that the counts are very much unbalanced (142 dementia cases versus more than 4000 without dementia) creates problems for the prediction: the confidence intervals are large.

(iii) To overcome this problem, one can duplicate the smaller category [32], see Table 2:

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(2)	0.75	0.72	0.73	0.73 0.78	0.70 0.75

Table 2: correct predictions due to GLM and NN for dementia cases (p_d), for people without dementia (p_{nd}), global p_g, and 95% confidence intervals after duplication.

After duplication, the widths of the 95% confidence intervals are reduced [0.71 0.76] instead of [0.50 0.85] for the future Alzheimer detection and [0.70 0.75] instead of [0.65 0.77] for the future non Alzheimer.

(iv) The prediction abilities of the two methods are very comparable on this example, with moderate dimension of the data set.

(v) Increasing the number of neurons from 2 to 3 in a unique hidden layer does not improve the prediction ability. The only consequence is increasing the time to get the result. Same comment when using two layers with respective number of neurons 3 and 2

As a perspective, one could use world wide data (Big Data) related to Alzheimer disease and the observed risk factors on the patients under survey. Then this would become a regression problem, estimating the time to onset of the disease as a function of the numerous available risk factors.

7. Conclusions and perspectives

(i) Analysis of risk is a crucial issue nowadays. The same mathematical methods to evaluate the impact of risk factors on the waiting time for an event to occur may be applied to technological aging systems [44], to public health problems [47], insurance, management [17], ecology and other fields.

(ii) The link between survival analysis and reliability, ignored for a long time, so that the terminology is different in both fields to name the same concepts, is now fully acknowledged. Only recently, their common points, more numerous than their differences, were recognized. The favorite model though for survival data is the Cox model and its extensions, while accelerated models are preferred in an industrial environment [2,3].

(iii) The increasing power of computers facilitates the development of non parametric and semi-parametric stochastic models, more greedy in time computation than the parametric ones, and also of complex algorithms in Artificial Intelligence to deal with Big Data. Many new packages appear every month in R and Python which are available on the web.

(iv) Statistical Learning and Big Data (SLBD):
Research statisticians turn now to Machine Learning, like Hastie, Tibshirani et al [18] and the reverse is also true. Machine learning researchers like Murphy [39] investigate the probability background of their manipulation of Big Data through algorithms. This forms finally what could be called Statistical Learning and Big Data.

(v) The increasing knowledge of the genome [34] provides high dimensional data [34]. The FHT (First Hitting Time) model, in use in this context, gives a special status to some risk factor.

(vi) Specific developments in medicine:
"Individualized medicine", which seems to be in total contradiction with classical statistical analysis is now seriously taken into account. Including all the data relative to a patient, such as medical doctors report (text data) leads to the development of new graphical models.

Acknowledgements
The authors dedicate this work to professor Mikhail Nikulin, from Saint Petersburg, Russia, who shared our research activities during many years.

References

[1] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[2] Vilijandas Bagdonavicius and Mikhail Nikulin. *Accelerated life models: modeling and statistical analysis*. Chapman and Hall/CRC, 2001.

[3] Vilijandas Bagdonavičius and Mikhail S Nikulin. Goodness-of-fit tests for accelerated life models. In *Goodness-of-fit tests and model validity*, pages 281–297. Springer, 2002.

[4] Jean Bretagnolle and Catherine Huber-Carol. Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics*, pages 125–138, 1988.

[5] Antoine Chambaz, Dominique Choudat, Catherine Huber, Jean-Claude Pairon, and Mark J Van der Laan. Analysis of the effect of occupational exposure to asbestos based on threshold regression modeling of case–control data. *Biostatistics*, 15(2):327–340, 2013.

[6] Francisco Chinesta, Amine Ammar, Adrien Leygue, and Roland Keunings. An overview of the proper generalized decomposition with applications in computational rheology. *Journal of Non-Newtonian Fluid Mechanics*, 166(11):578–592, 2011.

[7] Francisco Chinesta, Pierre Ladeveze, and Elías Cueto. A short review on model order reduction based on proper generalized decomposition. *Archives of Computational Methods in Engineering*, 18(4):395, 2011.

[8] David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

[9] David R Cox. *Analysis of survival data*. Chapman and Hall/CRC, 2018.

[10] Jean-Jacques Droesbeke, Société mathématique de France, Association pour la statistique et ses utilisations (France), and Journées d'étude en statistique (3: 1988: Marseille-Luminy). *Analyse statistique des durées de vie: modélisation des données censurées*. Economica, 1989.

[11] Bradley Efron. Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, 83(402):414–425, 1988.

[12] Shulamith T Gross and Catherine Huber. Matched pair experiments: Cox and maximum likelihood estimation. *Scandinavian Journal of Statistics*, pages 27–41, 1987.

[13] Shulamith T Gross and Catherine Huber-Carol. Regression models for truncated survival data. *Scandinavian Journal of Statistics*, pages 193–213, 1992.

[14] Shulamith T Gross and Tze Leung Lai. Bootstrap methods for truncated and censored data. *Statistica Sinica*, pages 509–530, 1996.

[15] Shulamith T Gross and Tze Leung Lai. Nonparametric estimation and regression analysis with left-truncated and right-censored data. *Journal of the American Statistical Association*, 91(435):1166–1180, 1996.

[16] Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

[17] Boris Harlamov. *Stochastic risk analysis and management*. John Wiley & Sons, 2017.

[18] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[19] Xin He and Mei-Ling Ting Lee. First-hitting-time based threshold regression. *International Encyclopedia of Statistical Science*, pages 523–524, 2011.

[20] Catherine Huber. Efficient regression estimation under general censoring and truncation. In *Mathematical and Statistical Models and Methods in Reliability*, pages 235–241. Springer, 2010.

[21] Catherine Huber. Robust versus nonparametric approaches and survival data analysis. In *Advances in Degradation Modeling*, pages 323–337. Springer, 2010.

[22] Catherine Huber and Joseph Lellouch. Estimation dans les tableaux de contingence a un grand nombre d'entrées. *International Statistical Review/Revue Internationale de Statistique*, pages 193–203, 1974.

[23] Catherine Huber and Mikhail S Nikulin. Remarques sur le maximum de vraisemblance. *Qüestiió: quaderns d'estadística i investigació operativa*, 21(1), 1997.

[24] Catherine Huber, Valentin Solev, and Filia Vonta. Estimation of density for arbitrarily censored and truncated data. In *Probability, statistics and modelling in public health*, pages 246–265. Springer, 2006.

[25] Catherine Huber, 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.

[26] Peter J Huber and Elvezio M Ronchetti. *Robust Statistics*. Wiley, New York, 1981.

[27] Catherine Huber-Carol, Narayanaswamy Balakrishnan, M Nikulin, and M Mesbah. *Goodness-of-fit tests and model validity*. Springer Science & Business Media, 2012.

[28] Catherine Huber-Carol, Shulamith T Gross, and Annick Alpérovitch. Within the sample comparison of prediction performance of models and submodels: Application to Alzheimer's disease. *Statistical Models and Methods for Reliability and Survival Analysis*, pages 95–109, 2013.

[29] Catherine Huber-Carol and Filia Vonta. Frailty models for arbitrarily censored and truncated data. *Lifetime Data Analysis*, 10(4):369–388, 2004.

[30] Catherine Huber-Carol and Filia Vonta. Semiparametric transformation models for arbitrarily censored and truncated data. In *Parametric and Semiparametric Models with Applications to Reliability, Survival Analysis, and Quality of Life*, pages 167–176. Springer, 2004.

[31] Jerald F Lawless. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, 2011.

[32] Yann Le Cun. *Personal communication*. Yann Le Cun, December, 2018.

[33] Yann Le Cun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[34] Mei-Ling Ting Lee. *Analysis of microarray gene expression data*. Springer Science & Business Media, 2007.

[35] Mei-Ling Ting Lee, GA Whitmore, and Bernard A Rosner. Threshold regression for survival data with time-varying covariates. *Statistics in Medicine*, 29(7-8):896–905, 2010.

[36] Mei-Ling Ting Lee and George A Whitmore. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Statistical Science*, 21(4):501–513, 2006.

[37] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[38] Govind S Mudholkar and Deo Kumar Srivastava. Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42(2):299–302, 1993.

[39] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[40] Mikhail Nikulin and Firoozeh Haghighi. A chi-squared test for the generalized power Weibull family for the head-and-neck cancer censored data. *Journal of Mathematical Sciences*, 133(3):1333–1341, 2006.

[41] Anthony Nouy. Low-rank tensor methods for model order reduction. *Handbook of uncertainty quantification*, pages 857–882, 2017.

[42] Roger Ohayon. Reduced models for fluid–structure interaction problems. *International Journal for Numerical Methods in Engineering*, 60(1):139–152, 2004.

[43] Odile Pons. Estimation in a Cox regression model with a change-point according to a threshold in a covariate. *The Annals of Statistics*, 31(2):442–463, 2003.

[44] Vladimir Rykov. *Reliability of engineering systems and technological Risk*. John Wiley & Sons, 2016.

[45] Terry M Therneau and Patricia M Grambsch. *Modeling survival data: extending the Cox model*. Springer Science & Business Media, 2013.

[46] Filia Vonta. Efficient estimation in a non-proportional hazards model in survival analysis. *Scandinavian Journal of Statistics*, pages 49–61, 1996.

[47] Filia Vonta, MS Nikulin, Nikolaos Limnios, and Catherine Huber-Carol. *Statistical models and methods for biomedical and technical systems*. Springer Science & Business Media, 2008.