SIMERG2E

Statistical Inference for the Management of Extreme Risks,
Genetics and Global Epidemiology
http://mistis.inrialpes.fr/simerge

- Joint team started in **January 2015** (SIMERGE),
- Extended in **January 2018** for 3 additional years (SIMERG2E).

Part of the Inria International Lab LIRIMA.

Outline

- 1. Scientific assessment,
- 2. Reduced-bias estimators of the Conditional Tail Expectation for heavy-tailed distributions,
- 3. A classification method for binary predictors combining similarity measures and mixture models,
- 4. Future of the partnership.

1. Scientific assessment

SIMERG2E is co-headed by

- Abdou Kâ Diongue, Université Gaston Berger, Saint-Louis, Sénégal,

- Stéphane Girard, Mistis, Inria Grenoble Rhône-Alpes, France.

Partners (SIMERGE & SIMERG2E)

➪ Sénégal

- Institut Pasteur de Dakar, 4-year group in Bioinformatics, Biostatistics and Modelling (G4-BBM).
- IRD (Institut de Recherche pour le Développement), Unité de Recherche sur les Maladies Infectieuses et Tropicales Emergentes (URMITE), Dakar.
- LERSTAD (Laboratoire d'études et de recherches en statistiques et développement), Université Gaston Berger, Saint-Louis.

➪ France

- Mistis, Inria Grenoble Rhône-Alpes.
- Modal, Inria Lille Nord-Europe.

Members (September 2018)

➪ Permanent staff

- Institut Pasteur de Dakar: Cheikh Loucoubar (research scientist).
- Mistis: Stéphane Girard (**co-head**, senior researcher), Florence Forbes (senior researcher).
- Modal: Sophie Dabo-Niang (professor), Serge Iovleff (assistant professor), Baba Thiam (assistant professor).
- LERSTAD: Abdou Kâ Diongue (**co-head**, professor), Aliou Diop (professor), El Hadji Deme (assistant professor).

➪ Non-permanent staff

- PhD students: Aboubacrène Ag Ahmad (LERSTAD & Mistis), Clément Albert (Mistis).
- Data scientist: Seydou-Nourou Sylla (Institut Pasteur de Dakar).

➪ Former PhD students

Seydou-Nourou Sylla (IRD, LERSTAD & Mistis), Alessandro Chiancone (Mistis), Pathé Ndao & Aladji Bassene (LERSTAD & Modal).

Research themes

The Associate team is built on two research themes in **statistics**:

⇨ 1. Spatial extremes, application to management of extreme risks

- Estimation of general risk measures in case of extreme losses (basing on the extreme-value theory).
- Estimation of such extreme risk measures able to deal with covariates (using nonparametric methods).
- Focus on the case where the covariate is a random field ($\Longrightarrow$ quantitatively characterize the spatial dependences between extreme climate events).

⇨ 2. Classification, application to global epidemiology

- Design of parsimonious multinomial probability distributions to model each class of the mixture.
- Introduction of a kernel function in the Gaussian mixture model to measure the similarity between binary data.
- Taking into account of hierarchical dependences between variables.

(International) Visits

➫ From Sénégal to France

	Visits	Context	Duration
Seydou-Nourou Sylla	3	PhD work	1 year
Aboubacrène Ag Ahmad	2	PhD work	14 weeks
El Hadji Deme	2	collaborative work$^{(*)}$	10 weeks
Total	7		18 months

➫ From France to Sénégal

	Visits	Context	Duration
Sophie Dabo-Niang	3	collaborative work	2 months
Serge Iovleff	2	collaborative work + courses	5 weeks
Stéphane Girard	1	collaborative work + courses	2 weeks
Total	6		$\simeq$ 4 months

$^{(*)}$ The first visit of El Hadji Deme was combined with the invitation of Prof. Abdelhakim Necir (Université de Biskra, Algérie) to Grenoble in the framework of a joint collaboration.

PhD defences

➪ 1. Spatial extremes, application to management of extreme risks

- Pathé Ndao, "Modélisation de valeurs extrêmes conditionnelles en présence de censure", Université Gaston Berger de Saint-Louis, Sénégal, August, 2015. He works as an ATER at Université Bretagne Sud.
- Aladji Bassene, "Contribution à la modélisation spatiale des événements extrêmes", Université Lille, May, 2016. He works now as a senior biostatistician at the Unité de Recherche Clinique Lariboisière-St Louis, Paris.

➪ 2. Classification, application to global epidemiology

- Alessandro Chiancone, "Réduction de dimension via Sliced Inverse Regression: Idées et nouvelles propositions", Université Grenoble-Alpes, October, 2016. He works now as a research engineer at Know-Center, Austria.
- Seydou-Nourou Sylla, "Modélisation et classification de données binaires en grande dimension - Application à l'autopsie verbale", Université Gaston-Berger, December, 2016. He works now as a data scientist at the Institut Pasteur, Dakar.

Joint organization of conferences

- Learning with functional data workshop in **Lille** (France, October 2016), co-funded by the CNRS "Défi Mastodons". Stéphane Girard and Serge Iovleff co-organized the workshop while Sophie Dabo-Niang was an invited speaker.
- Financial and Actuarial Mathematics conference in **Mbour** (Sénégal, July 2016). The conference was organized in collaboration with AIMS Sénégal (African Institute for Mathematical Science) and the SWMA (Senegalese Women in Mathematics Association). Stéphane Girard and Aliou Diop were part of the scientific committee while Sophie Dabo Niang was part of the organizing committee. El Hadji Deme was an invited speaker.
- Méthodes statistiques pour l'évaluation des risques extrêmes: application à l'environnement, l'alimentation et l'assurance school in **Saint-Louis** (Sénégal, April 2016). This CIMPA school was co-organized by Sophie Dabo-Niang and Aliou Diop. The scientific committee was headed by Stéphane Girard.

Courses given in Sénégal

- Stéphane Girard gave a course on *extreme-value analysis* (master level) at the CIMPA school in Saint-Louis (Sénégal), April 2016. Other courses were given by Liliane Bel (AgroParisTech), Patrice Bertail (Université Paris 10), Anthony Davison (EPFL), Thomas Mikosch (Copenhagen University), Jessica Tressou (AgroParisTech).
- Serge Iovleff gave courses (master level) on *stochastic processes* at AIMS, Mbour (Sénégal), November 2016 and on *Monte-Carlo methods for big data classification*, at Université Gaston Berger, Saint-Louis (Sénégal), April 2018.

Publications

Co-publications between members of SIMERGE or directly related to the project objectives

	2015	2016	2017	2018
PhD Theses	1	3	0	1 (*)
Journals	2	2	3	2
Conference proceedings	2	3	3	3
Technical reports	0	0	5	3

(*) Clément Albert should defend his PhD thesis in November, 2018.

Collaborations and External support

Collaborations

- **Axis 1:** A. Daouia (Univ. Toulouse, France), J.F. Dupuy (Univ. Rennes, France), A. Dutfoy (Electricité de France - EDF), L. Gardes & A. Guillou (Univ. Strasbourg, France), A. Necir (Biskra Univ., Algeria), G. Stupfler (Nottingham Univ., UK).
- **Axis 2:** J. Chanussot (Grenoble-INP, France), C. Sokhna (IRD Dakar, Sénégal).

External support

- **Visits:** Several visits Mistis, France were partially funded by the Service de Coopération et d'Action Culturelle de l'Ambassade de France à Dakar (SCAC), the Coopération et Mobilité Internationales Rhône-Alpes (CMIRA) program and the Centre d'Excellence Africain en Mathématiques, Informatique et TIC (CEA-MITIC).
- **PhD theses:** The PhD thesis of C. Albert is co-funded by EDF. The PhD thesis of A. Chiancone was funded by the LabEx Persyval-Lab, France through its Advanced Data Mining program.

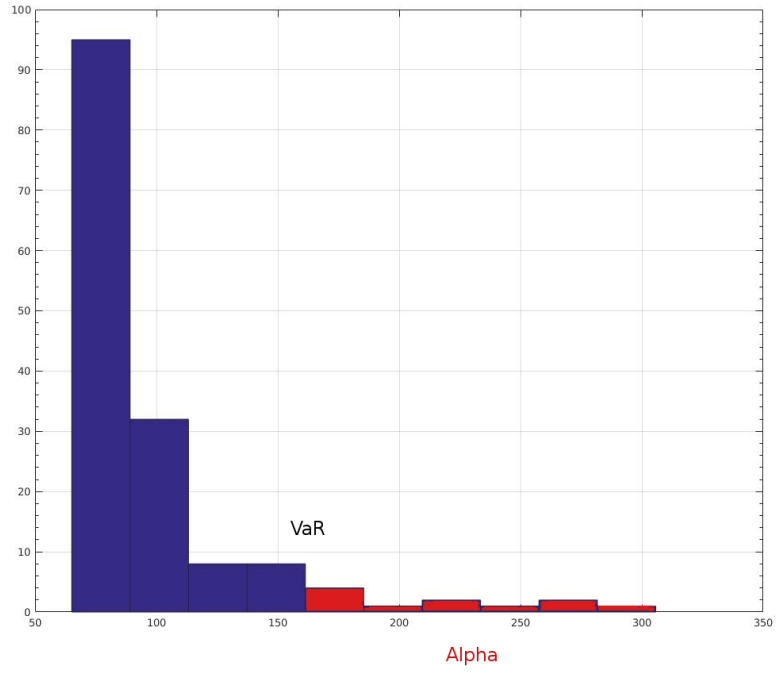
2. Reduced-bias estimators of the Conditional Tail Expectation for heavy-tailed distributions

E. Deme, A. Guillou & S. Girard, In M. Hallin et al, eds, *Mathematical Statistics and Limit Theorems*, pages 105–123, Springer, 2015.

Different risk measures have been proposed in the literature:

- The most popular one is the **Value-at-Risk** (VaR), it is defined as the $\alpha-$quantile $\mathbb{P}(X > q(\alpha)) = \alpha$ for $\alpha \in (0, 1)$ where X is the random variable of interest.
- A second important risk measure is the **Conditional Tail Expectation** (CTE) sometimes referred to as **Expected Shortfall** defined by $\mathbb{C}(\alpha) = \mathbb{E}(X|X > q(\alpha))$ when $\mathbb{E}(|X|) < \infty$.

The CTE satisfies all the desirable properties of a coherent risk measure (whereas the VaR does not) and it provides a more conservative measure of risk than the VaR for the same level of confidence α. For all these reasons, the CTE is preferable in many applications.

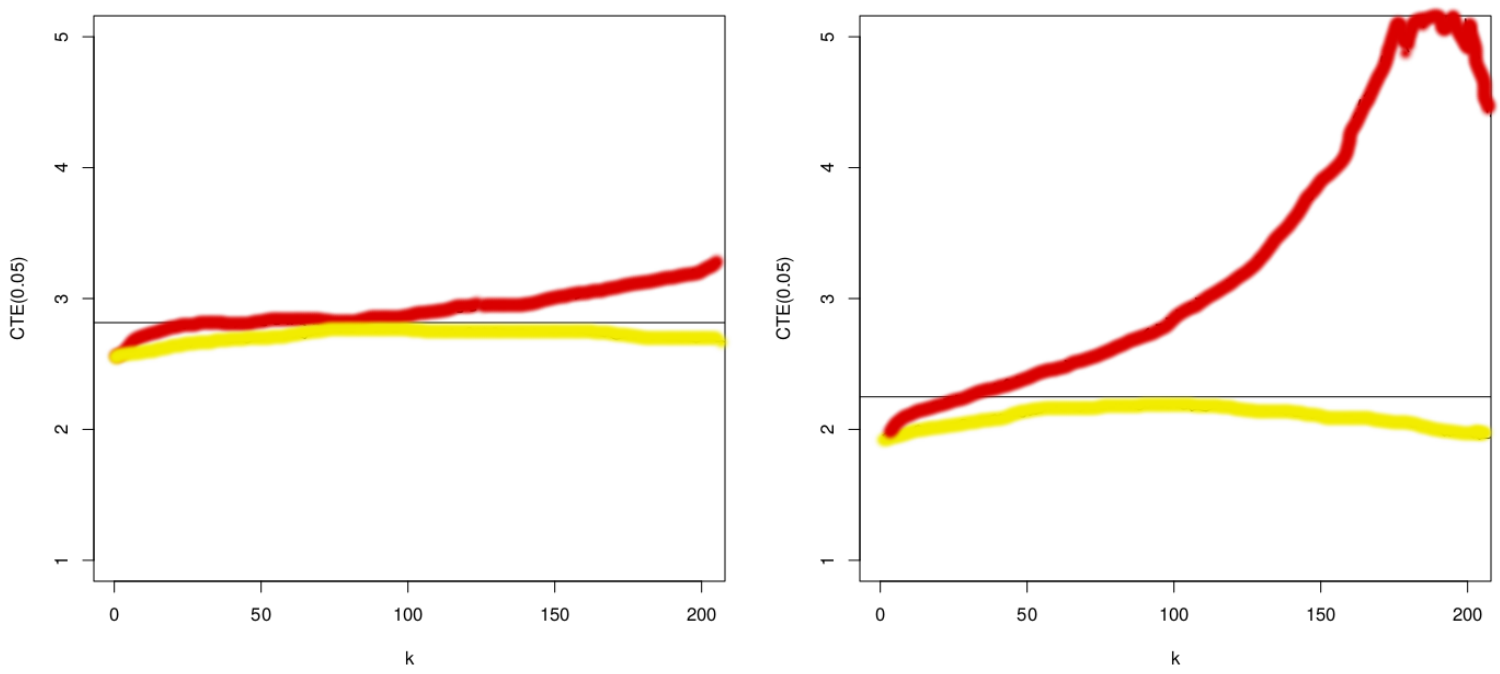


Here $\alpha = 7\%$, the VaR is $q(\alpha) \simeq 160$ and the CTE is the mean of the losses above the VaR (depicted in red), $\mathbb{C}(\alpha) \simeq 230$.

Assming that X is continuous, the CTE can be rewritten as

$$\mathbb{C}(\alpha) = \frac{1}{1-\alpha}\int_{\alpha}^{1} q(s)ds.$$

- Brazauskas *et al.* (2008) proposed an estimator of the CTE based on empirical quantiles to estimate $q(\cdot)$. The asymptotic behavior of this estimator is established when $\mathbb{E}(X^2) < \infty$ which is quite a **restrictive condition.**
- Necir *et al.* (2010) proposed an extension to the case $\mathbb{E}(X^2) = \infty$ using extreme-value based estimators of $q(\cdot)$. However, this estimator may suffer from a **high bias in finite sample situations**.
- In Deme *et al.* (2015), we evaluated the asymptotic bias and proposed a bias correction.

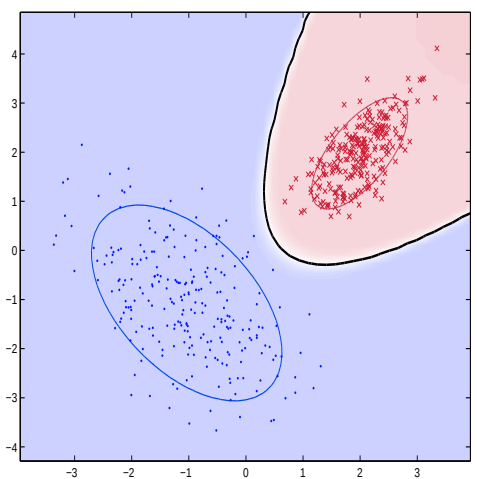


Horizontally: tuning parameter, Vertically: CTE. Red: Estimator from Necir *et al.* (2010), yellow: Estimator from Deme *et al.* (2015), horizontal line: True value. The estimators are computed on samples of size $n = 500$ from two Burr distributions (left & right).

3. A classification method for binary predictors combining similarity measures and mixture models

S. Sylla, S. Girard, A. Diongue, A. Diallo & C. Sokhna, *Dependence Modeling*, 3, 240–255, 2015.

- **Supervised classification** aims to build a decision rule for assigning an observation x in a space E with unknown class membership to one of L known classes $C_1, \ldots, C_L$. A learning dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is used, where an observation is denoted by $x_i \in E$ and its label is denoted by $y_i \in \{1, \ldots, L\}$ for $i = 1, \ldots, n$.
- **Model-based classification** assumes that the predictors $\{x_1, \ldots, x_n\}$ are independent realizations of a random vector X on E and that the class conditional distribution of X is parametric. When $E = \mathbb{R}^p$, the Gaussian distribution is often adopted.

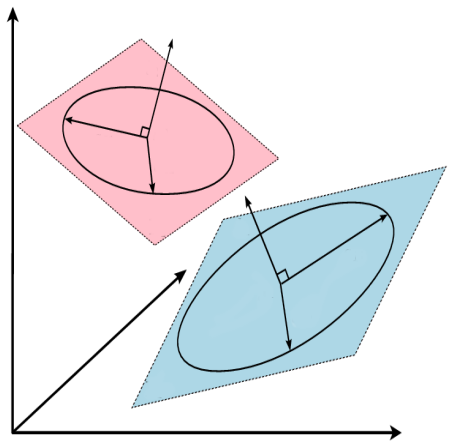


- Only few works exist to handle **categorical data**. They are usually based on multinomial or Dirichlet distributions.
- Recently, a new classification method, referred to as *parsimonious Gaussian process Discriminant Analysis* (pgpDA), has been proposed (Bouveyron *et al*, 2015) to tackle the case of **data of arbitrary nature**. The basic idea is to introduce a kernel function in the Gaussian classification rule.

⇨ Contribution:
We focus on the application of the pgpDA method to binary predictors. To this end, we show how new kernels can be built basing on similarity or dissimilarity measures. An application to verbal autopsy data is presented.

⇨ pgpDA:

- Most classification algorithms can be turned into kernel ones as far as they depend on the data only in terms of dot products. The dot product is simply changed to a kernel evaluation $K(\cdot, \cdot)$, leading to a transformation of linear algorithms to non-linear ones. Additionally, a nice property of kernel learning algorithms is the possibility to deal with any kind of data. The **first ingredient** of pgpDA is the application of this technique to the Gaussian classification rule.
- The **second ingredient** is the assumption that data of each class C_k live in its own subspace of dimension d_k in the kernel space.



These two points give rise to a nonlinear classification rule (not given here), which only depends on kernel evaluations $K(\cdot, \cdot)$ measuring the similarity between individuals.

⇨ Similarity measures

- Let x, x' be two vectors in $\{0,1\}^p$ and introduce $a =< x, x' >$, $b =< \mathbf{1} - x, x' >$, $c =< x, \mathbf{1} - x' >$ and $d =< \mathbf{1} - x, \mathbf{1} - x' >$, where $< ., . >$ is usual scalar product on $\mathbb{R}^p$ and $\mathbf{1} = (1, \ldots, 1)^T$ in $\mathbb{R}^p$. The integer a is often referred to as the **intersection** of x and x', $(b + c)$ is the **difference** and d is the **complement intersection**. Note that one always has $a + b + c + d = p$.
- The review article (Seung-Seok *et al*, 2010) lists 76 examples of similarity measures based on the 4 quantities a, b, c and d.
- We shall also consider the similarity measure

$$S_{\text{Sylla \& Girard}}(x, x') = \alpha a + (1 - \alpha)d.$$

It can be interpreted as an extension of Intersection, Russell & Rao, Sokal & Michener and Innerproduct measures, see (Seung-Seok *et al*, 2010). Here, the parameter α permits to balance the relative weights of positive and negative matches.

⇨ Kernels for binary predictors

- In the binary framework, the RBF kernel can be built from the Hamming similarity measure:

$$\begin{aligned} K_{\mathsf{RBF}}(x, x') &= \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) = \exp\left(-\frac{b+c}{2\sigma^2}\right) \\ &= \exp\left(\frac{S_{\mathsf{Hamming}}(x, x')}{2\sigma^2}\right), \end{aligned}$$

  for all x, $x' \in \{0,1\}^p$.

- We propose to extend this principle to any similarity measure S to define new kernels adapted to binary predictors:

$$K(x, x') := \exp\left(\frac{S(x, x')}{2\sigma^2}\right).$$

⇨ Application to verbal autopsy data

- The goal of verbal autopsy is to get some information from family about the circumstances of a death when medical certification is incomplete or absent. In such a situation, verbal autopsy can be used as a routine death registration. A list of p possible symptoms is established and the collected data $X = (X_1, \ldots, X_p)$ consist of the absence or presence (encoded as 0 or 1) of each symptom on the deceased person. The probable cause of death is assigned by a physician and is encoded as a qualitative random variable Y. **The aim of the classification method is assign causes of death Y from verbal autopsy data X.**
- We focus on data measured on the deceased persons during the period from 1985 to 2010 in the three IRD sites (Niakhar, Bandafassi and Mlomp) in Sénégal. The dataset includes $n = 2.500$ individuals (deceased persons) distributed in $L = 22$ classes (causes of death) and characterized by $p = 100$ variables (symptoms).

Kernel	**CCR (learning set)**	**CCR (test set)**
Euclid	88.0	83.8
Pearson	87.7	83.2
Hellinger	87.7	83.2
Dice	87.3	83.0
3w-Jaccard	87.2	82.9
Ochia1	87.2	82.8
Gower & Legendre	86.6	82.6
Roger & Tanimoto	85.9	82.4
Sylla & Girard	85.8	81.5
Godman & Kruskal	84.3	80.8
Sokal & Sneath 5	84.7	80.5
Sokal & Sneath 1	83.4	78.7

Top 12 results over the 76 kernels built from (Seung-Seok *et al*, 2010) in terms of CCR (Correct Classification Rate). The train set includes 63% individuals from the initial dataset. The results are satisfying. As a comparison, a classical multinomial model yields CCR$\simeq 50\%$.

	pgpDA		SVM		kNN	
α	**CCR learn. set**	**CCR test set**	**CCR learn. set**	**CCR test set**	**CCR learn. set**	**CCR test set**
0.1	86.9	76.3	85.3	74.6	64.5	53.1
0.2	86.6	76.4	79.9	70.8	67.4	57.6
0.3	86.1	76.0	79.5	70.4	68.3	59.5
0.4	86.1	76.1	76.0	67.9	69.1	60.9
0.5	85.8	76.1	72.7	65.3	69.0	61.0
0.6	84.3	74.9	70.3	63.5	69.2	61.8
0.7	83.4	74.2	69.2	62.6	68.3	60.9
0.8	83.3	74.1	68.7	62.2	68.5	60.9
0.9	82.8	73.7	68.2	61.7	67.7	59.8
1	82.1	72.0	67.6	61.2	64.6	56.4

Sylla & Girard kernel is plugged into pgpDA, Support Vector Machines (SVM) and k Nearest Neighbours (kNN) methods for $\alpha \in \{0.1, 0.2, \ldots, 1\}$. The CCR associated with the parameter α selected by a cross-validation procedure is in blue. On this application, it appears that pgpDA $>$ SVM $>$ kNN.

4. Future of the partnership

We wish to pursue these collaborations in the framework of SIMERG2E which is the follow-up of SIMERGE focused on (mainly) the same two axes:

1. Spatial extremes, application to management of extreme risks
2. Classification, application to genetics and global epidemiology

As emphasized above, genetic applications are also considered (see next slide). Moreover, some changes will be brought to the initial team composition:

- As already mentioned, Serge Iovleff (Modal) has joined the team. His expertise on model-based classification and software development should allow to strengthen our forces on the second axis.
- We included the 4-year group in Bioinformatics, Biostatistics and Modelling (G4-BBM) from the Institut Pasteur de Dakar. This team is headed by Cheikh Loucoubar and recruited our former PhD student Seydou Nourou Sylla as a data scientist.

In the first axis, we shall introduce semi-parametric models for extreme risk measures: Location and scale parameters may depend on the covariate, while the tail-index does not $\Longrightarrow$ Uncoupling of extreme and non-parametric effects for a better rate of convergence.

In the second axis, we shall address the challenge to build new probabilistic models sufficiently complex to handle complex dependences and yet sufficiently simple to be estimated in high dimension.

- SIMERGE: verbal autopsy data (Sylla *et al*, 2016): $p = 100$ variables measured on $n = 2,500$ individuals.
- SIMERG2E: SNP (Single Nucleotide Polymorphism) data: $p = 719,656$ variables measured on $n = 481$ individuals.

The situation $p \gg n$ is an obstacle to most statistical methods, all the more so as the individuals may be dependent due to kinship and/or repeated measurements. This phenomenon further reduces the number of independent observations.