Scientific contributions with PhD students

Stéphane Girard

Inria Rhône-Alpes & LJK (team MISTIS).

655, avenue de l'Europe, Montbonnot. 38334 Saint-Ismier Cedex, France

Stephane.Girard@inria.fr

Abstract: This document summarizes my past and present joint research works with my former PhD students.

1 Myriam Garrido (Researcher, INRA)

Extreme value theory is a branch of statistics dealing with the extreme deviations from the bulk of probability distributions. More specifically, it focuses on the limiting distributions for the minimum or the maximum of a large collection of random observations from the same arbitrary (unknown) distribution. Let $x_1 < \cdots < x_n$ denote n ordered observations from a random variable X representing some quantity of interest. A p_n-quantile of X is the value q_{p_n} such that the probability that X is greater than q_{p_n} is p_n, i.e. $P(X > q_{p_n}) = p_n$. When $p_n < 1/n$, such a quantile is said to be extreme since it is usually greater than the maximum observation x_n. To estimate such extreme quantiles requires therefore specific methods to extrapolate information beyond the observed values of X. Those methods are based on Extreme value theory. This kind of issues appeared in hydrology. One objective was to assess risk for highly unusual events, such as 100-year floods, starting from flows measured over 50 years. The decay of the survival function $P(X > x) = 1 - F(x)$, where F denotes the cumulative distribution function associated to X, is driven by a real parameter called the extreme-value index γ. We proposed Bayesian estimators of γ, see [1]. The choice of a tail model is an important issue, we proposed a goodness-of-fit test [2, 3], see [4] for its implementation.

2 Laurent Gardes (Professor, Strasbourg)

We proposed several estimators for the parameter γ, see [5, 6, 7]. When this parameter is positive, the survival function is said to be heavy-tailed, when this parameter is

negative, the survival function vanishes above its right end point. If the parameter γ is zero, then the survival function decreases to zero at an exponential rate. An important part of our work is dedicated to the study of such distributions. For instance, in reliability, the distributions of interest are included in a semi-parametric family whose tails are decreasing exponentially fast. These so-called Weibull tail-distributions encompass a variety of light-tailed distributions, such as Weibull, Gaussian, gamma and logistic distributions. Let us recall that a cumulative distribution function F has a Weibull tail if it satisfies the following property: There exists $\theta > 0$ such that for all $\lambda > 0$,

$$\lim_{y\to\infty} \frac{\log(1 - F(\lambda y))}{\log(1 - F(y))} = \lambda^{1/\theta}.$$

Dedicated methods have been proposed to estimate the Weibull tail-coefficient θ since the relevant information is only contained in the extreme upper part of the sample. More specifically, the estimators I proposed are based on the log-spacings between the upper order statistics [8, 9, 10, 11]. See also [12, 13, 14, 15] for the estimation of the associated extreme quantiles. We also addressed the estimation of extreme level curves. This problem is equivalent to estimating quantiles when covariate information is available and when their order converges to one as the sample size increases. We show that, under some conditions, these so-called "extreme conditional quantiles" can still be estimated through a kernel estimator of the conditional survival function. Sufficient conditions on the rate of convergence of their order to one are provided to obtain asymptotically Gaussian distributed estimators. Making use of this result, some estimators of the extreme-value parameters are introduced and extreme conditional quantiles estimators are deduced [16, 17, 18, 19, 20, 21, 22, 23, 24]. Finally, the tail copula is widely used to describe the dependence in the tail of multivariate distributions. In some situations such as risk management, the dependence structure may be linked with some covariate. The tail copula thus depends on this covariate and is referred to as the conditional tail copula. The aim of [18] is to propose a nonparametric estimator of the conditional tail copula and to establish its asymptotic normality. In the multivariate context, we focus on extreme geometric quantiles [25]. Their asymptotics are established, both in direction and magnitude, under suitable integrability conditions, when the norm of the associated index vector tends to one. Applications of extreme-value theory are found in hydrology [26, 27, 28] and more generally in risk estimation [29].

Sliced Inverse Regression (SIR) is an effective method for dimension reduction in high-dimensional regression problems. The original method, however, requires the inversion of the predictors covariance matrix. In case of collinearity between these predictors or small sample sizes compared to the dimension, the inversion is not possible and a regularization technique has to be used. The proposed approach is based on a Fisher Lecture given by R.D. Cook where it is shown that SIR axes can be interpreted as solutions of an inverse regression problem. A Gaussian prior is introduced on the

distribution on the unknown parameters of the inverse regression problem in order to regularize their estimation [30]. We showed that some existing SIR regularizations can enter this framework, which permits a global understanding of these methods [31]. Three new priors are proposed leading to new regularizations of the SIR method. A comparison on simulated data as well as an application to the estimation of Mars surface physical properties from hyperspectral images are provided [32].

3 Charles Bouveyron (Professor, Paris)

Clustering in high-dimensional spaces is a recurrent problem in many fields of science, for example in image analysis. Indeed, the data used in image analysis are often high-dimensional and this penalizes clustering methods. In this paper, we focus on model-based clustering method. Popular clustering methods are based on the Gaussian mixture model and show a disappointing behavior when the size of the dataset is too small compared to the number of parameters to estimate. This well-known phenomenon is called *curse of dimensionality*.

To avoid overfitting, it is necessary to find a balance between the number of parameters to estimate and the generality of the model. I proposed a Gaussian mixture model which takes into account the specific subspace in which each cluster is located and therefore limits the number of parameters to estimate. The Expectation-Maximization algorithm is used for parameter estimation and the intrinsic dimension of each group is determined automatically either with the scree-test of Cattell or by maximum likelihood [33]. This allows to derive a robust clustering method in high-dimensional spaces, called High Dimensional Data Clustering (HDDC) [34]. The method has also been adapted to supervised classification (HDDA – High Dimensional Data Analysis) [35, 36] and to the label noise situation [37]. In order to further limit the number of parameters, it is possible to make additional assumptions on the model. We can for example assume that classes are spherical in their subspaces or fix some parameters to be common between classes. Finally, HDDA and HDDC are evaluated and compared to standard clustering or classification methods on artificial and real datasets. These approaches are shown to outperform existing clustering methods [38]. The methods are implemented in a R package [39, 40] which is freely available on the CRAN archive. Finally, the extension to the classification of non necessarily quantitative data is investigated in [41, 42].

4 Alexandre Lekina (Engineer, Lille)

The PhD thesis of Alexandre Lekina was co-advised with Laurent Gardes, See Section 2 for a description of the associated publications [21, 22].

5 El-hadji Deme (Assistant Professor, Sénégal)

We focussed on the estimation of the extreme-value index [5] and of some risk measures (Conditional Tail Expectation and Proportional Hazard Premium) in case of heavy-tailed distributions [43, 44].

6 Jonathan Elmethni (Assistant Professor, Paris)

The PhD thesis of Jonathan Elmethni was co-advised with Laurent Gardes, See Section 2 for a description of the associated publications [29, 27, 13].

7 Gilles Stupfler (Assistant Professor, UK)

A part of our work focussed on the case where the parameter γ is negative and thus the survival function vanishes above its right end point. Some estimation methods for the right end point have been proposed in [45, 46]. When a covariate is available, the right end point is a function referred to as the frontier. The estimation of the frontier is addressed in [47, 48].

A popular way to study the tail of a distribution function is to consider its high or extreme quantiles. While this is a standard procedure for univariate distributions, it is harder for multivariate ones, primarily because there is no universally accepted definition of what a multivariate quantile should be. In [25, 49], we focus on extreme geometric quantiles. Their asymptotics are established, both in direction and magnitude.

8 Gildas Mazo (Postdoc, Belgium)

A bivariate copula defined on the unit square $[0,1]^2$ is a bivariate cumulative distribution function (cdf) with univariate uniform margins. Sklar's Theorem states that any bivariate distribution with cdf H and marginal cdf F and G can be written $H(x,y) = C(F(x), G(y))$, where C is a copula. This result justifies the use of copulas for building bivariate distributions. While there exist various families of bivariate copulas, the construction of flexible and yet tractable copulas suitable for high-dimensional applications is much more challenging. This is even more true if one is concerned with the analysis of extreme values. In [50, 51], we construct a class of one-factor copulas and a family of extreme-value copulas well suited for high-dimensional applications and exhibiting a good balance between tractability and flexibility. The inference for these copulas is performed by using a least-squares estimator based on dependence coefficients [52]. The modeling capabilities of the copulas are illustrated on simulated

and real datasets. This class of copula is extended in [53]. In [54], we propose a class of multivariate copulas based on products of transformed bivariate copulas. No constraints on the parameters refrain the applicability of the proposed class. Furthermore the analytical forms of the copulas within this class allow to naturally associate a graphical structure which helps to visualize the dependencies and to compute the likelihood efficiently even in high dimension.

We also worked on the application to extreme-value methods to hydrology [55].

References

[1] J. Diebolt, M. El-Aroui, M. Garrido, and S. Girard. Quasi-conjugate Bayes estimates for GPD parameters and application to heavy tails modelling. *Extremes*, 8:57–78, 2005.

[2] J. Diebolt, M. Garrido, and S. Girard. A goodness-of-fit test for the distribution tail. In M. Ahsanulah and S. Kirmani, editors, *Extreme Value Distributions*, pages 95–109. Nova Science, New-York, 2007.

[3] J. Diebolt, M. Garrido, and S. Girard. Asymptotic normality of the ET method for extreme quantile estimation. Application to the ET test. *Comptes-Rendus de l'Académie des Sciences, Série I*, 337:213–218, 2003.

[4] J. Diebolt, J. Ecarnot, M. Garrido, S. Girard, and D. Lagrange. Le logiciel Extremes, un outil pour l'étude des queues de distribution. *La revue de Modulad*, 30:53–60, 2003.

[5] E. Deme, L. Gardes, and S. Girard. On the estimation of the second order parameter for heavy-tailed distributions. *REVSTAT - Statistical Journal*, 11:277–299, 2013.

[6] L. Gardes and S. Girard. Asymptotic properties of a Pickands type estimator of the extreme value index. In Louis R. Velle, editor, *Focus on probability theory*, pages 133–149. Nova Science, New-York, 2006.

[7] L. Gardes and S. Girard. Asymptotic distribution of a Pickands type estimator of the extreme value index. *Comptes-Rendus de l'Académie des Sciences, Série I*, 341:53–58, 2005.

[8] L. Gardes, S. Girard, and A. Guillou. Weibull tail-distributions revisited: a new look at some tail estimators. *Journal of Statistical Planning and Inference*, 141(1):429–444, 2011.

[9] L. Gardes and S. Girard. Estimation of the Weibull tail-coefficient with linear combination of upper order statistics. *Journal of Statistical Planning and Inference*, 138:1416–1427, 2008.

[10] J. Diebolt, L. Gardes, S. Girard, and A. Guillou. Bias-reduced estimators of the Weibull tail-coefficient. *Test*, 17:311–331, 2008.

[11] L. Gardes and S. Girard. Comparison of Weibull tail-coefficient estimators. *REVSTAT - Statistical Journal*, 4(2):373–188, 2006.

[12] L. Gardes and S. Girard. Estimation de quantiles extrêmes pour les lois à queue de type Weibull : une synthèse bibliographique. *Journal de la Société Française de Statistique*, 154:98–118, 2013.

[13] J. El Methni, L. Gardes, S. Girard, and A. Guillou. Estimation of extreme quantiles from heavy and light tailed distributions. *Journal of Statistical Planning and Inference*, 142(10):2735–2747, 2012.

[14] J. Diebolt, L. Gardes, S. Girard, and A. Guillou. Bias-reduced extreme quantiles estimators of Weibull-tail distributions. *Journal of Statistical Planning and Inference*, 138:1389–1401, 2008.

[15] L. Gardes and S. Girard. Estimating extreme quantiles of Weibull tail-distributions. *Communication in Statistics - Theory and Methods*, 34:1065–1080, 2005.

[16] L. Gardes and S. Girard. Functional kernel estimators of conditional extreme quantiles. In F. Ferraty, editor, *Recent advances in functional data analysis and related topics*, pages 135–140. Springer, Physica-Verlag, 2011.

[17] L. Gardes and S. Girard. On the estimation of the functional Weibull tail-coefficient. *Journal of Multivariate Analysis*, 146:29–45, 2016.

[18] L. Gardes and S. Girard. Nonparametric estimation of the conditional tail copula. *Journal of Multivariate Analysis*, 137:1–16, 2015.

[19] A. Daouia, L. Gardes, and S. Girard. On kernel smoothing for extremal quantile regression. *Bernoulli*, 19:2557–2589, 2013.

[20] L. Gardes and S. Girard. Functional kernel estimators of large conditional quantiles. *Electronic Journal of Statistics*, 6:1715–1744, 2012.

[21] A. Daouia, L. Gardes, S. Girard, and A. Lekina. Kernel estimators of extreme level curves. *Test*, 20(14):311–333, 2011.

[22] L. Gardes, S. Girard, and A. Lekina. Functional nonparametric estimation of conditional extreme quantiles. *Journal of Multivariate Analysis*, 101:419–433, 2010.

[23] L. Gardes and S. Girard. A moving window approach for nonparametric estimation of the conditional tail index. *Journal of Multivariate Analysis*, 99:2368–2388, 2008.

[24] A. Daouia, L. Gardes, and S. Girard. Nadaraya's estimates for large quantiles and free disposal support curves. In I. Van Keilegom and P. Wilson, editors, *Exploring research frontiers in contemporary statistics and econometrics*, pages 1–22. Springer, 2012.

[25] S. Girard and A. G. Stupfler. Intriguing properties of extreme geometric quantiles. *REVSTAT - Statistical Journal*, 2016. to appear.

[26] J. Carreau, D. Ceresetti, E. Ursu, S. Anquetin, J.D. Creutin, L. Gardes, S. Girard, and G. Molinié. Evaluation of classical spatial-analysis schemes of extreme rainfall. *Natural Hazards and Earth System Sciences*, 12:3229–3240, 2012.

[27] J. El Methni, L. Gardes, and S. Girard. Estimation de mesures de risque pour des pluies extrêmes dans la région Cévennes Vivarais. *La Houille Blanche*, 4:46–51, 2015.

[28] L. Gardes and S. Girard. Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels. *Extremes*, 13(2):177–204, 2010.

[29] J. El Methni, L. Gardes, and S. Girard. Nonparametric estimation of extreme risks from conditional heavy-tailed distributions. *Scandinavian Journal of Statistics*, 41:988–1012, 2014.

[30] C. Bernard-Michel, L. Gardes, and S. Girard. Gaussian regularized sliced inverse regression. *Statistics and Computing*, 19:85–98, 2009.

[31] C. Bernard-Michel, L. Gardes, and S. Girard. A note on sliced inverse regression with regularizations. *Biometrics*, 64:982–986, 2008.

[32] C. Bernard-Michel, S. Douté, M. Fauvel, L. Gardes, and S. Girard. Retrieval of Mars surface physical properties from Omega hyperspectral images using regularized sliced inverse regression. *Journal of Geophysical Research - Planets*, 114, 2009. E06005.

[33] C. Bouveyron, G. Celeux, and S. Girard. Intrinsic dimension estimation by maximum likelihood in isotropic probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713, 2011.

[34] C. Bouveyron, S. Girard, and C. Schmid. High dimensional data clustering. *Computational Statistics and Data Analysis*, 52:502–519, 2007.

[35] C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. *Communication in Statistics - Theory and Methods*, 36(14):2607–2623, 2007.

[36] C. Bouveyron, S. Girard, and C. Schmid. Class-specific subspace discriminant analysis for high-dimensional data. In C. Saunder et al., editor, *Lecture Notes in Computer Science*, volume 3940, pages 139–150. Springer-Verlag, Berlin Heidelberg, 2006.

[37] C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658, 2009.

[38] J. Jacques, C. Bouveyron, S. Girard, O. Devos, L. Duponchel, and C. Ruckebusch. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Journal of Chemometrics*, 24:719–727, 2010.

[39] L. Bergé, C. Bouveyron, and S. Girard. HDclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, 46(6):1–29, 2012.

[40] C. Bouveyron and S. Girard. Classification supervisée et non supervisée des données de grande dimension. *La revue de Modulad*, 40:81–102, 2009.

[41] C. Bouveyron, M. Fauvel, and S. Girard. Kernel discriminant analysis and clustering with parsimonious Gaussian process models. *Statistics and Computing*, 25:1143–1162, 2015.

[42] M. Fauvel, C. Bouveyron, and S. Girard. Parsimonious Gaussian process models for the classification of hyperspectral remote sensing images. *IEEE Geoscience and Remote Sensing Letters*, 12(12):2423–2427, 2015.

[43] E. Deme, S. Girard, and A. Guillou. Reduced-bias estimator of the proportional hazard premium for heavy-tailed distributions. *Insurance: Mathematics and Economics*, 22:550–559, 2013.

[44] E. Deme, S. Girard, and A. Guillou. Reduced-bias estimators of the conditional tail expectation for heavy-tailed distributions. In M. Hallin et al., editor, *Mathematical Statistics and Limit Theorems*, pages 105–123. Springer, 2015.

[45] S. Girard, A. Guillou, and G. Stupfler. Estimating an endpoint with high order moments in the Weibull domain of attraction. *Statistics and Probability Letters*, 82:2136–2144, 2012.

[46] S. Girard, A. Guillou, and G. Stupfler. Estimating an endpoint with high order moments. *Test*, 21:697–729, 2012.

[47] S. Girard, A. Guillou, and G. Stupfler. Uniform strong consistency of a frontier estimator using kernel regression on high order moments. *ESAIM: Probability and Statistics*, 18:642–666, 2014.

[48] S. Girard, A. Guillou, and G. Stupfler. Frontier estimation with kernel regression on high order moments. *Journal of Multivariate Analysis*, 116:172–189, 2013.

[49] S. Girard and A. G. Stupfler. Extreme geometric quantiles in a multivariate regular variation framework. *Extremes*, 18(4):629–663, 2015.

[50] G. Mazo, S. Girard, and F. Forbes. A class of multivariate copulas based on products of bivariate copulas. *Journal of Multivariate Analysis*, 140:363–376, 2015.

[51] F. Durante, S. Girard, and G. Mazo. Copulas based on Marshall-Olkin machinery. In U. Cherubini et al., editor, *Marshall-Olkin Distributions. Advances in Theory and Applications*, volume 141 of *Springer Proceedings in Mathematics and Statistics*, pages 15–31. Springer, 2015.

[52] G. Mazo, S. Girard, and F. Forbes. Weighted least square inference based on dependence coefficients for multivariate copulas. *ESAIM: Probability and Statistics*, 19:746–765, 2015.

[53] F. Durante, S. Girard, and G. Mazo. Marshall-Olkin type copulas generated by a global shock. *Journal of Computational and Applied Mathematics*, 296:638–648, 2016.

[54] G. Mazo, S. Girard, and F. Forbes. A flexible and tractable class of one-factor copulas. *Statistics and Computing*, 2016. to appear.

[55] B. Barroca, P. Bernardara, S. Girard, and G. Mazo. Considering hazard estimation uncertain in urban resilience strategies. *Natural Hazards and Earth System Sciences*, 15:25–34, 2015.