Research highlights in 2017

Stéphane Girard

Inria Grenoble Rhône-Alpes & LJK (team MISTIS).

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

Stephane.Girard@inria.fr

Abstract

This note summarizes my research outputs in 2017. Two main research domains are investigated: High dimensional statistical learning and Extreme-value analysis,

1 High dimensional statistical learning

Sliced Inverse Regression (SIR) is an effective method for dimensionality reduction in high-dimensional regression problems. However, the method has requirements on the distribution of the predictors that are hard to check since they depend on unobserved variables. It has been shown that, if the distribution of the predictors is elliptical, then these requirements are satisfied. In case of mixture models, the ellipticity is violated and in addition there is no assurance of a single underlying regression model among the different components. Our approach clusterizes the predictors space to force the condition to hold on each cluster and includes a merging technique to look for different underlying models in the data [1]. Moreover, SIR is originally a model free method but it has been shown to actually correspond to the maximum likelihood of an inverse regression model with Gaussian errors. This intrinsic Gaussianity of standard SIR may explain its high sensitivity to outliers as observed in a number of studies. To improve robustness, the inverse regression formulation of SIR is therefore extended to non-Gaussian errors with heavy-tailed distributions. Considering Student distributed errors it is shown that the inverse regression remains tractable via an Expectation- Maximization (EM) algorithm. The algorithm is outlined and tested in the presence of outliers, both in simulated and real data, showing improved results in comparison to a number of other existing approaches [2].

Besides, I worked on the classification of grasslands using high resolution satellite image time series. Grasslands considered in this work are semi-natural elements in fragmented landscapes, i.e., they are heterogeneous and small elements. The first contribution of this study is to account for grassland heterogeneity while working at the object scale by modeling its pixels distributions by a Gaussian distribution. To measure the similarity between two grasslands, a new kernel is proposed as a second contribution: the a-Gaussian mean kernel. It allows to weight the influence of the covariance matrix when comparing two Gaussian distributions. This kernel is introduced in Support Vector Machine for the supervised classification of grasslands from south-west France. A dense intra-annual multispectral time series of Formosat-2 satellite is used for the classification of grasslands management practices, while an inter-annual NDVI time series of Formosat-2 is used for permanent and temporary grasslands discrimination. Results are compared to other existing pixel- and object-based approaches in terms of classification accuracy and processing time. The proposed method shows to be a good compromise between processing speed and classification accuracy. It can adapt to the classification constraints and it encompasses several similarity measures known in the literature. It is appropriate for the classification of small and heterogeneous objects such as grasslands [3, 4, 5, 6, 7].

Finally, I worked on the application of dimension reduction methods dedicated to high dimensional classification [8] or regression [9, 10] to astrophysics [11] and medicine [12].

2 Extreme-value analysis

The decay of the survival function is driven by a real parameter called the extreme-value index. When this parameter is positive, the survival function is said to be heavy-tailed. In [13], a new estimator of the extreme-value index dedicated to this context was introduced. It was applied to ecological data in [14].

If the extreme-value index is zero, then the survival function decreases to zero at an exponential rate. An important part of my work is dedicated to the study of such distributions [15]. 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 include Gaussian, gamma, exponential and Weibull distributions, among others.

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. I focussed on extreme geometric quantiles. Their asymptotics are established, both in direction and magnitude, under suitable integrability conditions, when the norm of the associated index vector tends to one [16, 17].

I also investigated the asymptotic behavior of the (relative) extrapolation error associated with some estimators of univariate extreme quantiles based on extreme-value theory. It is shown that the extrapolation error can be interpreted as the remainder of a first order Taylor expansion. Necessary and sufficient conditions are then provided such that this error tends to zero as the sample size increases. Interestingly, in case of the so-called Exponential Tail estimator, these conditions lead to a subdivision of Gumbel maximum domain of attraction into three subsets. In constrast, the extrapolation error associated with Weissman estimator has a common behavior over the whole Fréchet maximum domain of attraction. First order equivalents of the extrapolation error are then derived and their accuracy is illustrated numerically [18, 19, 20].

References

[1] A. Chiancone, S. Girard, and J. Chanussot. Collaborative sliced inverse regression. *Communication in Statistics - Theory and Methods*, 46(12):6035–6053, 2017.

[2] A. Chiancone, F. Forbes, and S. Girard. Student sliced inverse regression. *Computational Statistics and Data Analysis*, 113:441–456, 2017.

[3] M. Lopes, M. Fauvel, A. Ouin, and S. Girard. Spectro-temporal heterogeneity measures from dense high spatial resolution satellite image time series: Application to grassland species diversity estimation. *Remote Sensing*, 9(993), 2017.

[4] M. Lopes, M. Fauvel, S. Girard, and D. Sheeren. Object-based classification of grasslands from high resolution satellite image time series using Gaussian mean map kernels. *Remote Sensing*, 9(7), 2017.

[5] S. Girard, M. Lopes, M. Fauvel, and D. Sheeren. Object-based classification of grassland management practices from high resolution satellite image time series with Gaussian mean map kernels. In *27th Annual Conference of the International Environmetrics Society*, Bergame, Italy, juillet 2017.

[6] M. Lopes, M. Fauvel, A. Ouin, and S. Girard. Potential of Sentinel-2 and SPOT5 (Take5) time series for the estimation of grasslands biodiversity indices. In *9th International Workshop on the Analysis of Multitemporal Remote Sensing Images*, Bruges, Belgium, juin 2017.

[7] M. Lopes, M. Fauvel, A. Ouin, and S. Girard. Evaluation de la biodiversité des prairies semi-naturelles par télédétection hyperspectrale. In *5ème colloque scientifique du groupe thématique hyperspectral de la Société Française de Photogrammétrie et Télédétection*, Brest, mai 2017.

[8] D. Fraix-Burnet, C. Bouveyron, S. Girard, and J. Arbel. Unsupervised classification in high dimension. In *European Week of Astronomy and Space Science (EWASS)*, Prague, République Tchèque, juin 2017.

[9] M. Fauvel, S. Girard, S. Douté, and L. Gardes. Machine learning methods for the inversion of hyperspectral images. In A. Reimer, editor, *Horizons in World Physics*, volume 290, pages 51–77. Nova Science, New-York, 2017.

[10] V. Watson, J-F. Trouilhet, F. Paletou, and S. Girard. Inference of an explanatory variable from observations in a high-dimensional space: Application to high-resolution spectra of stars. In *IEEE International Workshop of Electronics, Control, Measurement, Signals and their application to Mechatronics*, San Sebastian, Spain, mai 2017.

[11] J. Arbel, D. Fraix-Burnet, and S. Girard. Les écoles d'astrostatistique "statistics for astrophysics". In *Colloque Francophone International sur l'Enseignement de la Statistique*, Grenoble, septembre 2017.

[12] S. Sylla, S. Girard, A. Diongue, A. Diallo, and C. Sokhna. Hierarchical kernel applied to mixture model for the classification of binary predictors. In *61st ISI World Statistics Congress*, Marrakech, Morocco, juillet 2017.

[13] P. Jordanova, Z. Fabian, P. Hermann, L. Strelec, A. Rivera, S. Girard, S. Torres, and M. Stehlik. Weak properties and robustness of t-Hill estimators. *Extremes*, 19:591–626, 2016.

[14] M. Stehlik, P. Aguirre, S. Girard, P. Jordanova, J. Kiselak, S. Torres, Z. Stadovsky, and A. Rivera. On ecosystems dynamics. *Ecological Complexity*, 29:10–29, 2017.

[15] S. Girard and L. Gardes. Estimation of the functional Weibull tail-coefficient. In *10th International Conference on Extreme Value Analysis*, Delft, Netherlands, juin 2017.

[16] S. Girard and G. Stupfler. Intriguing properties of extreme geometric quantiles. *REVSTAT - Statistical Journal*, 15:107–139, 2017.

[17] G. Stupfler and S. Girard. Some negative results on extreme multivariate quantiles defined through convex optimisation. In *10th International Conference of the ERCIM WG on computing and statistics*, London, UK, décembre 2017.

[18] C. Albert, S. Girard, and A. Dutfoy. On the relative approximation error of extreme quantiles by the block maxima method. In *10th International Conference on Extreme Value Analysis*, Delft, Netherlands, juin 2017.

[19] C. Albert, A. Dutfoy, and S. Girard. On the extrapolation limits of extreme-value theory for risk management. In *10th International Conference on Mathematical Methods in Reliability*, Grenoble, juillet 2017.

[20] C. Albert, A. Dutfoy, and S. Girard. Etude de l'erreur relative d'approximation des quantiles extrêmes. In *49èmes Journées de Statistique organisées par la Société Française de Statistique*, Avignon, juin 2017.