Applications of statistics

Stéphane Girard

Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

Stephane.Girard@inria.fr

Abstract: This report summarizes my contributions in seven fields of applied statistics.

1 Environmental and climate applications

A large part of my work is dedicated to the estimation of risk measures associated with extreme weather events. It is assumed that the event distribution is heavy-tailed and depends on a covariate. The estimation method thus combines nonparametric kernel methods with extreme-value statistics. The asymptotic distribution of the estimators is established and their finite sample behavior is illustrated on simulated data. Such estimators are applied to environmental datasets such as daily rainfalls in the Cévennes-Vivarais region (France) [1, 2, 3] or to river flows (Franche-Comté region, France) [4, 5].

2 Econometrics

The estimation of optimal support boundaries under the monotonicity constraint is relatively unexplored and still in full development. In [6], a new extreme-value based model is examined. It provides a valid alternative for completely envelopment frontier models that often suffer from lack of precision, and for purely stochastic ones that are known to be sensitive to model specification. A new characterization of partial boundaries of a free disposal multivariate support, lying near the true support curve, is introduced in [7] by making use of large quantiles of a simple transformation of the underlying multivariate distribution. Pointwise empirical and smoothed estimators of the full and partial support curves are built as extreme sample and smoothed quantiles. The extreme-value theory holds then automatically for the empirical frontiers and we show that some fundamental properties of extreme order statistics carry over to Nadaraya's estimates of upper quantile-based frontiers. Different motivating applications are presented including the estimation of the minimal cost in production activity and the assessment of the reliability of nuclear reactors.

Many different premium principles have been proposed in the literature. In [8], we focus on the Proportional Hazard Premium. Its asymptotic normality has been established in the literature under suitable conditions which are not fulfilled in case of heavy tailed distributions. We thus focus on this framework and propose a reduced-bias approach for the classical estimators.

3 Image analysis

In [9], we describe a new method for the detection and reconstruction of building in dense urban areas using high resolution aerial images. Our approach begins with the generation of a dense digital elevation model (DEM). A sparse disparity map is density using a region-based segmentation of the left aerial image: each detected region is tested to be planar in the disparity map. A strategy is proposed to optimize the generation of these planar surfaces taking into account the noise present in the sparse disparity map and the robustness and complexity of different algorithms for planar approximation. The second step of our approach deals with the generation of building hypotheses. Based on the DEM previously computed, geometric and colorimetric criteria are used for the fusion of parallel regions, for the detection of symmetrical regions in the 3D object space and for the reconstruction of roof buildings. Experimental results are presented on a scene in the suburb of Bruxelles with colour images at the resolution of 10cm/pixel.

In [10], we address the problem of identifying the projection of an object from incomplete data extracted from its radiographic image. We assume that the unknown object is a particular sample of a flexible object. Our approach consists firstly in designing a deformation model able to represent and to simulate a great variety of samples of the flexible object radiographic projection. This modelization is achieved using a training set of complete data. Then, given the incomplete data, the identification task consists in estimating the observed object using the deformation model. The proposed modeling extracts from the training set, not only the deformation modes, but also other relevant information (such as probability distributions on the deformations, relations between deformations) in order to use it to regularize the identification step.

In [11], we are given a set of points in a high dimensional space. For instance, this set can represent many visual appearances of an object, a face or a hand. We address the problem of approximating this set by a manifold in order to have a compact representation of the object appearance. When the scattering of this set is approximately an ellipsoid, then the problem has a well-known solution given by Principal Components Analysis (PCA). Yet, in some situations like object deplacement learning or face learning this linear technique can be ill-adapted and nonlinear approximation must be introduced. The method we propose can be seen as a Non Linear PCA (NLPCA) [12, 13, 14], the main difficulty being that the data points are not ordered.

We propose an index to find projection axes encouraging the choice of axes which preserve as well as possible the structure of the closest point neighborhood. These axes determine an order for visiting all the points when smoothing. Finally, a new criterion, called "generalization error" is introduced to determine the smoothing rate, that is the spline number of knots in this case. The method is tested on artificial data and on two data sets coming from databases used in visual learning.

In the supervised classification framework, human supervision is required for labeling a set of learning data which are then used for building the classifier. However, in many applications, human supervision is either imprecise, difficult or expensive. In [15], the problem of learning a supervised multi-class classifier from data with uncertain labels is considered and a model-based classification method is proposed to solve it. The idea of the proposed method is to confront an unsupervised modelling of the data with the supervised information carried by the labels of the learning data in order to detect inconsistencies. The method is able afterward to build a robust classifier taking into account the detected inconsistencies into the labels. Experiments on artificial and real data are provided to highlight the main features of the proposed method as well as an application to object recognition under weak supervision.

Grasslands represent a significant source of biodiversity that is important to monitor over large extents. The Spectral Variation Hypothesis (SVH) assumes that the Spectral Heterogeneity (SH) measured from remote sensing data can be used as a proxy for species diversity. In [16], we argue the hypothesis that the grassland's species differ in their phenology and, hence, that the temporal variations can be used in addition to the spectral variations. The purpose of this study is to attempt verifying the SVH in grasslands using the temporal information provided by dense Satellite Image Time Series (SITS) with a high spatial resolution. Our method to assess the spectro-temporal heterogeneity is based on a clustering of grasslands using a robust technique for high dimensional data. We propose new SH measures derived from this clustering and computed at the grassland level. We compare them to the Mean Distance to Centroid (MDC). The method is experimented on 192 grasslands from southwest France using an intra-annual multispectral SPOT5 SITS comprising 18 images and using single images from this SITS. The combination of two of the proposed SH measuresthe within-class variability and the entropyin a multivariate linear model explained the variance of the grasslands' Shannon index more than the MDC. However, there were no significant differences between the predicted values issued from the best models using multitemporal and monotemporal imagery. We conclude that multitemporal data at a spatial resolution of 10 m do not contribute to estimating the species diversity. The temporal variations may be more related to the effect of management practices. The paper [17] deals with 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.

4 Spectroscopy data

Hyperspectral remote sensing, also known as imaging spectroscopy, is a promising space technology regularly selected by agencies with regard to the exploration and observation of planets, to earth's geology or to the monitoring of the environment. It allows to collect for each pixel of a scene, the intensity of light energy reflected from planets as it varies across different wavelengths. More than one hundred spectels in the visible and near infra-red are typically recorded, making it possible to observe a continuous spectrum for each image cell. Usually, in space exploration, the analysis of these spectral signatures allows to retrieve the physical, chemical or mineralogical properties of surfaces and of atmospheres that may help to understand the geological and climatological history of planets. We propose in this paper a statistical method to evaluate the physical properties of surface materials on Mars from hyperspectral images collected by the OMEGA instrument aboard the Mars express spacecraft. The approach we develop in [18] is based on the estimation of the functional relationship F between some physical parameters and observed spectra. For this purpose, a database of synthetic spectra is generated by a physical radiative transfer model and used to estimate F. The high dimension of spectra is reduced by using Gaussian regularized sliced inverse regression (GRSIR) [19, 20] to overcome the curse of dimensionality and consequently the sensitivity of the inversion to noise (ill-conditioned problems). Compared with a naive spectrum matching approach such as the k-nearest neighbors algorithm, estimates are more accurate and realistic [21].

In [22], a family of generative Gaussian models designed for the supervised classification of high-dimensional data is presented as well as the associated classification method called High Dimensional Discriminant Analysis (HDDA). The advantages of these Gaussian models are: i) the representation of the input density model is smooth; ii) the data of each class are modeled in a specific subspace of low dimensionality; iii) each class may have its own covariance structure; iv) regularization is coupled to the classification criterion to avoid data over-fitting. To illustrate the abilities of the method, HDDA is applied on complex high-dimensional multi-class classification problems in mid-infrared and near infrared spectroscopy and compared to state-of-the-art methods. Similarly, a family of parsimonious Gaussian process models for classification is proposed in [23]. A subspace assumption is used to build these models in the kernel feature space. By constraining some parameters of the models to be common between classes, parsimony is controlled. Experimental results are given for three real hyperspectral data sets, and comparisons are done with three others classifiers. The proposed models show good results in terms of classification accuracy and processing time.

5 Biological applications

In order to obtain reference curves for data sets when the covariate is multidimensional, we propose in [24] a new procedure based on dimension-reduction and nonparametric estimation of conditional quantiles. This semiparametric approach combines sliced inverse regression (SIR) and a kernel estimation of conditional quantiles. The asymptotic convergence of the derived estimator is shown. By a simulation study, we compare this procedure to the classical kernel nonparametric one for different dimensions of the covariate. The semiparametric estimator shows the best performance. The usefulness of this estimation procedure is illustrated on a real data set collected in order to establish reference curves for biophysical properties of the skin of healthy French women [25, 26, 27].

In [28], it is shown how a dynamical system given by a t-score function for some class of monotonic data transformations generates consistent extreme value estimators. The variation of their values increases the uncertainty of proper assessment of climate change. Two important examples illustrate the methodology: mass balance measurements on Guanaco glacier, Chile, and extreme snow loads in Slovakia. We experience singular learning of the transitions in ecosystems.

6 Power management

In [29], the optimal choice of the waiting period (or timeout) that a device should respect before entering sleep mode, so as to optimize a tradeoff between power consumption and user impact. The optimal timeout is inferred by appropriate statistical modeling of the times between user requests. In a test approach, these times are supposed independent, and a constant optimal timeout is inferred accordingly. In a second approach, some dependency is introduced through a hidden Markov chain, which also models specific activity states, like business hours or night periods. This model leads to a statistical framework for computing adaptive optimal timeout values. Different strategies are assessed using real datasets, on the basis of the power consumption, user impact and the frequency of wrong decisions.

The dramatic increase in leakage current has become a major issue for future IC designs. Moreover, as process variability in nano-scaled CMOS technologies induces a large spread of leakage power, leakage variability cannot be neglected anymore. In [30], the predominant physical process parameters for static power consumption variation are analyzed for a 32 nm technology node. The presented results are confirmed by a Principal Component Analysis. In addition, a Sliced Inverse Regression method is used to study the evolution of the impact of several parameters, like the gate-length, the oxide thickness and the doping, with the supply-voltage. To conclude, a comparative analysis with a 45nm technology is presented.

7 Network monitoring

In the context of network traffic analysis [31], we address the problem of estimating the tail index of flow size distribution from the observation of a sampled population of packets. We give an exhaustive bibliography of the existing methods and show the relations between them. The main contribution of this work is then to propose a new method to estimate the tail index from sampled data, based on the resolution of the maximum likelihood problem. To assess the performance of our method, we present a full performance evaluation based on numerical simulations, and also on a real traffic trace corresponding to internet traffic recently acquired.

References

[1] 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.

[2] 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.

[3] 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.

[4] B. Barroca, P. Bernardara, S. Girard, and G. Mazo. Considering hazard estimation uncertain in urban resilience strategies. In K. Etingoff, editor, *Ecological Resilience, Response to Climate Change and Natural Disasters*, pages 197–220. Apple Academic Press, 2016.

[5] 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.

[6] A. Daouia, S. Girard, and A. Guillou. A γ-moment approach to monotonic boundaries estimation: with applications in econometric and nuclear fields. *Journal of Econometrics*, 178:727–740, 2014.

[7] 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.

[8] 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.

[9] S. Girard, P. Guérin, H. Maître, and M. Roux. Building detection from high resolution colour images. In S.B. Serpico, editor, *Image and Signal Processing for Remote Sensing IV*, volume 3500, pages 278–289. SPIE, 1998.

[10] S. Girard, J-M. Dinten, and B. Chalmond. Building and training radiographic flexible prior models for object identification from incomplete data. *IEE proceedings on Vision, Image and Signal Processing*, 143(4):257–264, 1996.

[11] B. Chalmond and S. Girard. Nonlinear modeling of scattered multivariate data and its application to shape change. *IEEE Pattern Analysis and Machine Intelligence*, 21(5):422–432, 1999.

[12] S. Girard. A nonlinear PCA based on manifold approximation. *Computational Statistics*, 15(2):145–167, 2000.

[13] S. Girard and S. Iovleff. Auto-associative models and generalized principal component analysis. *Journal of Multivariate Analysis*, 93(1):21–39, 2005.

[14] S. Girard and S. Iovleff. Auto-associative models, nonlinear principal component analysis, manifolds and projection pursuit. In A. Gorban et al., editor, *Principal Manifolds for Data Visualisation and Dimension Reduction*, volume 28, pages 205–222. LNCSE, Springer-Verlag, 2007.

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

[16] 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.

[17] 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.

[18] 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.

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

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

[21] 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.

[22] 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.

[23] 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.

[24] A. Gannoun, S. Girard, C. Guinot, and J. Saracco. Sliced inverse regression in reference curves estimation. *Computational Statistics and Data Analysis*, 46(1):103–122, 2004.

[25] A. Gannoun, S. Girard, C. Guinot, and J. Saracco. Reference ranges based on nonparametric quantile regression. *Statistics in Medicine*, 21(20):3119–3135, 2002.

[26] A. Gannoun, S. Girard, C. Guinot, and J. Saracco. Trois méthodes non paramétriques pour l'estimation de courbes de référence - application à l'analyse de propriétés biophysiques de la peau. *Revue de Statistique Appliquée*, L(1):65–89, 2002.

[27] A. Gannoun, S. Girard, C. Guinot, and J. Saracco. Implémentation en C d'estimateurs non-paramétriques de quantiles conditionnels. Application au tracé de courbes de référence. *La revue de Modulad*, 31:59–70, 2004.

[28] 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.

[29] J.B. Durand, S. Girard, V. Ciriza, and L. Donini. Optimization of power consumption and user impact based on point process modeling of the request sequence. *Journal of the Royal Statistical Society series C*, 62:151–165, 2013.

[30] S. Joshi, A. Lombardot, P. Flatresse, C. D'Agostino, A. Juge, E. Beigne, and S. Girard. Statistical estimation of dominant physical parameters for leakage variability in 32nanometer CMOS under supply voltage variations. *Journal of Low Power Electronics*, 8:113–124, 2012.

[31] P. Loiseau, P. Gonçalves, S. Girard, F. Forbes, and P. Primet Vicat-Blanc. Maximum likelihood estimation of the flow size distribution tail index from sampled packet data. In *SIGMETRICS–Joint International Conference on Measurement and Modeling of Computer Systems*, Seattle, USA, june 2009.