Contributions to high dimensional statistical learning

Stéphane Girard

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

Stephane.Girard@inria.fr

November 21, 2023

Abstract: This report summarizes my contributions to high dimensional learning. Four research topics are addressed: Unsupervised nonlinear dimension reduction, high dimensional classification, high dimensional regression and copulas construction.

Image analysis and computer vision are two important application domains for high dimensional data analysis and, more precisely, for dimension reduction methods. Indeed, a $M \times M$ grey-level image can be represented as a $p-$dimensional vector with $p = M^2$ or by a set of local descriptors. In both case, even with moderate image sizes, one obtains data living in very high-dimensional spaces. Principal Component Analysis (PCA) is usually an efficient tool for reducing the dimension of such data. However, even simple transformations between images can yield strong non-linearities in the p-dimensional space and thus strongly reduce the PCA efficiency.

To overcome this problem, we have introduced Auto-Associative models allowing to build new nonlinear dimension reduction methods. The dataset is approximated by a differentiable manifold generalizing PCA's linear subspaces [1, 2, 3, 4, 5, 6, 7]. The approximation algorithm is simple: it consists in incrementing the dimension of the manifold step by step. When the dataset is scattered into several groups, we have proposed a parametrization of the Gaussian mixture model. It is assumed that the high-dimensional data live in subspaces with intrinsic dimensions smaller than the dimension of the original space and that the data of different classes live in different subspaces with different intrinsic dimensions. New high-dimensional data classifiers are introduced on the basis of this model in both supervised and unsupervised contexts [8, 9, 10, 11, 12, 13]. The extension to non necessarily quantitative data is investigated in [14, 15, 16] and the application to the classification of grasslands from high resolution satellite image time series is addressed in [17, 18]. The challenging situation of irregularly sampled time series is tackled using multivariate Gaussian processes [19, 20]. We also refer to [21, 22, 23] for the estimation of the intrinsic dimension of a set points.

Another aspect of multivariate data analysis is the modeling of dependence between variables. The theory of copulas provides a relevant tool to build multivariate probability laws, from fixed marginal distributions and required degree of dependence. From Sklar's Theorem, the dependence properties of a continuous multivariate distribution can be entirely summarized, independently of its margins, by a copula. We have introduced a new semiparametric family of bivariate copulas. The family is generated by a univariate function, determining the symmetry (radial symmetry, joint symmetry) and dependence property (quadrant dependence, total positivity, ...) of the copulas [24, 25, 26]. An extension of this family is introduced in [27]. Inference is addressed in [28]. 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. In [29, 30, 31], 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 either by using a least-squares estimator based on dependence coefficients [32] or using Bayesian methods [33]. In [34], we propose a class of multivariate copulas based on products of transformed bivariate copulas. 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 [35] is to propose a nonparametric estimator of the conditional tail copula and to establish its asymptotic normality. Besides, the definition of new families of copulas thanks to a nonlinear mapping in investigated in [36].

Finally, I developped dimension reduction methods based on SIR for high dimensional regression problems [37, 38, 39, 40, 41, 42, 43, 44]. Two major issues are addressed: regularization for very high dimensional problems and sequential learning for very large datasets. See [45] for a review and [46, 47, 48, 49] for applications in astrophysics. Finally, a new version of the PLS method dedicated to conditional distribution tails has been proposed in [50].

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] S. Girard, B. Chalmond, and J-M. Dinten. Position of principal component analysis among auto-associative composite models. *Comptes-Rendus de l'Académie des Sciences, Série I*, 326:763–768, 1998.

[7] S. Girard, B. Chalmond, and J-M. Dinten. Une ACP non-linéaire basée sur l'approximation par variétés. *Revue de Statistique Appliquée*, XLVI(3):5–19, 1998.

[8] 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.

[9] 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.

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

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

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

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

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

[15] 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.

[16] S. Sylla, S. Girard, A. Diongue, A. Diallo, and C. Sokhna. A classification method for binary predictors combining similarity measures and mixture models. *Dependence Modeling*, 3:240–255, 2015.

[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] 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.

[19] A. Constantin, M. Fauvel, and S. Girard. Joint supervised classification and reconstruction of irregularly sampled satellite image times series. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022.

[20] A. Constantin, M. Fauvel, and S. Girard. Mixture of multivariate Gaussian processes for classification of irregularly sampled satellite image time-series. *Statistics and Computing*, 32:79, 2022.

[21] 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.

[22] O. Chelly, L. Amsaleg, T. Furon, S. Girard, M. Houle, K. Kawarabayashi, and M. Nett. Extreme-value-theoretic estimation of local intrinsic dimensionality. *Journal of Data Mining and Knowledge Discovery*, 32:1768–1805, 2018.

[23] L. Amsaleg, O. Chelly, T. Furron, S. Girard, M. Houle, and K. Kawarabayashi. Estimating local intrinsic dimensionality. In *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 29–38, Sydney, Australia, august 2015.

[24] C. Amblard and S. Girard. A new extension of bivariate FGM copulas. *Metrika*, 70:1–17, 2009.

[25] C. Amblard and S. Girard. Symmetry and dependence properties within a semiparametric family of bivariate copulas. *Nonparametric Statistics*, 14(6):715–727, 2002.

[26] C. Amblard and S. Girard. A semiparametric family of symmetric bivariate copulas. *Comptes-Rendus de l'Académie des Sciences, Série I*, 333:129–132, 2001.

[27] C. Amblard, S. Girard, and L. Menneteau. Bivariate copulas defined from matrices. `http://hal.archives-ouvertes.fr/hal-00875303`, 2013.

[28] C. Amblard and S. Girard. Estimation procedures for a semiparametric family of bivariate copulas. *Journal of Computational and Graphical Statistics*, 14(2):1–15, 2005.

[29] G. Mazo, S. Girard, and F. Forbes. A flexible and tractable class of one-factor copulas. *Statistics and Computing*, 26:965–979, 2016.

[30] 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.

[31] 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.

[32] 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.

[33] J. Arbel, M. Crispino, and S. Girard. Dependence properties and Bayesian inference for asymmetric multivariate copulas. *Journal of Multivariate Analysis*, 174:104530, 2019.

[34] 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.

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

[36] S. Girard. Transformation of a copula using the associated co-copula. *Dependence Modeling*, 6:298–308, 2018.

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

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

[39] R. Coudret, S. Girard, and J. Saracco. A new sliced inverse regression method for multivariate response. *Computational Statistics and Data Analysis*, 77:285–299, 2014.

[40] M. Chavent, S. Girard, V. Kuentz-Simonet, B. Liquet, T. M. N. Nguyen, and J. Saracco. A sliced inverse regression approach for data stream. *Computational Statistics*, 29:1129–1152, 2014.

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

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

[43] 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.

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

[45] S. Girard, H. Lorenzo, and J. Saracco. Advanced topics in sliced inverse regression. *Journal of Multivariate Analysis*, 188:104852, 2022.

[46] 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.

[47] S. Girard and J. Saracco. Supervised and unsupervised classification using mixture models. In D. Fraix-Burnet and S. Girard, editors, *Statistics for astrophysics - Clustering and classification*, volume 77, pages 69–90. EDP Sciences, 2016.

[48] S. Girard and J. Saracco. An introduction to dimension reduction in nonparametric kernel regression. In D. Fraix-Burnet and D. Valls-Gabaud, editors, *Regression methods for astrophysics*, volume 66, pages 167–196. EDP Sciences, 2014.

[49] 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.

[50] M. Bousebata, G. Enjolras, and S. Girard. Extreme Partial Least-Squares. *Journal of Multivariate Analysis*, 194:105101, 2023.