Irene Albarrán y Aurea Grané

“Selección de factores de riesgo en la dependencia y longevidad”

Albarrán, Alonso y Grané (2015) han desarrollado técnicas estadísticas basadas en distancias para bases de datos moderadamente grandes, con datos de tipo mixto y ponderado. Dichas técnicas se aplicaron para la determinación de los perfiles de dependencia en niños españoles que se compararon con lo que establece la legislación española vigente. En Albarrán, Alonso y Arribas-Gil (2016) se obtuvieron curvas de evolución de la dependencia para la población española y utilizando las técnicas de alineamiento de curvas para funciones de tipo escalón desarrolladas en Arribas-Gil y Müller (2014), se distinguieron comportamientos diferenciados según sexo y edad. Nuestro objetivo es estudiar el efecto de la discapacidad sobre la esperanza de vida libre de dependencia para proponer una metodología de tarificación del seguro de dependencia.

Referencias:

Albarrán, I., Alonso, P., Arribas-Gil, A. (2016) Dependence evolution in Spanish disabled population: A functional data analysis approach.**Journal of the Royal Statistical Society Series A- Statistics in Society**, vol. 180, 657–677. DOI: 10.1111/rssa.12228

Albarrán, I., Alonso, P. and Grané, A. (2015) Profile identification via weighted related metric scaling: An application to dependent Spanish children. **Journal of the Royal Statistical Society Series A- Statistics in Society**, vol. 178, 1—26. DOI: 10.1111/rssa.12084

Arribas-Gil, A., Müller, H.-G. (2014) Pairwise dynamic time warping for event data. **Computational Statistics and Data Analysis**, 69, 255-268.

Stefano Cabras y Mike Wiper

“Improving stochastic model search by analyzing algoritihm outcomes with non-parametric Bayesian methods”

Stochastic model search can be unfeasible when the number of predictor is large, say 1 million, as there are at least 2^(10^6) posible models. Even sophisticated model search algorithm can explore only a fraction of the possible models. Nonetheless, using already proposed model research strategies, it could be possible to estimate the portion of the model space with higher posterior probability. The idea is to use non-parametric Bayesian methods to analyze such outcomes to approximate the posterior probabilities on the full (and large) model space.

Carlos Ruiz Mora

“A predictive and prescriptive framework for the optimal integration of renewable generation in electricity systems”

In this work we will develop a novel data-driven forecasting and optimization framework to assist electricity producers, consumers and system operators in their decision making under uncertainty. The aim is to integrate the maximum amount of renewable resources in the electricity system while guaranteeing the energy supply and the revenue adequacy of all the agents. The forecasting model will combine ideas from Machine Learning (ML) and classical time series analysis techniques to obtain accurate predictions (point or probabilistic) of the uncertain parameters (renewables capacity, demand load, electricity price, etc.). These predictions will be adapted, in the form of a set of scenarios or uncertainty regions, to be the input of the prescriptive tool, which will be formulated either as a two-stage stochastic programing problem or as a robust optimization framework to derive optimal energy policies.

Juan Miguel Marin y Helena Veiga

"Bayesian estimation for generalized state space models"

State-space models (SSM), also known as hidden Markov models (HMM), are a very popular class of time series models that have been applied to very diverse fields as ecology, econometrics, engineering and environmental sciences. The popularity of state- space models is due to their flexibility and easy interpretability.

A noteworthy application of the SSM is stochastic volatility (SV), which has been suggested in the literature for modeling the time-varying volatility present in financial time series. Other applications are biochemical network models where the purpose is to model the dependence and relationships of various biochemical products when there are imprecise measurements of such products. Environmetrics is another area of application of these models that has been very successful; see Libonati et al. (2008).

The estimation of these models is not straightforward, given that the likelihood cannot be computed explicitly and so simulation based procedures, such as importance sampling (Geweke, 1989; Durbin and Koopman, 1997, 2001), MCMC (Jacquier et al., 1994), particle methods, also known as Sequential Monte Carlo (SMC) (Kantas et al., 2009; Creal, 2012) methods are commonly used to estimate model parameters.

In particular, MCMC methods are considered one of the most efficient estimation methods. However, its implementation can be very computationally demanding. This could be very hard for practitioners interested in fast answers and available computational programs. Therefore, the aim of this thesis is to propose estimation methods based on Bayesian inference like the INLA or ABC methodologies that are still efficient but that are not so computationally and time demanding.

Referencias:

Drew Creal (2012) A Survey of Sequential Monte Carlo Methods for Economics and Finance, *Econometric Reviews* 31(3), 245-296.

Durbin, J. and Koopman, S. J. (1997) Monte Carlo maximum likelihood estimation for non-Gaussian state space models, *Biometrika* 84 (3), 669–684.

Durbin, J. and Koopman, S. J. (2001) *Time Series Analysis by State Space Methods*, Oxford University Press, Oxford

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometric, 57*(6), 1317-1339.

Jacquier, E., Polson, N.G. and Rossi, P.E. (1994). Bayesian analysis of stochastic volatility models, *Journal of Business Economics and Statistics* 12, 371-389.

Kantas, N and Doucet, A and Singh, SS and Maciejowski, JM (2009) *An overview of sequential Monte Carlo methods for parameter estimation in general state-space models.* In: 15th IFAC Symposium on System Identification, SYSID 2009, 2009-7-6 to 2009-7-8, Saint-Malo, France.

Libonati, R., Trigo, I. and Camara, C. (2008) Correction of 2 m-temperature forecasts using Kalman Filtering technique, Atmospheric Research 87, 183–197.

Helena Veiga

"New estimators of realized variance and applications to financial predictability"

A recurrent question in the financial literature is whether stock returns are predictable. There is no clear evidence of this predictability and it often depends on the predictors used; see Wilcox [2007], Lettau and Van Nieuwerburgh [2008], Bollerslev et al. [2009, 2011, 2012, 2014], Drechsler and Yaron [2011], Galaix [2012], Bekaer and Hoerova [2014] and Kelly and Jiang [2014].

One well-known potential predictor of the stock returns is the variance risk premium that is the different between the VIX (volatility index) squared and the conditional variance often measured by the realized variance. Since the availability of high-frequency data, realized variance (RV) has been one of the main focus of research to account for uncertainty in financial investments. Amongst many other applications, the realized variance is considered a proxy of economic uncertainty [see Bekaer and Hoerova, 2014] and it has a primordial role in the estimation and forecast of variance risk premium which, at the same time, is a measure of risk aversion to uncertainty. Therefore, obtain good measures of realized variance is crucial for research.

Barndorf-Nielsen et al. [2010] propose the realized semivariance as the first asymmetric measure of realized variance, which decomposes the realized variance into a positive component and a negative component corresponding to positive and negative high-frequency returns, respectively. However, this measure of asymmetry does not capture how unsettling positive or negative returns are likely to be for volatility and current investment decisions.

To overcome this caveat, the idea is to propose new asymmetric measures of the realized variance based on the net estimators in Hamilton [1996] and Ramos and Veiga [2011] and on the scale estimators in Park and Ratti [2008]. The proposal is to substitute the daily positive and negative realized variance terms in the heterogeneous autoregressive (HAR) model of Corsi et al. [2012] resulting in new asymmetric HAR-type specifications which are evaluated in terms of stock returns, economic and financial instability predictability. Furthermore, the proposed specifications are extended to allow for flexible time-varying coefficients which are estimated nonparametrically.

References

O. E. Barndorff-Nielsen, S. Kinnebrock, and N. Shephard. Measuring downside risk? Realised semivariance. In T. Bollerslev, J. Russell, and M. Watson, editors, Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle. Oxford University Press, New York, 2010

G. Bekaer and M. Hoerova. The VIX, the variance premium and stock market volatility. Journal of Econometrics, 183(2):181–192, 2014.

T. Bollerslev, G. Tauchen, and H. Zhou. Expected stock returns and variance risk premia. The Review of Financial Studies, 22(11):44–63, 2009.

T. Bollerslev, M. Gibson, and H. Zhou. Dynamic estimation of volatility risk premia

and investor risk aversion from option-implied and realized volatilities. Journal of Econometrics, 160(1): 235–245, 2011.

T. Bollerslev, N. Sizova, and G. Tauchen. Volatility in equilibrium: Asymmetries and dynamic dependencies. Review of Finance, 16(1):31–80, 2012.

T. Bollerslev, J. Marrone, L. Xu, and H. Zhou. Stock return predictability and variance risk premia: Statistical inference and international evidence. Journal of Financial and Quantitative Analysis, 49(3):633–661, 2014.

F. Corsi, F. Audrino, and R. Reno. HAR modeling for realized volatility forecasting. In Handbook of Volatility Models and Their Applications, pages 363–382. John Wiley & Sons, Inc., New Jersey, USA, 2012.

I. Drechsler and A. Yaron. What's vol got to do with it. The Review of Financial Studies, 24 (1):1–45, 2011.

X. Galaix. Variable rare disasters: An exactly solved framework for ten puzzles in macro- finance. Quartely Journal of Economics, 127:645–700, 2012.

J. D. Hamilton. This is what happened to the oil price-macroeconomy relationship. Journal of Monetary Economics, 38(2):215–220, 1996.

B. Kelly and H. Jiang. Tail risk and asset prices. The Review of Financial Studies, 27(10): 2841–2871, 2014.

M. Lettau and S. Van Nieuwerburgh. Reconciling the return predictability evidence. The Review of Financial Studies, 21(4):1607–1652, 2008.

J. Park, R. A. Ratti. Oil price shocks and stock markets in the U.S. and 13 European countries. Energy Economics, 30(5): 2587-2608.

S. B. Ramos and H. Veiga. Risk factors in oil and gas industry returns: International evidence. Energy Economics, 33(3):525–542, 2011.

S. E. Wilcox. The adjusted earnings yield. Financial Analysts Journal, 63:54–68, 2007

Helena Veiga and Marc Vorsatz

"Experiments with applications to financial markets"

The 2008 crisis has been visible on many dimensions: some financial institutions collapsed, housing markets eroded, and several European countries faced a sovereign debt crisis after bailing out their banks. The consequences of the financial crisis and the implied economic downturn were, among others, higher unemployment rates, more inequality, and a reduction of the welfare state in general (less expenditure in education, health care, and pensions). It took several countries a decade to regain their pre-crisis economic activity.

Traditional theoretical models have a hard time to explain the 2008 financial crisis due to the limitation of the assumptions underlying individual behavior. Instead, laboratory experiments can be design to account for the fact that not all individuals are rational. The reality is that only very few studies use experiments to analyze the outcome of financial markets. One reason why the experimental method has not been widely adopted in finance could be the general belief that the participants in real-life markets are highly paid professionals who face a very tough competition and that the competition in the market eliminates the worse traders and thereby price inefficiencies. Yet, if this argument was really true and economic agents were indeed highly rational, then it would not have been possible to sustain a serious mispricing of assets over a very long period of time. This apparent contradiction shows the need for a behavioral approach, in order to better understand the functioning of financial markets. Note that there are variables that are not empirically observable such as private information, fundamental values of assets, and beliefs, and they can be controlled for or extracted in the laboratory.

The aim of this thesis is to analyze financial markets via laboratory experiments. One important question is the financial contagion across markets. We speak of financial contagion if the correlation of prices across markets is too high. The theoretical literature on financial contagion has identified several channels that are not caused by the linkage of financial institutions across countries. In King and Wadhwani (1990), informational spillovers give traders incentives to observe the price in one market and trade in the other market on the basis of this information. Calvo (1999) examines liquidity shocks; that is, a trader who needs to close a position in one market may liquidate her position in other markets as well in order to meet the liquidity requirements. In Kodres and Prisker (2002), contagion happens because a shock in one market causes the traders to rebalance their portfolio across markets, a hypothesis that has been tested successfully in the laboratory by Cipriani et al. (2013). Finally, in Fostel and Geanakoplos (2008), market

incompleteness and heterogeneity of preferences cause prices of a priori uncorrelated assets to be correlated. Since it has been shown experimentally by Veiga and Vorsatz (2009, 2010) that an uninformed manipulator is more likely to successfully distort the process of information transmission the more incomplete markets are, we conjecture that financial contagion is also linked to this dimension (as suggested by Fostel and Geanakoplos 2008). To test this hypothesis, we are going to implement several experiments in laboratory and analyze the data.

Methodology: The student has to develop first the experimental design. This phase is crucial for the success and has to be carefully executed. In particular, the student has to develop the simplest possible environment where the research question can be tested. Also, the student has to define different treatment conditions that isolate the effects to be studied. Since the proposed experiments are all conducted via computer terminals, programs have to be written. Thanks to Z-Tree, a free computer software toolbox for running laboratory experiments, the effort can be realized in a reasonable amount of time. Once the data is obtained, it has to be analyzed using statistical inference.

References:

Calvo, G. (1999). Contagion in emerging markets: when Wall Street is a carrier. WP University of Maryland.

Cipriani, M., G. Gardenal, and A. Guarino (2013). Financial contagion in the laboratory: The crossmarket rebalancing channel. Journal of Banking and Finance 37: 4310-4326.

Fostel, A. and J. Geanakoplos (2008). Leverage Cycles and the Anxious Economy. American Economic Review 98: 1211-1244.

King, M. and S. Wadwani (1990). Transmission of volatility between stock markets. Review of Financial Studies 3: 5-33.

Kodres, L. and M. Pritsker (2002). A rational expectations model of financial contagion. Journal of Finance 57: 769-799

Veiga H. and M. Vorsatz (2009). Price manipulation in an experimental asset market. European Economic Review 53: 327-342.

Veiga H. and M. Vorsatz (2010). Information aggregation in experimental asset markets in the presence of a manipulator. Experimental Economics 13: 379-398.

Isabel Molina y Juan Miguel Marin

“Bayesian poverty mapping in small áreas”

For estimation of non-linear parameters in small areas such as poverty and/or inequality indicators (poverty mapping), a common approach is to model the incomes of individuals through unit level models, see e.g. Molina and Rao (2010) for a frequentist approach or Molina, Nandram and Rao (2014) for a hierarchical Bayes analogue. These models assume normality of the log-incomes, but even after log transformation, the distribution of income is not fitting well a normal distribution. Graf, Marín and Molina (2018) proposed a unit level model for small area estimation based on the GB2 distribution that fits income much better, and obtained empirical best (EB) estimators of small area quantities, including a family of poverty indicators. They used a parametric bootstrap procedure for mean squared error estimation and a Monte Carlo procedure for approximation of the EB estimators of complex target parameters, which can be computationally very intensive. But taking advantage of the mixture representation of the GB2 distribution, the Bayesian approach can be also applied. Thus, we propose to consider a Bayesian version of the mentioned unit level model based on the GB2 distribution, which will avoid the use of the bootstrap and therefore will be much faster than the frequentist counterpart, while providing similar estimates through the use of non-informative (or very vaguely informative) priors.

References:

- Graf, M., Marín, J.M. and Molina, I. (2018). A generalized mixed model for skewed distributions applied to small area estimation, Test, https://doi.org/10.1007/s11749-018-0594-2
- Molina, I. and Rao, J.N.K. (2010). Small area estimation of poverty indicators, Canadian Journal of Statistics 38:369–385.
- Molina, I., Nandram, B. and Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach, The Annals of Applied Statistics:8, 852–885.

Daniel Peña and Andrés M. Alonso

For an extended versión, please contact the proposing professors

"On time series clustering"

Most procedures for clustering time series look at the similarity of the elements of a set of times series and build measures of distance using the univariate features of the series. Some recent reviews are Liao (2005) and Aghabozorgi et al (2015). These methods are useful when we have independent time series and the objective is to cluster them by the similarity among their univariate models, in a parametric framework, or by its periodogram or autocorrelations, in a nonparametric framework. However, in many applications, the independence assumption is not fulfilled. The main goal of this project is to develop procedures that take into account the cross-dependence when clustering time series. The developed procedures will be use to study meteorological, environmental as well as financial and economical time series.

Some reviews on this topic:

- Aghabozorgi, S., Shirkhorshidi, A.S. and Wah, T.Y. (2015) Time-series clustering – A decade review, *Information Systems*, 53, 16–38.
- Liao, T.W. (2005) Clustering of time series data-a survey, *Pattern Recognition*, 38, 1857–1874.

A (biased) list of references on this topic:

1. Alonso, A.M., Berrendero, J.R., Hernández, A. and Justel, A. (2006) Time series clustering based on forecast densities, *Computational Statistics and Data Analysis*, 51, 762–766.
2. Alonso, A.M., Galeano, P. and Peña, D. (2017) Outlier detection and robust estimation in dynamic factor models with cluster structure, *Preprint*, 25pp.
3. Alonso, A.M. and Maharaj, E.A. (2006) Comparison of time series using subsampling, *Computational Statistics and Data Analysis*, 50, 2589–2599.
4. Alonso, A.M. and Peña, D. (2017) Clustering time series by dependency, *Preprint*, 34pp.
5. Barbosa, S., Gouveia, S., Scotto, M. and Alonso, A.M. (2016) Wavelet-based clustering of sea level records, *Mathematical Geosciences*, 48, 149–162.
6. Caiado, J., Crato, N. and Peña, D. (2006) A periodogram-based metric for time series classification, *Computational Statistics and Data Analysis*, 50, 2668–2684.
7. D'Urso, P., Maharaj, E.A. and Alonso, A.M. (2017) Fuzzy clustering of time series using extremes, *Fuzzy Sets and Systems*, 318, 56–79.
8. Galeano, P. and Peña, D. (2000) Multivariate Analysis in vector Time Series, *Resenhas*, 4, 383–404.
9. Peña, D. and Prieto, F.J. (2001) Cluster identification using projections, *Journal*

of American Statistical Association, 96, 1433–1445.

10. Scotto, M., Barbosa, S. and Alonso, A.M. (2011) Extreme value and cluster analysis of European daily temperature series, *Journal of Applied Statistics*, 38, 2793–2804.

11. Maharaj, E.A., Alonso, A.M. and DUrso, P. (2015) Clustering seasonal time series using extreme value analysis: An application to Spanish temperature time series, *Communications in Statistics Case Studies and Data Analysis*, 1, 175-191.

12. Vilar-Fernández, J.A., Alonso, A.M. and Vilar-Fernández, J.M. (2010) Nonlinear time series clustering based on nonparametric forecast densities, *Computational Statistics and Data Analysis*, 54, 2850–2865.

Rosa Lillo y Pedro Galeano

"Applications of the functional Mahalanobis semi-distance"

Functional data consist of observed functions or curves evaluated at a finite interval of the real line. In a conceptual sense, functional data are intrinsically infinite dimensional and thus, classical methods designed for multivariate observations are no longer applicable. Consequently, there is a need to develop special techniques for this type of data. The books by Ramsay and Silverman (2005) and Ferraty and Vieu (2006) offer comprehensive introductions to FDA and its applications, while Horváth and Kokoszka (2012), Hsing and Eubank (2015) and Wang, Chiou and Müller (2016) review some recent developments for functional data.

It is well-known that distances play a major role in many statistical techniques as supervised classification, clustering or prediction problems. Reviewing the notion of distances and semi-distances for functional data, it is easy to see that the Mahalanobis distance proposed by Mahalanobis (1936) had not yet been extended to the functional framework until Galeano et al. (2015), who generalize the definition of the multivariate Mahalanobis distance to the functional case. In this way, it can be used to solve functional statistical problems that may require the use of distances. Galeano et al. (2014) focus on supervised classification and Joseph et al. (2017) considers the hypothesis testing problem.

The objective of the thesis proposal is to continue researching the usefulness of the functional Mahalanobis semi-distance in other statistical problems based on distances. In particular, it would interesting to face the problem of cluster analysis in the functional framework or to focus on the problem of outlier detection in functional data.

References:

- Galeano P., Joseph, E, Lillo, R. E. (2015). The Mahalanobis distance for functional data with applications to classification. *Technometric*, 57, 281-291.
- Ferraty, F. and Vieu, P. (2006). Nonparametric functional data analysis. New York: Springer.
- Horváth, L. and Kokoszka, P. (2012).Inference for functional data with applications. New York: Springer.
- Hsing, T. and Eubank, R. (2015). Theoretical foundations of functional data analysis, with an introduction to linear operators. Wiley Series in Probability and Statistics, Wiley & Sons,
- Joseph, E., Galeano, P. y Lillo, R. E. (2015). The functional two-sample Hotelling's T2 statistic. Submitted.
- Ramsay, J. O. and Silverman, B. W. (2005). Functional data analysis, Second edition. New York: Springer.
- Wang, J. L., Chiou, J. M. and Müller, H. G. (2016). Functional data analysis. Annual Review of Statistics and its Applications, 3, 257-295.

Rosa Lillo

"Multivariate Directional quantiles in Finance and/or Hydrology"

The aim of the work is to continue with open problems related to the thesis "The multivariate directional approach: high level quantiles estimation and applications to finance and environmental phenomena" defended in December 2016 by Raúl Torres in the Statistics Department. The basic idea of the thesis is to introduce a directional multivariate approach to analyze extremes that allow us to introduce manager preferences or external information available for the system of interest. The key definition in which this thesis is based on, is the notion of directional multivariate quantiles.

In Torres et al (2016), we introduce a directional multivariate risk measure which is a multivariate extension of the well-known univariate risk measure Value at Risk (VaR). Properties for the proposed multivariate risk measure are provided as extensions of the axiomatic for univariate risk measures given in the literature. This paper also highlights the importance of using directions thanks to a result providing a conservative bound (upper bound) of the total risk in a portfolio investment by using the direction of the weights of investment to analyze such loss.

In Torres et al. (2017a), we focus on the formal definition and estimation of the directional multivariate extremes. Given that environmental science possesses different phenomena where joint behavior of variables may cause disasters, two real cases of study are analyzed. In the literature, it is possible to find copula theory to model those dependencies, which leads us to introduce the directional approach to the copula framework. Thus, advantages and disadvantages between non-parametric approaches and theoretical copula approaches are highlighted in this work. Moreover, it is presented a proposal to choose a suitable direction of analysis by considering the direction of the maximum variability on the data, which links the use of Principal Component Analysis (PCA). Applications are performed on the real cases of study of flood risk at a dam (3 dimensional case) and sea storms (5 dimensional case).

In extreme value theory, it is known that standard non-parametric methods cannot be applied to estimate quantiles at high levels. Therefore, a different approach known as out-sample estimation must be considered. In this sense, Torres et al. (2017b) introduce the necessary background to face the multivariate extreme value theory. Then, results including the directional approach to the multivariate extreme value theory are given. An estimator of the directional multivariate quantiles is provided and its asymptotic normality is also proved. Finally, it is presented a nonparametric methodology based on bootstrap to accomplish the goal of estimation, with illustrations in both simulated and real data.

However, there exist several research lines related to the previous works that can be explored for a Ph.D student: (1) we have a relationship with researchers of the Environmental Hydraulics Institute in Cantabria that provided us interesting multivariate data in order to be explored from a directional perspective. (2) We have collaboration with Gloria González-Rivera (University of California) to extend the Value in Stress that she introduced in the literature using a dynamic and directional approach. In summary, research in both environmental and financial areas can be developed.

References:

- Torres R., Lillo R.E., and Laniado H. (2015). A directional multivariate Value at Risk. *Insurance: Mathematics and Economics*, 65, 111-123.
- Torres R., De Michele C, Laniado H., Lillo, R.E. (2017a). Directional Multivariate Extremes in Enviromental Phenomena. *Environmetrics*, Vol. 28 (2), 1-15.
- Torres R, Di Bernardino E, Laniado H; Lillo, Rosa E. (2017b). A bootstrap-based method to estimate directional extreme risk regions at high levels. Submitted.

Helena Veiga, Michael Wiper y Juan de Dios Tena

"New approaches to competitivensess and efficiency modeling in football!

Competitiveness in sport has various different aspects. Firstly, from the viewpoint of an individual club, it is important to constantly maintain high league rankings in order to gain more gate and television money, access to lucrative European competitions and also to sign high quality players who wish to play in successful teams. Also, from the point of view of a given league, it is important for the league to be both internally and externally competitive so that in the first case, fans do not get bored watching the same teams always winning and, in the second case, leagues with clubs that do well in international competitions attract more funding. Therefore, we propose to explore new measures of competitiveness in this sector following e.g. Corona et al. (2017) and Santamaria (2016).

It is also important for football clubs to be efficient. Efficiency in this context has a number of different aspects: we can think of revenue efficiency or cost efficiency or even managerial efficiency, but the main objectives are to do as well as possible given the resources available to a club. In this context, it is possible to apply stochastic frontier models to compare the efficiencies of different players (Tiedemann et al. 2011), teams (Barros et al 2015) or even to compare different leagues which is one of the objectives of this thesis.

Bibliography & Recommended Reading

- Barros, C.P. Wänke, P. and Figueiredo, O. (2015). The Brazilian Soccer Championship: an efficiency analysis. *Applied Economics*, **47**, 906–915,
- Corona, F., de Dios Tena, J., Forrest, D. and Wiper, M.P. (2017). Evaluating significant effects from alternative seeding systems: a Bayesian approach, with an application to the UEFA Champions League. *UC3M Working Papers. Statistics and Econometrics* **17-03**.
- Corona, F., de Diós Tena, J. and Wiper, M. P. (2017). On the Importance of the Probabilistic Model in Identifying the Most Decisive Games in a Tournament *Journal of Quantitative Analysis of Sports*, **13**, 11-23.
- Jara, M., Paolini, D. and de Dios Tena Horrillo, J. (2015). Management Efficiency in Football: An Empirical Analysis of Two Extreme Cases. *Managerial and Decision Economics*, **36**, 286-298.
- Santamaria, R. (2016). Competitividad en las ligas de fútbol europeas. Trabajo Fin de Grado dirigido por M.P. Wiper, Grado en Estadística y Empresa, Universidad Carlos III de Madrid.
- Tiedemann, T., Franckson, T. and Latacz-Lohmann, U. (2011). Assessing the performance of German Bundesliga football players: a non-parametric metafrontier approach. *Central European Journal of Operations Research*, **19**, 571–587.

Michael Wiper, Helena Veiga y Sofía Ramos

"Advances in Bayesian stochastic frontier analysis and applications to finance"

Stochastic frontier models, originally proposed by Aigner et al. (1977) and Meeusen and van den Broeck (1977), are crucial in efficiency measurement and were the starting point of an active research area with relevant theoretical developments and applications of interest for managers and policy makers. Examples may be found in many fields of study including insurance sector, economics, banking and finance. In particular, policy makers are interested in regional competitiveness evaluation given that the economic efficiency of regions represents the basis of a countries economic success.

Another application of interest is cost efficiency estimation of the banking sector since efficiency measurement and relative efficiency comparison of banks are crucial questions for analysts as well as for economic policy creators. After the financial crisis of 2008 the costs associated with the bank bailouts increased the policy debate concerning the role and benefits of bank size and the influence of public safety net subsidies.

The aim of this thesis is to propose new methodology in the context of Bayesian stochastic frontier analysis and its application to real and actual finance cases of crucial interest in both cross-sectional and panel data scenarios

Bibliography & Recommended Reading

Aigner, D.,Lovell, C. A. K. and Schmidt, P. (1977). Formulation and estimation of stochastic frontier production function models, ***Journal of Econometrics***, 6, 21-37.

Gálan, J., Ramos, S. and Veiga, H. (2017). An analysis of the dynamics of efficiency of mutual funds, Universidad Carlos III de Madrid, WP 15-17.

Gálan, J., Veiga, H. and Wiper, M. P. (2014). Bayesian estimation of inefficiency heterogeneity in stochastic frontier, ***Journal of Productivity Analysis***, 42 (1), 85-101.

Gálan, J., Veiga, H. and Wiper, M. P. (2015). Dynamic Effects in Inefficiency: Evidence from the Colombian Banking Sector, ***European Journal of Operational Research***, 20(2), 562-571.

Meeusen, W. and Van Den Broeck, J. (1977). Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error, ***International Economic Review***, 18(2), 435-444.

Michael Wiper y Eduardo García Portugués

"Flexible modelling of high-dimensional directional data"

The term *directional data* refers to data whose support is a sphere of arbitrary dimension q. Most common cases are $q = 1$ (circular data) and $q = 2$ (spherical data). Circular variables appear naturally in several applied disciplines: proteomics (angles in the structure of proteins); environmental sciences (wind direction, direction of waves); biology (animal orientation); cyclic phenomena (arrival times at a care unit, seasonality in freezing and thawing). Spherical data is present in astronomy (positions of stars, satellite tracking), paleontology (bedrock orientations), and forest science (orientations of wildfires). High-dimensional directional data plays a relevant role in *text mining* (e.g., see Banerjee *et al.* (2005) or García-Portugués *et al.* (2016)). Remarkably, in the last years, *toroidal data* (*i.e.*, multivariate circular data) has played a key role in the development of probabilistic models for protein structure prediction (Boomsma *et al.* 2008; Hamelryck *et al.* 2012; Golden *et al.*, 2017).

The collection of statistical techniques intended to analyze directional data is referred as **directional statistics** (see Mardia and Jupp (2000) or Jammalamadaka and SenGupta (2001)) and presents numerous differences with respect to classical statistics. For example, the circular mean differs substantially from the usual mean. Most of the statistical methods to analyze directional data are parametric. They have excellent properties whenever the assumptions on which they are based hold, which unfortunately is rarely the case in real-life applications. As a consequence, flexible methods that do not rely on hard-to-satisfy assumptions are of primary interest for practitioners, with applications ranging from mere data visualization to the goodness-of-fit of parametric models.
The objective of this PhD project is to **develop and implement new methodology for flexible modelling of directional data**. Specifically, the project aims to provide contributions along three research lines in the context of directional data:

1. **Kernel Density Estimation** (KDE; see Wand and Jones (1995) for an introduction). KDE with directional data (see Hall *et al.*, (1987) and García-Portugués (2013) for an overview) is underdeveloped when compared to the Euclidean setting, and is also more challenging due to the nonlinearity of the support. In particular, there is a lack of sophisticated bandwidth selectors for KDE with toroidal data. The first research direction is to develop and study a bandwidth selector similar to the ones proposed in García-Portugués (2013) for KDE with directional data. A second research direction involves the development of a software package implementing a fast and efficient KDE for high-dimensional toroidal data.

2. **Flexible parametric modelling**. The mixtures of directional distributions, such as von Mises (Banerjee *et al.*, 2005), have been a popular modelling tool among practitioners due to their flexibility. An application of mixtures of distributions is the approximation of the Kullback-Leibler divergence between two unknown densities, which is an untractable quantity with important statistical applications (see Hershey and Olsen (2007)). This research line will seek to find directional distributions that yield analytically tractable Kullback-Leibler approximations.

3. **Visualization of high-dimensional data**. High-dimensional multivariate data is routinely present in massive molecular dynamic simulations of proteins. As a consequence, there is an urge for dimension reduction methods that are able to provide insightful visualization

of the main protein conformations hidden the simulation outputs. This research line aims to explore and implement convenient adaptations of the popular t-SNE of Maaten and Hinton (2008) for analyzing large volumes of toroidal data.

For more information on the project, contact one of the advisors.

References

- Banerjee, A., Dhillon, I. S., Ghosh, J. and Sra, S. (2005) Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6, 1345-1382.
- Boomsma, W., Mardia, K. V., Taylor, C. C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008) A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105, 8932-8937.
- García-Portugués, E. (2013) Exact risk improvement of bandwidth selectors for kernel density estimation with directional data. *Electronic Journal of Statistics*, 7, 1655-1685.
- García-Portugués, E., Van Keilegom, I., Crujeiras, R. and González-Manteiga, W. (2016) Testing parametric models in linear-directional regression. *Scandinavian Journal of Statistics*, 43(4), 1178-1191.
- Golden, M., García-Portugués, E., Sørensen, M., Mardia, K., Hamelryck, T. and Hein, J. (2017) A generative angular model of protein structure evolution. *Molecular Biology Evolution*, in press.
- Hall, P. and Watson, G. S. and Cabrera, J. (1987) Kernel density estimation with spherical data. *Biometrika*, 74(4), 751-762.
- Hamelryck, T., Mardia, K. V. and Ferkinghoff-Borg, J. (Eds.) (2012) *Bayesian Methods in Structural Bioinformatics*. Springer.
- Hershey, J. R. and Olsen, P. A. (2007) Approximating the Kullback Leibler divergence between Gaussian mixture models. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 4, 317-320.
- Jammalamadaka, S. R. and SenGupta, A. (2001) *Topics in Circular Statistics*. World Scientific Publishing.
- van der Maaten, L. J. P. and Hinton, G. E. (2008) Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9(Nov), 2579-2605.
- Mardia, K. V. and Jupp, P. E. (2000) *Directional Statistics*. John Wiley & Sons.
- Wand, M. P. and Jones, M. C. (1995) *Kernel smoothing*. Chapman & Hall.