

IWSM 2011

Proceedings of the

26th International Workshop on Statistical Modelling

Valencia (Spain), July 11-15, 2011



Editors:

David Conesa

Anabel Forte

Antonio López-Quílez

Facundo Muñoz

Proceedings of the 26th International Workshop on Statistical Modelling.
València, July 11-15, 2011
David Conesa, Anabel Forte, Antonio López-Quílez, Facundo Muñoz, eds.
València 2011.
ISBN 978-84-694-5129-8
Depósito Legal V-2498-2011

Editors:

David Conesa¹, David.V.Conesa@uv.es
Anabel Forte², forte@eco.uji.es
Antonio López-Quílez¹, Antonio.Lopez@uv.es
Facundo Muñoz¹, Facundo.Munoz@uv.es

¹ Departament d'Estadística i Investigació Operativa
Universitat de València (Estudi General)
Facultat de Matemàtiques

Dr. Moliner 50, 46100 Burjassot, Spain.

² Departamento de Economía

Universitat Jaume I

Facultad de Ciencias Jurídicas y Económicas
Campus del Riu Sec, E-12071 Castelló de la Plana, Spain.

Cover photo: Victor Roda

Printed by Copiformes S.L.

Contents

Part 1. Invited papers

Berger et al. Risk Assessment for Pyroclastic Flows: Combining Deterministic and Statistical Modeling	3
Firth Quasi-variances and extensions	10
Gómez Some theoretical thoughts when using a composite endpoint to prove the efficacy of a treatment	14
Green et al. Identifying influential model choices in Bayesian hierarchical models	22
Jørgensen et al. The Ecological Footprint of Taylor's Universal Power Law	27

Part 2. Contributed papers

Aerts et al. Incomplete Clustered Data and Non-Ignorable Cluster Size	35
Alvaro-Meca et al. Bayesian Lee-Carter Model: A Spatio-Temporal Approach.	41
Andrés-Ferrer and Ney From Empirical Bayes to Leaving-One-Out	45
Aregay et al. Model Based Estimates of Long-Term Persistence of Induced HPV Antibodies: A Flexible Subject-Specific Approach	49
Armero et al. Bayesian model selection for assessing the progression of chronic kidney disease in transplanted children.	53
Badiella et al. Area under the ROC curve using logistic regression with random effects: Estimation and Inference	57
Barber et al. Optical properties of fresh date palm in different stages of maturity	63
Bárceña et al. Measuring the real estate bubble: a house price index for Bilbao.	67

Baxter et al. Missing data, multiple imputation and the UK National Vascular Database	71
Belgrave et al. A Comparison of Frequentist and Bayesian Approaches to Latent Class Modelling of Susceptibility to Asthma and Patterns of Antibiotic Prescriptions in Early Life	75
Boixadera et al. Who uses Complementary and Alternative Medicine? An analysis for cancer patients	79
Bowman and Crujeiras Assessing isotropy with the variogram	83
Brechmann et al. Simplified regular vines for modeling high-dimensional financial risk data	87
Brewer et al. Climate Envelopes for Species Distribution Models	93
Burke and MacKenzie XD survival regression models with frailty	99
Caballero-Águila et al. Least-squares signal estimation using correlated delayed observations transmitted by different sensors	105
Caballero-Águila et al. Filtering algorithm for fractional order discrete systems with uncertain observations	109
Carrasco et al. The Log-Generalized Modified Weibull Regression Model	113
Castillo and Serra An exponential dispersion family to modelling critical phenomenon	117
Catelan and Biggeri Hierarchical Bayesian modelling to assess divergence in disease mapping	121
Conde and MacKenzie LASSO Penalised Likelihood in High-Dimensional Contingency Tables	127
Conesa et al. Describing the geography of Spanish bank branching.	133
Corberán-Vallet and Lawson Spatio-temporal disease modeling and surveillance with Bayesian hierarchical Poisson models ...	137
Corberán-Vallet et al. Time series modeling and Bayesian forecasting with exponential smoothing models	141
Costa and Dias Assessment of e-government maturity in Portuguese municipalities using regression and clustering approaches	146

Creemers et al. <i>Joint Modeling Longitudinal Health Care Costs and Time-to-Event Data in Matched Pairs</i>	150
Cysneiros <i>Bartlett-type Correction in Heteroscedastic Symmetric Nonlinear Models</i>	156
Cysneiros et al. <i>A Symbolic Robust Regression Model</i>	160
Czado et al. <i>Bayesian inference for copula based GARCH models</i>	164
Dejardin et al. <i>Bayesian Dose Escalation in phase I studies of Combinations of Drugs with Control</i>	169
De Rooi and Eilers <i>Using text mining tools to compose structure priors for inferring gene networks</i>	173
Djennad et al. <i>Markov-Switching Multifractal models within GAMLSS</i>	178
Djeundje and Currie <i>Smooth mixed models for nested curves</i> .	183
Dondelinger et al. <i>A Bayesian regression and multiple changepoint model for systems biology</i>	189
Dooley et al. <i>Analysis of an Observational Study</i>	195
Eilers et al. <i>Sea Level Trend Estimation by Seemingly Unrelated Penalized Regressions</i>	200
Fabio et al. <i>Generalized random intercept log-gamma exponential family models</i>	206
Faria and Gonçalves <i>Modelling Financial Data using Poisson Mixture Approach</i>	210
Finazzi et al. <i>A multivariate space-time model for heterogeneous air quality networks</i>	214
Fonseca et al. <i>Predictive distributions for non-regular parametric models</i>	220
Forte et al. <i>Objective Bayes Criteria for Variable Selection</i>	224
Franco-Villoria et al. <i>Conditional Probability of Flood Risk in Scotland</i>	228
Fried et al. <i>Outliers and interventions in INGARCH time series</i>	234
Furche et al. <i>Bivariate Ordinal Regression Models for the Analysis of Neural Data</i>	240

Gallego et al. <i>Modelling endocytosis by means of non-homogeneous temporal Boolean models</i>	244
García-Donato et al. <i>A Prior for multiplicity control and closed-form Bayes factors in variable selection</i>	248
García-Mora et al. <i>Approximated Survival function in the Sum of Two Independent Homogeneous Markov Processes: Application to Bladder Carcinoma</i>	249
Gargoum <i>On using the Hellinger distance in checking the validity of approximations based on dynamic generalized linear models</i>	253
George and Ünlü <i>Parameter Estimation in Skills-based Knowledge Space Theory and Cognitive Diagnosis Models: A Comparison</i>	258
Gilchrist et al. <i>Forecasting film revenues using GAMLSS</i>	263
Gilthorpe et al. <i>Importance of correctly specifying the random structure in growth mixture models</i>	269
Gomes et al. <i>Modeling swimming marks through Blocks and POT methods</i>	273
Gonçalves and Costa <i>Improvement of surface water quality variables modelling that incorporates a hydro-meteorological factor: a state-space approach</i>	276
Gottard et al. <i>Modelling fertility and education in Italy in the presence of time-varying frailty component</i>	281
Grisotto et al. <i>Empirical Bayes models to estimate contextual effects</i>	287
Habteab Ghebretinsae et al. <i>Generalized Frailty Model for Comet Assays</i>	292
Ha et al. <i>Interval Estimation of Random Effects in Frailty Models</i>	298
Haggarty et al. <i>Functional Clustering of Water Quality Data in Scotland</i>	303
Hasso and Matawie <i>Using Probability Models to Classify Software Patterns</i>	308
Hernandez et al. <i>Linear Model comparison with structured mean and dispersion parameters</i>	312
Huertas et al. <i>Joint Modelling of Two Sequential Times to Events With Longitudinal Information</i>	316

Ibacache Pulgar and Paula Elliptical semiparametric mixed models	322	Moreira and Machado An R Package for the Estimation of the Bivariate Distribution for Censored Gap Times	410
Kelly The change-point problem in regression with correlated data and change in variance	326	Muggeo and Lovison Testing for a breakpoint in segmented regression: a pseudo-score approach	415
Komárek Capabilities of R package <code>mixAK</code> for clustering based on multivariate continuous and discrete longitudinal data	330	Muñoz and López-Qúlez Geostatistical modelling with non-Euclidean distances	419
Lambert Additive location-scale model when the response and some covariates are interval censored	334	Murawska et al. Multi-state models for non Markov process ..	423
Letón and Molanes-López Second order delta method for estimating the Youden index and optimal threshold	338	Mutsvari et al. Some approaches to correct for misclassification in the absence of an internal validation data set	427
Little et al. Modeling growth patterns of the swift tern using nonlinear mixed effect models	342	Nicholls and Ryder Phylogenetic models for Semitic vocabulary. 431	
Loquiha et al. Zero-Inflated Poisson and Negative Binomial Models Applied to Maternal Mortality Rate in Mozambique	346	Nicholls and Watt Partial Order Models for Episcopal Social Status in 12th Century England	437
Lynch and MacKenzie On Bivariate Survival Regression Models	352	Nysen et al. Testing Goodness-of-Fit of Parametric Models for Censored Data	441
Marchetti et al. Regression graph models: an application to joint modelling of fertility intentions among childless couples	358	Oller and Gómez Testing against ordered alternatives with interval-censored data	445
Martínez-Beneito et al. A spatio-temporal monitoring system for Influenza-Like Illness incidence	364	Palarea-Albaladejo and Martín-Fernández Examining distance-based grouping on the simplex sample space: the fuzzy clustering case	450
Martínez-Coscollà et al. Bayesian hierarchical modelling for analyzing the efficiency in the European banking system.	368	Pardo and Pérez The use of GEE for analyzing housing prices .	454
Marx et al. Multidimensional Single-Index Signal Regression ...	372	Peng and MacKenzie Precision of estimators in interval censored parametric survival models	458
Mauff and Little Multivariate Nonlinear Multi-Level Mixed Effect Models: Techniques and Application to Pharmacokinetic Data	378	Pennino et al. A Bayesian spatial approach to modelling fish species occurrence.	464
Mayr et al. Boosting Generalized Additive Models for Location, Scale and Shape	384	Pereira et al. The truncated inflated beta regression	468
Menten et al. Estimation of Infection Rates from Repeated ELISA Optical Density Data using Hidden Markov Models	390	Perra et al. A Bayesian analysis of survival times for stage IV non-small cells lung cancer	472
Mirkov and Friedl Nonlinear and Spline Regression Models for Forecasting Gas Flow on Exits of Gas Transmission Networks	394	Pfeifer On probabilities of avalanches triggered by alpine skiers. Models with random effects taking the stratified data into account.	476
Mohd Din et al. Prediction of the rheumatoid arthritis activity score: a joint modeling approach	400	Pita-Fernández et al. Cancer incidence in kidney transplant recipients	480
Molanes-López et al. Covariate-adjusted inference for the Youden index and associated classification threshold	404	Pomann et al. Evaluating Change Detection in Data Streams .	486

Porcu et al. Modelling the Timing of Divorce in Italy: a survival analysis based on regression quantiles	490	Stöber and Czado A Markov switching model for vine copulas	581
Prieto et al. Estimation of the density of the Antarctic Blue whales population using their sequences of sounds	494	Sweeney and Haslett Bayesian residual analysis in Poisson regression models	587
Ramsey and Futschik Optimal DNA Pooling for the Detection of Single Nucleotide Polymorphisms	499	Tamura and Giampaoli Prediction for an observation in a new cluster for Multilevel Logistic Regression considering k random coefficients	593
Riebler et al. Modelling seasonal patterns in longitudinal profiles with correlated circular random walks	503	Taylor and Einbeck Multivariate regression smoothing through the “falling net”	597
Rippe and Eilers Segmented smoothing with an L_0 penalty	509	Tharmaratnam and Claeskens Robust model selection in additive penalized regression splines models	603
Rodríguez-Álvarez et al. Testing for covariate effects in ROC-GAM regression models based on bootstrap methods	515	Thompson Statistical modeling of geographic risks for very low birth weights near Texas superfund sites	607
Rodríguez-Díaz et al. D-Optimum designs in random effect logistic regression models	519	Ugarte et al. Spatio-temporal risk smoothing and forecasting with P-splines	612
Rosen et al. Adaptive Spectral Estimation for Nonstationary Time Series	523	Urbano et al. Bioassays models with natural mortality and random effects	616
Rushworth et al. Distributed lag models for hydrological data	529	Usuga et al. A study to compare HGLM and GAMLSS in mixed linear models	622
Russo et al. Exact and approximate inferences for nonlinear mixed-effects heavy-tailed models	534	Van den Hout et al. A latent-class semi-parametric change point model for cognitive ability in older age	626
Sabanés Bové et al. Hyper- g Priors for Generalised Additive Model Selection	538	Van Oirbeek and Lesaffre Measuring the Brier score for frailty models	632
Schnabel et al. Optimal time scaling for plant growth analysis	544	Ventrucci et al. A Dipole Model for MEG Data	636
Sellers Introducing a Model to Determine True Counts via the Conway-Maxwell-Poisson Distribution	548	Ventura and Racugno A Bayesian adjustment of the modified profile likelihood	642
Sikorska et al. Fast genome-wide association analysis in longitudinal studies	553	Waldmann and Kneib Bayesian Structured Additive Quantile Regression	648
Singh and Huzubazar Analysis of Gene Duplication Data	557	West et al. Groups within networks	652
Slaets et al. Flexible Modelling of Functional Data using Continuous Wavelet Dictionaries	561	Worton and McLellan Robust mixture modelling of telemetry data in wildlife studies of home range	656
Smith and Bowman Boundary identification in 3D images	565	Yee and Hadi Row-Column Association Models	660
Sobotka et al. Confidence intervals for geoadditive expectile regression models	571	Ziegler-Graham and Rohde Use of Marginal Likelihoods in Statistical Inference	666
Stefanova Measuring Efficiency of Trial Designs with Unreplicated or Partially Replicated Test Lines	577		

impacts upon subject classification more than the model's fixed effects, as these are simply the average of individual fitted curves whilst subject classification is based on individual curves. Subject classification is key to the utility of GMMs; models that capture model-generated autocorrelation within the GMM framework are thus preferred. Whilst the exact choice of parameterization remains open, our findings suggest that some kind of explicit modelling of autocorrelation is warranted in these types of models. In any event, correct parameterization of the random structure is needed for growth mixture modelling of outcomes that exhibit less within-subject than between-subject heterogeneity.

Berkey, C.S. & Colditz, G.A. (2007). Adiposity in adolescents: change in actual BMI works better than change in BMI z score for longitudinal studies. *Ann. Epidemiol.*, **17**(1), 44-50.

Bollen, K. & Curran, P. (2006). *Latent curve models*, 2nd ed. New York, Wiley.

Duncan, T.E., Duncan, S.E., & Stryker, L.A. (2006). *An introduction to latent variable growth curve modeling*, 2nd ed. Mahwah, NJ, Lawrence Erlbaum Associates Inc.

Goodman, E., Adler, N.E., Daniels, S.R., Morrison, J.A., Slap, G.B., & Dolan, L.M. (2003). Impact of objective and subjective social status on obesity in a biracial cohort of adolescents. *Obes. Res.*, **11**(8), 1018-1026.

Kreuter, F. & Muthen, B. (2008). Analyzing Criminal Trajectory Profiles: Bridging Multilevel and Group-based Approaches Using Growth Mixture Modeling. *Journal of Quantitative Criminology*, **24**(1), 1-31.

Kuczmański, R.J., Ogden, C.L., Grummer-Strawn, L.M., Flegal, K.M., Guo, S.S., Wei, R., Mei, Z., Curtin, L.R., Roche, A.F., & Johnson, C.L. (2000). CDC growth charts: United States. *Adv. Data*, **314**, 1-27.

Li, C., Goran, M.I., Kaur, H., Nollen, N., & Ahluwalia, J.S. (2007). Developmental trajectories of overweight during childhood: role of early life factors. *Obesity (Silver. Spring)*, **15**(3), 760-771.

Mustillo, S., Worthman, C., Erkanli, A., Keeler, G., Angold, A., & Costello, E.J. (2003). Obesity and psychiatric disorder: developmental trajectories. *Pediatrics*, **111**(4-1), 851-859.

Needham, B.L., Epel, E.S., Adler, N.E., & Kiefe, C. (2010). Trajectories of change in obesity and symptoms of depression: the CARDIA study. *Am. J. Public Health*, **100**(6), 1040-1046.

Modeling swimming marks through Blocks and POT methods

Dulce Gomes¹, Júlia Teles², Luísa Canto e Castro³

¹ Departamento de Matemática/CIMA, Escola de Ciências e Tecnologia, Universidade de Évora, Rua Romão Ramalho, 59, 7000-671 Évora, Portugal (email: dmog@uevora.pt)

² Departamento de Métodos Matemáticos/CIPER, Faculdade de Motricidade Humana, Universidade Técnica de Lisboa, Estrada da Costa, 1495-688 Cruz Quebrada-Dafundo, Portugal (email: jteles@fmh.utl.pt)

³ Departamento de Estatística e Investigação Operacional/CEAUL, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Edifício C6, 1749-016 Lisboa, Portugal (email: luísa.loura@deio.fc.pt)

Abstract: The swimming marks in the 100m men's freestyle long course are modelled using extreme value theory. Using the statistical package R, extreme value and generalized Pareto models were adjusted in order to estimate the left endpoint of these models. The left endpoint can be interpreted as the best mark that can ever be reached, admitting that swimming pool conditions, athlete's equipment and training methods remain the same.

Keywords: Extreme value models; Generalized Pareto models; Blocks method; Peaks over threshold method; Swimming marks.

1 Introduction

Extreme value models are frequently used for the analysis of samples of maximum or minimum and generalized Pareto models are commonly used to analysing the samples of exceedances over a high or low threshold.

The swimming marks of 100m men's freestyle (long course), that appear in FINA ("La Federation Internationale de Natation") Website (www.fina.org), are the personal best in a very large sample of marks, so we could say that we are in the presence of extreme value — in this case minima. In this sense, we consider analysing and modeling this type of data by means of extreme value models.

Using two methods of extreme value analysis — the blocks method (De Haan and Ferreira, 2001) and the peaks over threshold (POT) method (Pickands, 1975; Robinson and Tawn, 1995) — we are going to model the swimming marks and estimate their left endpoint. This left endpoint can be interpreted as the best mark that can ever be reached, admitting that swimming pool conditions, athlete's equipment and methods of training remain the same.

TABLE 1. The three best annual marks (in seconds), through 1948 to 2010, of Men's Long Course World Records in 100m freestyle.

year	rank	athlete	mark	nationality
1948	1	Wally Ris	57.3	USA
1948	2	Keith Carter	57.6	USA
1948	3	Alan Ford	57.8	USA
...
1999	1	Pieter van den Hoogenband	48.35	NED
1999	2	Michael Klim	48.73	AUS
1999	3	Alexander Popov	48.82	RUS
2000	1	Pieter van den Hoogenband	47.84	NED
2000	2	Michael Klim	48.18	AUS
2000	3	Alexander Popov	48.27	RUS
...
2010	1	Simon Burnett	48.54	GBR
2010	2	William Meynard	48.56	FRA
2010	3	Kyle Richardson	48.69	AUS

A draft of the dataset with the three best annual marks (in seconds) of the men's long course world records in 100m freestyle are presented in the Table 1. The information was available from 1948 to 2010, with missing value for 1950 and 1951.

In Figure 1, the marks are plotted against the year and by ranking. As we expected, there is a decreasing trend in the marks. So, the relevant question is: Until when these marks could fall? In order to give answer to this question two different approaches of extreme value theory were used.

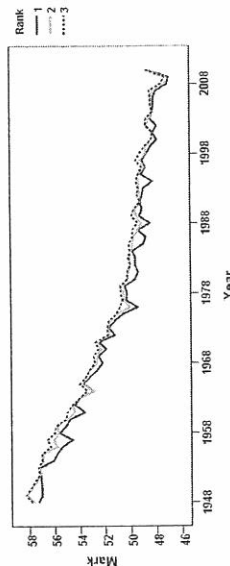


FIGURE 1. Scatter plot of mark against year by ranking.

2 Two different approaches

To apply the extreme value methodology we must have independent and identically distributed observations (Reiss and Thomas, 2001). As we can observe in Table 1, there are some athletes (e.g., Michael Klim and Alexander Popov) which contributed with more than one mark. So to use this type

of analysis we only select the best mark of each athlete. In order to adjust an extreme value or generalized Pareto model, the trend also needs to be removed.

In the POT approach the inference is based in the exceedances over a high threshold that is unknown. Our empirical way of choosing this threshold was through the analysis of the diagram of the shape parameter's estimates. To apply the POT method we adjust a model in the family of generalized Pareto models. For different shape, location and scale parameters we obtain three different submodel families: exponential, Pareto and beta.

In the block method we adjust a model in the family of extreme value models. Also depending on the shape, location and scale parameters we could reach the Gumbel, Fréchet or Weibull submodels.

Acknowledgments: This research was partially support by the Center of Mathematics and Applications, University of Évora, by the Interdisciplinary Centre for the Study of Human Performance, Technical University of Lisbon, and by the Center of Statistics and Applications, University of Lisbon, through the Programs FCT/POCTI, FCT/POCH2010 and POCI/FEDER.

References

- De Haan, L., and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Boston: Springer.
- Pickands, J. (1995). Statistical inference using extreme value order statistics. *Annals of Statistics*, **3**, 119-131.
- R Development Core Team (2010). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria. (Available at <http://www.R-project.org/>).
- Reiss, R.D., and Thomas, M. (2001). *Statistical Analysis of Extreme Values*, 2nd edition. Basel: Birkhäuser.
- Robinson, M.E., and Tawn, J.A. (1995). Statistics for expected athletics records. *Applied Statistics*, **44**, 499-511.