Contents

BJPS is an IMS supported journal

ISSN 0103-0752

Brazilian Statistical Association

Brazilian Journal of Probability and Statistics

Volume 37 • Number 4 • December 2023

Brazilian Journal of Probability and Statistics

Volume 37 • Number 4 • December 2023

www.imstat.org/bjps

ISSN 0103-0752 (Print) ISSN 2317-6199 (Online), Volume 37, Number 4, December 2023. Published quarterly by the Brazilian Statistical Association.

POSTMASTER:

Send address changes to Brazilian Journal of Probability and Statistics, Institute of Mathematical Statistics, Dues and Subscriptions Office, PO Box 729, Middletown, Maryland 21769, USA.

Brazilian Statistical Association members should send address changes to Rua do Matão, 1010 sala 250A, 05508-090 São Paulo/SP Brazil (address of the BSA office).

Printed in the United States of America

Partial financial support:
CNPq and CAPES (Brazil).

Brazilian Journal of Probability and Statistics
2023, Vol. 37, No. 4, 643–666
https://doi.org/10.1214/23-BJPS585
This research was funded, in part, by the GAMBAS project funded by the French National Research Agency (ANR-18-CE02-0025).

On quasi Pólya thinning operator

Jean Peyhardi[a]

IMAG, University of Montpellier, CNRS, Montpellier, France, [a]*jean.peyhardi@umontpellier.fr*

Abstract. Thinning operation is a stochastic operation that shrinks a random count variable into a smaller one. This kind of random operation has been intensively studied during the seventies to characterize some count distributions, such as the Poisson distribution using the binomial thinning operator (also named binomial damage model). Then, the closure under thinning operator has been studied in order to define some classes of integer valued autoregressive (INAR) models for count time series. These two properties will be studied in this paper for the new class of quasi Pólya thinning operators. Classical results concerning the binomial thinning operator are recovered as a special case. The quasi Pólya thinning operator is related to the new class of quasi Pólya splitting distributions, defined for multivariate count data. The probabilistic graphical model (PGM) of these multivariate distributions is characterized. Finally a general class of integer valued autoregressive models is introduced, including the usual cases of Poisson marginal or generalized Poisson as a special cases and the generalized negative binomial as a new case.

References

Al-Osh, M. A. and Alzaid, A. A. (1991). Binomial autoregressive moving average models. *Stochastic Models* **7**, 261–282. MR1107412 https://doi.org/10.1080/15326349108807188

Alzaid, A. A. and Al-Osh, M. A. (1993). Some autoregressive moving average processes with generalized Poisson marginal distributions. *Annals of the Institute of Statistical Mathematics* **45**, 223–232. MR1232490 https://doi.org/10.1007/BF00775809

Bol'shev, L. N. (1965). On a characterization of the Poisson distribution and its statistical applications. *Theory of Probability and Its Applications* **10**, 446–456.

Castañer, A., Claramunt, M. M., Lefèvre, C. and Loisel, S. (2015). Discrete Schur-constant models. *Journal of Multivariate Analysis* **140**, 343–362. MR3372573 https://doi.org/10.1016/j.jmva.2015.06.003

Davis, R. A., Fokianos, K., Holan, S. H., Joe, H., Livsey, J., Lund, R., Pipiras, V. and Ravishanker, N. (2021). Count time series: A methodological review. *Journal of the American Statistical Association* **116**, 1533–1547. MR4309291 https://doi.org/10.1080/01621459.2021.1904957

Gupta, R. C. (1974). Modified power series distribution and some of its applications. *Sankhya Series B*, 288–298. MR0391334

Jambunathan, M. V. (1954). Some properties of beta and gamma distributions. *The Annals of Mathematical Statistics*, 401–405. MR0061317 https://doi.org/10.1214/aoms/1177728800

Janardan, K. G. (1973). A new four urn model with predetermined strategy. Technical report 37-I, Sangamon State University, Springfield IL.

Janardan, K. G. (1975). Markov–Polya urn-model with pre-determined strategies. *Gujarat Statistical Review* **2**, 17–32. MR0391217

Janardan, K. G. and Rao, R. B. (1982). Characterization of generalized Markov–Polya and generalized Polya–Eggenberger distributions. *Communications in Statistics—Theory and Methods* **11**, 2113–2124. MR0673456 https://doi.org/10.1080/03610928208828374

Janardan, K. G. and Rao, R. B. (1986). Identifiability of a generalized Mabkov–Polya damage model. *Statistics* **17**, 303–310. MR0838720 https://doi.org/10.1080/02331888608801940

Joe, H. (1996). Time series models with univariate margins in the convolution-closed infinitely divisible class. *Journal of Applied Probability* **33**, 664–677. MR1401464 https://doi.org/10.2307/3215348

Key words and phrases. Quasi Polya distribution, binomial thinning operator, probabilistic graphical model.

Jones, M. C. and Marchand, É. (2019). Multivariate discrete distributions via sums and shares. *Journal of Multivariate Analysis* **171**, 83–93. MR3886395 https://doi.org/10.1016/j.jmva.2018.11.011

Jorgensen, B. (1997). *The Theory of Dispersion Models*. Boca Raton: CRC Press. MR1462891

Jourdan, A. and Kokonendji, C. C. (2002). Surdispersion et modèle binomial négatif généralisé. *Revue de Statistique Appliquée* **50**, 73–86.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge: MIT press. MR2778120

Lauritzen, S. L. (1996). *Graphical Models, Vol. 17*. Oxford: Clarendon Press. MR1419991

Lukacs, E. (1955). A characterization of the gamma distribution. *The Annals of Mathematical Statistics* **26**, 319–324. MR0069408 https://doi.org/10.1214/aoms/1177728549

McKenzie, E. (1985). Some simple models for discrete variate time series 1. *Journal of the American Water Resources Association* **21**, 645–650. MR1973555 https://doi.org/10.1016/S0169-7161(03)21018-X

Moran, P. A. P. (1952). A characteristic property of the Poisson distribution. In *Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 48* 206–207. Cambridge: Cambridge University Press. MR0045968 https://doi.org/10.1017/s0305004100027535

Ovaskainen, O. and Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R*. Cambridge: Cambridge University Press.

Patil, G. P. and Ratnaparkhi, M. V. (1975). Problems of damaged random variables and related characterizations. In *A Modern Course on Statistical Distributions in Scientific Work* 255–270. Berlin: Springer.

Peyhardi, J. (2023). Supplement to "On quasi Pólya thinning operator." https://doi.org/10.1214/23-BJPS585SUPP

Peyhardi, J. and Fernique, P. (2017). Characterization of convolution splitting graphical models. *Statistics & Probability Letters* **126**, 59–64. MR3634580 https://doi.org/10.1016/j.spl.2017.02.018

Peyhardi, J., Fernique, P. and Durand, J.-B. (2021). Splitting models for multivariate count data. *Journal of Multivariate Analysis* **181**, 104677. MR4161849 https://doi.org/10.1016/j.jmva.2020.104677

Peyhardi, J., Laroche, F. and Mortier, F. (2022). Pólya-splitting distributions as stationary solutions of multivariate birth-death processes under extended neutral theory. Available at SSRN 4412745.

Puig, P. and Valero, J. (2007). Characterization of count data distributions involving additivity and binomial subsampling. *Bernoulli*, 544–555. MR2331263 https://doi.org/10.3150/07-BEJ6021

Rao, B. R. and Janardan, K. G. (1984). The use of the generalized Markov–Polya distribution as a random damage model and its identifiability. *Sankhya Series A*, 458–462. MR0798051

Rao, C. R. (1949). On some problems arising out of discrimination with multiple characters. *Sankhya The Indian Journal of Statistics*, 343–366. MR0033501

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. *Sankhya Series A*, 311–324. MR0208736

Rao, C. R. and Rubin, H. (1964). On a characterization of the Poisson distribution. *Sankhya Series A*, 295–298. MR0184320

Scotto, M. G., Weiß, C. H. and Gouveia, S. (2015). Thinning-based models in the analysis of integer-valued time series: A review. *Statistical Modelling* **15**, 590–618. MR3441230 https://doi.org/10.1177/1471082X15584701

Shanbhag, D. N. and Panaretos, J. (1979). Some results related to the Rao–Rubin characterization of the Poisson distribution. *Australian Journal of Statistics* **21**, 78–83. MR0531070 https://doi.org/10.1111/j.1467-842x.1979.tb01122.x

Sprott, D. A. (1965). A class of contagious distributions and maximum likelihood estimation. *Sankhya Series A*, 369–382. MR0212934

Steutel, F. W., Kent, J. T., Bondesson, L. and Barndorff-Nielsen, O. (1979). Infinite divisibility in theory and practice [with discussion and reply]. *Scandinavian Journal of Statistics*, 57–64.

Steutel, F. W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *Annals of Probability*, 893–899. MR0542141

Tang, Z.-Z. and Chen, G. (2019). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics* **20**, 698–713. MR4019726 https://doi.org/10.1093/biostatistics/kxy025

Teicher, H. (1954). On the convolution of distributions. *The Annals of Mathematical Statistics*, 775–778. MR0065052 https://doi.org/10.1214/aoms/1177728664

Wang, T. and Zhao, H. (2017). A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics* **73**, 792–801. MR3713113 https://doi.org/10.1111/biom.12654

Weiß, C. H. (2008). Thinning operations for modeling time series of counts—a survey. *AStA Advances in Statistical Analysis* **92**, 319–341. MR2426093 https://doi.org/10.1007/s10182-008-0072-3

Xekalaki, E. (1986). The multivariate generalized Waring distribution. *Communications in Statistics—Theory and Methods* **15**, 1047–1064. MR0832095 https://doi.org/10.1080/03610928608829168

Brazilian Journal of Probability and Statistics
2023, Vol. 37, No. 4, 667–692
https://doi.org/10.1214/23-BJPS586

Two-stage Walsh-average-based robust estimation and variable selection for partially linear additive spatial autoregressive models

Zitong Li and Yunquan Song[a]

College of Science, China University of Petroleum, Qingdao 266580, P.R. China, [a]*statistics99@163.com*

Abstract. In this paper, we consider the estimation and variable selection problem for partially linear additive spatial autoregressive models (PLASARM). We propose a robust estimation of two-stage Walsh-average regression (2SWAR) based on Walsh-average regression and instrumental variable method. Under some mild conditions, we obtain and theoretically prove the asymptotic normality of finite parameter vectors and the convergence rate of the nonparametric part. In addition, We also propose a robust variable selection method and further demonstrate its ability to consistently identify real models. We further carry out Monte Carlo simulation and real data analysis, both of which yield promising numerical results.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Berlin: Springer.

Anselin, L. and Bera, A. K. (1995). Spatial dependence in linear regression models with an introduction to spatial econometrics.

Bowden, R. J. and Turkington, D. A. (1990). *Instrumental Variables*. **8**. Cambridge: Cambridge University Press. MR1113481

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* **80**, 580–598. MR0803258

Case, A. C. (1991). Spatial patterns in household demand. *Econometrica: Journal of the Econometric Society*, 953–965.

Cheng, S. and Chen, J. (2019). Estimation of partially linear single-index spatial autoregressive model. *Statistical Papers* **62**, 495–531. MR4220889 https://doi.org/10.1007/s00362-019-01105-y

Cliff, A. D. (1973). Spatial autocorrelation. Technical report.

Du, J., Sun, X., Cao, R. and Zhang, Z. (2018). Statistical inference for partially linear additive spatial autoregressive models. In *Spatial Statistics*. MR3809255 https://doi.org/10.1016/j.spasta.2018.04.008

Du, P., Cheng, G. and Liang, H. (2012). Semiparametric regression models with additive nonparametric components and high dimensional parametric components. *Computational Statistics & Data Analysis* **56**, 2006–2017. MR2892394 https://doi.org/10.1016/j.csda.2011.12.007

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

Feng, L., Zou, C. and Wang, Z. (2012). Local Walsh-average regression. *Journal of Multivariate Analysis* **106**, 36–48. MR2887678 https://doi.org/10.1016/j.jmva.2011.12.003

Guo, J., Tang, M.-L., Tian, M. and Zhu, K. (2013). Variable selection in high-dimensional partially linear additive models for composite quantile regression. *Computational Statistics & Data Analysis* **65**, 56–67. MR3064943 https://doi.org/10.1016/j.csda.2013.03.017

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management* **5**, 81–102.

Hettmansperger, T. P. and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. MR1604954

Huang, J. Z., Wu, C. O. and Zhou, L. (2003). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. MR2087972

Kai, B., Li, R. and Zou, H. (2011). New efficient estimation and variable selection methods for semiparametric varying-coefficient partially linear models. *The Annals of Statistics* **39**, 305. MR2797848 https://doi.org/10.1214/10-AOS842

Key words and phrases. Partially linear additive spatial autoregressive models, robust estimation, robust variable selection, Walsh-average regression, instrumental variable.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* **17**, 99–121.

Kelejian, H. H. and Prucha, I. R. (1999). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* **40**, 509–533. MR1690289 https://doi.org/10.1111/1468-2354.00027

Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* **104**, 219–257. MR1864417 https://doi.org/10.1016/S0304-4076(01)00064-1

Kim, M.-O. (2007). Quantile regression with varying coefficients. *The Annals of Statistics* **35**, 92–108. MR2332270 https://doi.org/10.1214/009053606000000966

Koch, M. and Krisztin, T. (2011). Applications for asynchronous multi-agent teams in nonlinear applied spatial econometrics. *Journal of Internet Technology* **12**, 1007–1014.

Krisztin, T. (2017). The determinants of regional freight transport: A spatial, semiparametric approach. *Geographical Analysis* **49**, 268–308.

Lee, L. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews* **22**, 307–335. MR2018884 https://doi.org/10.1081/ETC-120025891

Lee, L. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72**, 1899–1925. MR2095537 https://doi.org/10.1111/j.1468-0262.2004.00558.x

LeSage, J. P. (1997). Bayesian estimation of spatial autoregressive models. *International Regional Science Review* **20**, 113–129.

LeSage, J. P. (2010). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis* **32**, 19–35.

LeSage, J. P. and Parent, O. (2005). Bayesian model averaging for spatial econometric models. *Econometrics eJournal.*

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics* **361**, 261–286. MR2387971 https://doi.org/10.1214/009053607000000604

Ling, N., Aneiros, G. and Vieu, P. (2020). knn estimation in functional partial linear modeling. *Statistical Papers* **61**, 423–444. MR4056810 https://doi.org/10.1007/s00362-017-0946-0

Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models (ss-09-140). *Statistica Sinica* **21**, 1225–1248. MR2827522 https://doi.org/10.5705/ss.2009.140

Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *Journal of Statistical Planning and Inference* **141**, 204–219. MR2719488 https://doi.org/10.1016/j.jspi.2010.05.028

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23. MR0035933 https://doi.org/10.1093/biomet/37.1-2.17

Pace, R. K. and Gilley, O. W. (1997). Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics* **14**, 333–340.

Su, L. and Jin, S. (2010). Profile quasi-maximum likelihood estimation of partially linear spatial autoregressive models. *Journal of Econometrics* **157**, 18–33. MR2652276 https://doi.org/10.1016/j.jeconom.2009.10.033

Su, L. and Yang, Z. (2007). Instrumental variable quantile estimation of spatial autoregressive models. MR2354954 https://doi.org/10.1016/j.jeconom.2005.10.004

Terpstra, J. and McKean, J. W. (2005). Rank-based analysis of linear models using R. *Journal of Statistical Software* **14**, 1–26.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Methodological)* **58**, 267–288. MR1379242

Wang, L., Kai, B. and Li, R. (2009). Local rank inference for varying coefficient models. *Journal of the American Statistical Association* **104**, 1631–1645. MR2597005 https://doi.org/10.1198/jasa.2009.tm09055

Wang, L. and Li, R. (2009). Weighted wilcoxon-type smoothly clipped absolute deviation method. *Biometrics* **65**. MR2751481 https://doi.org/10.1111/j.1541-0420.2008.01099.x

Xie, H. and Huang, J. (2009). Scad-penalized regression in high-dimensional partially linear models. *The Annals of Statistics* **37**, 673–696. MR2502647 https://doi.org/10.1214/07-AOS580

Xue, H., Lam, K. F. and Li, G. (2004). Sieve maximum likelihood estimator for semiparametric regression models with current status data. *Journal of the American Statistical Association* **99**, 346–356. MR2062821 https://doi.org/10.1198/016214504000000313

Yao, W. and Wang, Q. (2013). Robust variable selection through mave. *Computational Statistics & Data Analysis* **63**, 42–49. MR3040249 https://doi.org/10.1016/j.csda.2013.01.021

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

Zhang, H. H., Cheng, G. and Liu, Y. (2011). Linear or nonlinear? Automatic structure discovery for partially linear models. *Journal of the American Statistical Association* **106**, 1099–1112. MR2894767 https://doi.org/10.1198/jasa.2011.tm10281

Zhang, Y. and Shen, D. (2015). Estimation of semi-parametric varying-coefficient spatial panel data models with random-effects. *Journal of Statistical Planning and Inference* **159**, 64–80. MR3299089 https://doi.org/10.1016/j.jspi.2014.11.001

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802

Brazilian Journal of Probability and Statistics
2023, Vol. 37, No. 4, 693–719
https://doi.org/10.1214/23-BJPS587

Unit gamma regression models for correlated bounded data

João Victor B. de Freitas[1,a]**, Juvêncio S. Nobre**[2]**, Patricia L. Espinheira**[3] **and Leandro C. Rêgo**[2]

[1]*Departamento de Estatística, Instituto de Matemática, Estatística e Computação Científica, Universidade Estadual de Campinas, Brazil,* [a]*jvbfreitas@ime.unicamp.br*
[2]*Departamento de Estatística e Matemática Aplicada, Universidade Federal do Ceará, Brazil*
[3]*Departamento de Estatística, Universidade Federal da Bahia, Brazil*

Abstract. Experiments with repeated measures are the ones where more than one observation per subject is available. To model of such experiments, dependency within subjects needs to be taken into consideration. In cases where the variable of interest is bounded in (a, b) with $a < b$ known reals, there are few proposals to model correlated bounded data most part being based on Beta, Simplex and Unit gamma distributions. In particular, for marginal modeling of the mean and precision/dispersion, Simplex and Beta models based on Generalized Estimating Equations (GEE) are used. In this paper, to take account of possible within-subject dependence using the GEE approach, we proposed an Unit Gamma regression model used to modeling bounded data in a unit interval. In this paper, we developed residuals and influence diagnostic tools to the Simplex and Unit Gamma models for correlated bounded data. Furthermore, To assess the finite-sample performance of the proposed estimators, we conducted a Monte Carlo simulation study. The methodology is illustrated with the analysis of a real data set. An R package was developed for all the new methodology described in this paper.

References

Abbasi, T. and Abbasi, S. A. (2012). *Water Quality Indices*. Amsterdam: Elsevier.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* (E. Parzen, K. Tanabe and G. Kitagawa, eds.). *Springer Series in Statistics (Perspectives in Statistics)*. New York: Springer. MR1486823

Akdur, H. T. K. (2021). Unit-Lindley mixed-effect model for proportion data. *Journal of Applied Statistics* **48**, 2389–2405. MR4330746 https://doi.org/10.1080/02664763.2020.1823946

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. *Oxford Statistical Science Series*. Oxford: Clarendon.

Bonat, W. H., Petterle, R. R., Hinde, J. and Demétrio, C. G. B. (2019). Flexible quasi-beta regression models for continuous bounded data. *Statistical Modelling* **19**, 617–633. MR4025764 https://doi.org/10.1177/1471082X18790847

Bonat, W. H. and Jørgensen, B. (2016). Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society Series C Applied Statistics* **65**, 649–675. MR3564998 https://doi.org/10.1111/rssc.12145

Bonat, W. H., Lopes, J. E., Shimakura, S. E. and Ribeiro Jr, P. J. (2018). Likelihood analysis for a class of simplex mixed models. *Chilean Journal of Statistics* **9**, 3–17. MR3886127

Bonat, W. H., Ribeiro Jr, P. J. and Zeviani, W. M. (2015). Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics* **42**, 252–266. MR3276943 https://doi.org/10.1080/02664763.2014.947248

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing* **16**, 1190–1208. MR1346301 https://doi.org/10.1137/0916069

Cadigan, N. G. and Farrell, P. J. (2002). Generalized local influence with applications to fish stock cohort analysis. *Journal of the Royal Statistical Society Series C Applied Statistics* **51**, 469–483. MR1974073 https://doi.org/10.1111/1467-9876.00281

Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society, Series B, Methodological* **48**, 133–169. MR0867994

Key words and phrases. Unit Gamma distribution, correlated data, generalized estimating equations, longitudinal data, repeated measures, bounded data.

Crowder, M. (1987). On linear and quadratic estimating functions. *Biometrika* **74**, 591–597. MR0909363 https://doi.org/10.1093/biomet/74.3.591

Cui, J. (2007). QIC program and model selection in GEE analyses. *Stata Journal* **7**, 209–220.

Cui, J. and Qian, G. (2007). Selection of working correlation structure and best model in GEE analyses of longitudinal data. *Communications in Statistics Simulation and Computation* **36**, 987–996. MR2415699 https://doi.org/10.1080/03610910701539617

de Freitas, João Victor, B., Nobre, J. S., Bourguignon, M. and Santos-Neto, M. (2022). A new approach to modeling positive random variables with repeated measures. *Journal of Applied Statistics* **49**, 3784–3803. MR4502139 https://doi.org/10.1080/02664763.2021.1963422

de Freitas, J. V. B., Nobre, J. S., Espinheira, P. L. and Rêgo, L. C. (2023). Supplement to "Unit gamma regression models for correlated bounded data." https://doi.org/10.1214/23-BJPS587SUPP

Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford: Oxford University Press. MR2049007

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**, 236–244.

Espinheira, P. L. and de Oliveira Silva (2020). Residual and influence analysis to a general class of simplex regression. *Test* **29**, 523–552. MR4095039 https://doi.org/10.1007/s11749-019-00665-3

Espinheira, P. L. and Silva, A. O. (2019). Residual and influence analysis to a general class of simplex regression. *Test* **29**, 523–552. MR4095039 https://doi.org/10.1007/s11749-019-00665-3

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**, 799–815. MR2095753 https://doi.org/10.1080/0266476042000214501

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. P. (2013). Mixed beta regression: A Bayesian perspective. *Computational Statistics & Data Analysis* **61**, 137–147. MR3063006 https://doi.org/10.1016/j.csda.2012.12.002

Godambe, V. P. (1997). *Estimating Functions: A Synthesis of Least Squares and Maximum Likelihood Methods*. *Lecture Notes—Monograph Series* **32**, 5–15. MR1837793 https://doi.org/10.1214/lnms/1215455034

Grassia, A. (1977). On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions. *Australian Journal of Statistics* **19**, 108–114. MR0458671 https://doi.org/10.1111/j.1467-842x.1977.tb01277.x

Griffiths, D. and CSIRO, C. S. (1981). Closeness of Grassia's transformed gammas and the beta distribution. *Australian Journal of Statistics* **23**, 240–246. https://doi.org/10.1111/j.1467-842X.1981.tb00783.x

Guedes, A. C., Cribari-Neto, F. and Espinheira, P. L. (2020). Modified likelihood ratio tests for unit gamma regressions. *Journal of Applied Statistics* **47**, 1562–1586. MR4115292 https://doi.org/10.1080/02664763.2019.1683152

Hardin, J. W. and Hilbe, J. M. (2012). *Generalized Estimating Equations*, 2nd ed. Boca Raton: Chapman and Hall/CRC. MR3134775

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226–252. MR0055650 https://doi.org/10.2307/3001853

Jørgensen, B. (1997). *The Theory of Dispersion Models*. London: Chapman and Hall. MR1462891

Jørgensen, B. and Labouriau, R. S. (1992). *Famílias exponenciais e Inferência Teórica*. *Monografias de Matemática no 52*. Rio de Janeiro: IMPA. MR1183933

Jørgensen, B., Lundbye-Christensen, S., Song, P. X.-K. and Sun, L. (1996). State-space models for multivariate longitudinal data of mixed types. *Canadian Journal of Statistics* **24**, 385–402. MR1421864 https://doi.org/10.2307/3315747

Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. *Statistical Modelling* **3**, 193–213. MR2005473 https://doi.org/10.1191/1471082X03st053oa

Kong, M., Xu, S., Levy, S. M. and Datta, S. (2015). GEE type inference for clustered zero-inflated negative binomial regression with application to dental caries. *Computational Statistics & Data Analysis* **85**, 54–66. MR3299782 https://doi.org/10.1016/j.csda.2014.11.014

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963–974.

Lee, S.-Y. and Xu, L. (2004). Influence analyses of nonlinear mixed-effects models. *Computational Statistics & Data Analysis* **45**, 321–341. MR2045635 https://doi.org/10.1016/S0167-9473(02)00303-1

Lemonte, A. J. and Bazán, J. L. (2016). New class of Johnson distributions and its associated regression model for rates and proportions. *Biometrical Journal* **58**, 727–746. MR3527412 https://doi.org/10.1002/bimj.201500030

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22. MR0836430 https://doi.org/10.1093/biomet/73.1.13

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* **46**, 673–687. MR1085815 https://doi.org/10.2307/2532087

Mitnik, P. A. and Baek, S. (2013). The Kumaraswamy distribution: Median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers* **54**, 177–192. MR3016961 https://doi.org/10.1007/s00362-011-0417-y

Mousa, A. M., El-Sheikh, A. A. and Abdel-Fattah, M. A. (2016). A gamma regression for bounded continuous variables. *Advances and Applications in Statistics* **49**, 305–326.

Nobre, J. S., Singer, J. M. and Sen, P. K. (2013). U-tests for variance components in linear mixed models. *Test* **22**, 580–605. MR3122324 https://doi.org/10.1007/s11749-013-0316-8

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120–125. MR1833297 https://doi.org/10.1111/j.0006-341X.2001.00120.x

Petterle, R. R., Taconeli César, A., da Silva, J. L. P., da Silva, G. P., Laureano, H. A. and Bonat, W. H. (2021). Unit gamma mixed regression models for continuous bounded data. *Journal of Statistical Computation and Simulation* **0**, 1–19. MR4575532 https://doi.org/10.1080/00949655.2021.1970164

Petterle, R. R., Bonat, W. H. and Scarpin, C. T. (2019). Quasi-beta longitudinal regression model applied to water quality index data. *Journal of Agricultural, Biological, and Environmental Statistics* **24**, 346–368. MR3945283 https://doi.org/10.1007/s13253-019-00360-8

Poon, W.-Y. and Poon, Y. S. (1999). Conformal normal curvature and assessment of local influence. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **61**, 51–61. MR1664096 https://doi.org/10.1111/1467-9868.00162

Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalised estimating equations. *Biometrika* **83**, 551–562.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–839. MR1141951 https://doi.org/10.2307/2532642

Qiu, Z., Song, P. X.-K. and Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics* **35**, 577–596. MR2468863 https://doi.org/10.1111/j.1467-9469.2008.00603.x

R Core Team (2019). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. **44**, 50–57. MR0024111

Ratnaparkhl, M. V. and Mosimann, J. E. (1990). On the normality of transformed beta and unit-gamma random variables. *Communications in Statistics—Theory and Methods* **19**, 3833–3854.

Ribeiro, V. O., Nobre, J. S., dos Santos, J. R. and Azevedo, C. L. N. (2021). Beta-rectangular regression models to longitudinal data. *Brazilian Journal of Probability and Statistics* **35**, 851–874. MR4350964 https://doi.org/10.1214/21-bjps511

Rocha, S. S., Espinheira, P. L. and Cribari-Neto, F. (2020). Residual and local influence analyses for unit gamma regressions. *Statistica Neerlandica*. MR4245906 https://doi.org/10.1111/stan.12229

Sarvi, F., Moghimbeigi, A. and Mahjub, H. (2019). GEE-based zero-inflated generalized Poisson model for clustered over or under-dispersed count data. *Journal of Statistical Computation and Simulation* **89**, 2711–2732. MR3979726 https://doi.org/10.1080/00949655.2019.1632857

Schwarzmann, B. (1991). A connection between local-influence analysis and residual diagnostics. *Technometrics* **33**, 103–104.

Sen, P. K., Singer, J. M. and de Lima, A. C. P. (2009). *From Finite Sample to Asymptotic Methods in Statistics*, 1st ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **29**. Cambridge: Cambridge University Press. MR2566692

Silva, A. R. S., Azevedo, C. L. N., Bazán, J. L. and Nobre, J. S. (2021). Augmented-limited regression models with an application to the study of the risk perceived using continuous scales. *Journal of Applied Statistics* **48**, 1998–2021. MR4296776 https://doi.org/10.1080/02664763.2020.1783518

Simas, A. B. Barreto-Souza, W. and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis* **54**, 348–366. MR2756431 https://doi.org/10.1016/j.csda.2009.08.017

Singer, J. M., Rocha, F. M. M. and Nobre, J. S. (2017). Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review* **85**, 290–324. MR3686569 https://doi.org/10.1111/insr.12178

Song, P. X.-K., Qiu, Z. and Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal* **46**, 540–553. MR2101142 https://doi.org/10.1002/bimj.200110052

Song, P. X.-K. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics* **56**, 496–502.

Tadikamalla, P. R. (1981). On a family of distributions obtained by the transformation of the gamma distribution. *Journal of Statistical Computation and Simulation* **13**, 209–214.

Tan, M., Qu, Y. and Kutner, M. H. (1997). Model diagnostics for marginal regression analysis of correlated binary data. *Communications in Statistics Simulation and Computation* **26**, 539–558.

Tsuyuguchi, A. B., Paula, G. A. and Barros, M. (2020). Analysis of correlated Birnbaum–Saunders data based on estimating equations. *Test* **29**, 661–681. MR4140778 https://doi.org/10.1007/s11749-019-00675-1

Venezuela, M. K. (2008). Generalized estimating equation and local influence to beta regression models with repeated measures. Unpublished Ph.D. thesis, Departamento de Estatística, Universidade de São Paulo, Brazil (in Portuguese).

Venezuela, M. K., Botter, D. A. and Sandoval, M. C. (2007). Diagnostic techniques in generalized estimating equations. *Journal of Statistical Computation and Simulation* **77**, 879–888. MR2409950 https://doi.org/10.1080/10629360600780488

Venezuela, M. K., Sandoval, M. C. and Botter, D. A. (2011). Local influence in estimating equations. *Computational Statistics & Data Analysis* **55**, 1867–1883. MR2748685 https://doi.org/10.1016/j.csda.2010.10.020

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* **54**, 426–482. MR0012401 https://doi.org/10.2307/1990256

Brazilian Journal of Probability and Statistics
2023, Vol. 37, No. 4, 720–734
https://doi.org/10.1214/23-BJPS589

Divide-and-conquer Metropolis–Hastings samplers with matched samples

Hang Qian[a]

The MathWorks, Inc. 55 Centre Street, Natick, Massachusetts 01760, USA, [a]*matlabist@gmail.com*

Abstract. Divide-and-conquer methods for scalable Bayesian inference divide the massive data into subsets, sample from the subset posterior distributions, and then combine the results. We develop an asymptotically exact recombination method by matched samples. Subset posterior densities calculated by the Metropolis–Hastings samplers are recycled for evaluating the importance weight to reduce the computational burden. Our computationally efficient aggregation algorithm features a collection of consistent estimators of expectations with respect to the full posterior distribution. Weight degeneracy of the importance sampling is resolved by the matched-sample resample-move method, which handles heterogeneous and non-overlapping subposteriors. Numeric examples and real-world mortgage data applications demonstrate excellent performance of the novel approach.

References

Banterle, M., Crazian, C., Lee, A. and Robert, C. (2019). Accelerating Metropolis–Hastings algorithms by delayed acceptance. *Foundations of Data Science* **1**, 103–128. MR4617837 https://doi.org/10.3934/fods.2019005

Bardenet, R., Doucet, A. and Holmes, C. (2017). On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research* **18**, 1–43. MR3670492

Bierkens, J., Fearnhead, P. and Roberts, G. (2019). The zig-zag process and super-efficient sampling for Bayesian analysis of big data. *The Annals of Statistics* **47**, 1288–1320. MR3911113 https://doi.org/10.1214/18-AOS1715

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika* **89**, 539–551. MR1929161 https://doi.org/10.1093/biomet/89.3.539

Gilks, W. R. and Berzuini, C. (2001). Following a moving target—Monte Carlo inference for dynamic Bayesian models. *Journal of the Royal Statistical Society, Series B* **63**, 127–146. MR1811995 https://doi.org/10.1111/1467-9868.00280

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109. MR3363437 https://doi.org/10.1093/biomet/57.1.97

Korattikara, A., Chen, Y. and Welling, M. (2014). Austerity in MCMC land: Cutting the Metropolis–Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning 1* 181–189.

Maclaurin, D. and Adams, R. (2014). Firefly Monte Carlo: Exact MCMC with subsets of data. In *Uncertainty in Artificial Intelligence—Proceedings of the 30th Conference, UAI 2014* 4289–4295.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087–1092.

Minsker, S., Srivastava, S., Lin, L. and Dunson, D. (2017). Robust and scalable Bayes via a median of subset posterior measures. *Journal of Machine Learning Research* **18**, 1–40. MR3763758

Neiswanger, W., Wang, C. and Xing, E. P. (2014). Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence* 623–632. Arlington: AUAI Press.

Nemeth, C. and Sherlock, C. (2018). Merging MCMC subposteriors through Gaussian-process approximations. *Bayesian Analysis* **13**, 507–530. MR3780433 https://doi.org/10.1214/17-BA1063

Quiroz, M., Kohn, R., Villani, M. and Tran, M. (2019). Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association* **114**, 831–843. MR3963184 https://doi.org/10.1080/01621459.2018.1448827

Scott, S. L. (2017). Comparing consensus Monte Carlo strategies for distributed Bayesian computation. *Brazilian Journal of Probability and Statistics* **31**, 668–685. MR3738171 https://doi.org/10.1214/17-BJPS365

Key words and phrases. Bayes, Big data, Markov chain Monte Carlo, Parallel computing.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management* **11**, 78–88.

Smets, F. and Wouters, R. (2007). Shocks and frictions in US business cycles: A Bayesian DSGE approach. *American Economic Review* **97**, 586–608. MR2221894 https://doi.org/10.1002/jae.834

Srivastava, S., Li, C. and Dunson, D. (2018). Scalable Bayes via barycenter in Wasserstein space. *Journal of Machine Learning Research* **19**, 1–35. MR3862415

Teh, Y. W., Thiery, H. A. and Vollmer, S. J. (2016). Consistency and fluctuations for stochastic gradient Langevin dynamics. *Journal of Machine Learning Research* **17**, 1–33. MR3482927

Wang, X. and Dunson, D. (2013). Parallelizing MCMC via Weierstrass sampler. arXiv preprint. Available at arXiv:1312.4605.

Wang, X., Guo, F., Heller, K. and Dunson, D. (2015). Parallelizing MCMC with random partition trees. *Advances in Neural Information Processing Systems* 451–459.

Brazilian Journal of Probability and Statistics
2023, Vol. 37, No. 4, 735–755
https://doi.org/10.1214/23-BJPS590

Recognition and variable selection in sparse spatial panel data models with fixed effects

Xuan Liu[1,a] **and Jianbao Chen**[2,b]

[1]*School of Mathematics and Information Science, Nanchang Normal University, Nanchang, 330032, P.R.China,* [a]*liuxuanyg@163.com*

[2]*School of Mathematics and Statistics, Fujian Normal University, Fuzhou, 350117, P.R.China,* [b]*jbjy2006@126.com*

Abstract. This paper is concerned with sparse spatial panel data models under fixed effects and increasing dimensions of covariates. We develop a nonconcave selection approach for spatial effects recognition and covariates selection of the models. Theoretical results show that the proposed method has the oracle property in the sense that the estimators have consistency, sparsity and asymptotic normality under suitable conditions. Furthermore, we present an coordinate descent algorithm to deal with the nonlinearity that arises in the optimization procedure. Numerical experiments show that the recognition and selection procedure can be used to select important covariates, identify spatial effects and estimate unknown parameters simultaneously. At the same time, the benefits of the proposed method is assessed by comparing different analyses of real spatial penal data.

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika* **60**, 255–265. MR0326953 https://doi.org/10.1093/biomet/60.2.255

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Netherlands: The Kluwer Academic Publishers.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica* **77**, 1229–1279. MR2547073 https://doi.org/10.3982/ECTA6135

Baltagi, B., Song, S. H. and Koh, W. (2003). Testing panel data regression models with spatial error correlation. *Journal of Econometrics* **117**, 123–150. MR2002299 https://doi.org/10.1016/S0304-4076(03)00120-9

Chu, T., Zhu, J. and Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics* **39**, 2607–2625. MR2906880 https://doi.org/10.1214/11-AOS919

Cliff, A. D. and Ord, J. K. (1973). *Spatial Autocorrelation*. London: Pion Ltd.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: John Wiley and Sons. MR1239641 https://doi.org/10.1002/9781119115151

Cuaresma, J. C., Doppelhofer, G. and Feldkircher, M. (2014). The determinants of economic growth in European regions. *Regional Studies* **48**, 44–67.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Mathematical Challenges Lecture* **1**, 1–32.

Elhorst, J. (2003). Specification and estimation of spatial panel data models. *International Regional Science Review* **26**, 244–268.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association* **99**, 710–723. MR2090905 https://doi.org/10.1198/016214504000001060

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467–5484. MR2849368 https://doi.org/10.1109/TIT.2011.2158486

Key words and phrases. Variable selection, penalty methods, spatial panel data model, fixed effects, oracle property.

Fang, Y., Xie, M., Bi, D., et al (2014). An exploratory analysis on the spatial correlation characteristics and influential factors of inbound tourism in China: A geographical weighted regression perspective. *Tourism Science* **28**, 22–35.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* **22**, 1947–1975. MR1329177 https://doi.org/10.1214/aos/1176325766

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *American Statistician* **64**, 325–334. MR2758564 https://doi.org/10.1198/tast.2010.10052

Hu, Y., Guo, C. and Lv, J. (2015). SCAD penalized rank regression with a diverging number of parameters. *Journal of Multivariate Analysis* **133**, 321–333. MR3282034 https://doi.org/10.1016/j.jmva.2014.09.014

Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics* **36**, 587–613. MR2396808 https://doi.org/10.1214/009053607000000875

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics* **1**, 799–821. MR0356373

Kapoor, M., Kelejian, H. H. and Prucha, I. R. (2007). Panel data models with spatially correlated error components. *Journal of Econometrics* **140**, 97–130. MR2395918 https://doi.org/10.1016/j.jeconom.2006.09.004

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* **17**, 99–121.

Kelejian, H. H. and Prucha, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* **104**, 219–257. MR1864417 https://doi.org/10.1016/S0304-4076(01)00064-1

Kock, A. B. and Callot, L. (2015). Oracle inequalities for high dimensional vector autoregressions. *Journal of Econometrics* **186**, 325–344. MR3343790 https://doi.org/10.1016/j.jeconom.2015.02.013

Lee, L. F. and Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. *Journal of Econometrics* **154**, 165–185. MR2558959 https://doi.org/10.1016/j.jeconom.2009.08.001

LeSage, J. P. and Fischer, M. (2008). Spatial growth regressions, model specification, estimation, and interpretation. *Spatial Economic Analysis* **3**, 275–304.

LeSage, J. P. and Pace, R. (2009). *Introduction to Spatial Econometrics*. Boca Raton: CRC Press. MR2485048 https://doi.org/10.1201/9781420064254

LeSage, J. P. and Parent, O. (2007). Bayesian model averaging for spatial econometric models. *Geographical Analysis* **39**, 241–267.

Li, Y. and Zeng, X. (2014). Research on variable selection method of penalized likelihood for panel data model. *Statistical Research* **31**, 83–89.

Liang, H. and Li, R. (2009). Variable selection for partially linear models with measurement errors. *Journal of the American Statistical Association* **104**, 234–248. MR2504375 https://doi.org/10.1198/jasa.2009.0127

Liu, X., Chen, J. and Cheng, S. (2018). A penalized quasi-maximum likelihood method for variable selection in the spatial autoregressive model. *Spatial Statistics* **25**, 86–104. MR3809257 https://doi.org/10.1016/j.spasta.2018.05.001

Mazumder, R., Friedman, J. H. and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* **106**, 1125–1138. MR2894769 https://doi.org/10.1198/jasa.2011.tm09738

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7**, 308–313. MR3363409 https://doi.org/10.1093/comjnl/7.4.308

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32. MR0025113 https://doi.org/10.2307/1914288

Pan, W. (2001). Akaikes information criterion in generalized estimating equations. *Biometrics* **57**, 120–125. MR1833297 https://doi.org/10.1111/j.0006-341X.2001.00120.x

Piribauer, P. (2016). Heterogeneity in spatial growth clusters. *Empirical Economics* **51**, 659–680.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464. MR0468014

Steel, M. F. (2020). Model averaging and its use in economics. *Journal of Economic Literature* **58**, 644–719.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological* **58**, 267–288. MR1379242

Wang, H., Li, R. and Tsai, C. L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568. MR2410008 https://doi.org/10.1093/biomet/asm053

Wang, H. and Zhu, J. (2009). Variable selection in spatial regression via penalized least squares. *Canadian Journal of Statistics* **37**, 607–624. MR2588952 https://doi.org/10.1002/cjs.10032

Wang, L. and Huang, L. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association* **103**, 1556–1569. MR2504204 https://doi.org/10.1198/016214508000000788

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society, Series B, Methodological* **71**, 177–190. MR2655529 https://doi.org/10.1111/j.1467-9868.2008.00679.x

Wang, L., Zhou, J. and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* **68**, 353–360. MR2959601 https://doi.org/10.1111/j.1541-0420.2011.01678.x

Xie, T., Cao, R. and Du, J. (2020). Variable selection for spatial autoregressive models with a diverging number of parameters. *Statistical Papers* **61**, 1125–1145. MR4096410 https://doi.org/10.1007/s00362-018-0984-2

Xu, D., Huang, Z. and Huang, R. (2019). The spatial effects of haze on tourism flows of Chinese cities: Empirical research based on the spatial panel econometric model. *Acta Geographica Sinica* **74**, 814–830.

Yang, W. and Zhang, H. (2015). A spatial econometric analysis of China's inbound tourism. *Tourism Tribune* **30**, 16–24.

Yu, J., Jong, R. and Lee, L. F. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *Journal of Econometrics* **146**, 118–134. MR2459648 https://doi.org/10.1016/j.jeconom.2008.08.002

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

Zhu, J., Huang, H. and Reyes, P. E. (2010). On selection of spatial linear models for lattice data. *Journal of the Royal Statistical Society, Series B, Methodological* **72**, 389–402. MR2758118 https://doi.org/10.1111/j.1467-9868.2010.00739.x

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429. MR2279469 https://doi.org/10.1198/016214506000000735

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B, Methodological* **67**, 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802

Brazilian Journal of Probability and Statistics
2023, Vol. 37, No. 4, 756–771
https://doi.org/10.1214/23-BJPS592

Multivariate zero-inflated Bell–Touchard distribution for multivariate counts: An application to COVID-related data

Artur J. Lemonte[a]

Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal/RN, Brazil,
[a]*arturlemonte@gmail.com*

Abstract. Multivariate count data with an excessive number of zeros occur commonly in practice. To deal with this kind of data, we introduce a simple and tractable parametric multivariate zero-inflated distribution, which corresponds to a multivariate extension of the univariate zero-inflated Bell–Touchard distribution. Various distributional properties are derived and, in particular, the marginal distributions are univariate zero-inflated Bell–Touchard distributions. The unknown parameters of the proposed multivariate zero-inflated distribution are estimated using the maximum likelihood estimation procedure. An empirical application that employs COVID-related real data is provided to illustrate the new multivariate zero-inflated distribution in practice, and comparison with some multivariate zero-inflated distributions is made.

References

Bell, E. T. (1934a). Exponential polynomials. *Annals of Mathematics* **35**, 258–277. MR1503161 https://doi.org/10.2307/1968431

Bell, E. T. (1934b). Exponential numbers. *The American Mathematical Monthly* **41**, 411–419. MR1523147 https://doi.org/10.2307/2300300

Cameron, A. C. and Trivedi, P. K. (2013). *Regression Analysis of Count Data*, 2nd ed. New York: Cambridge University Press. MR3155491 https://doi.org/10.1017/CBO9781139013567

Castellares, F., Ferrari, S. L. P. and Lemonte, A. J. (2018). On the Bell distribution and its associated regression model for count data. *Applied Mathematical Modelling* **56**, 172–185. MR3744937 https://doi.org/10.1016/j.apm.2017.12.014

Castellares, F., Lemonte, A. J. and Moreno-Arenas, G. (2020). On the two-parameter Bell–Touchard discrete distribution. *Communications in Statistics - Theory and Methods* **49**, 4834–4852. MR4135412 https://doi.org/10.1080/03610926.2019.1609515

Corless, R. M., Gonnet, G. H., Hare, D. E. G., Jeffrey, D. and Knuth, D. E. (1996). On the Lambert W function. *Advances in Computational Mathematics* **5**, 329–359. MR1414285 https://doi.org/10.1007/BF02124750

Johnson, N. and Kotz, S. (1969). *Distributions in Statistics: Discrete Distributions*. Boston: Houghton Mifflim. MR0268996

Kollo, T. (2008). Multivariate skewness and kurtosis measures with an application in ICA. *Journal of Multivariate Analysis* **99**, 2328–2338. MR2463392 https://doi.org/10.1016/j.jmva.2008.02.033

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34**, 1–14.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. New York: Springer. MR1639875

Lemonte, A. J. (2022). On the mean-parameterized Bell–Touchard regression model for count data. *Applied Mathematical Modelling* **105**, 1–16. MR4360238 https://doi.org/10.1016/j.apm.2021.12.035

Lemonte, A. J. (2023). Supplement to "Multivariate zero-inflated Bell–Touchard distribution for multivariate counts: An application to COVID-related data." https://doi.org/10.1214/23-BJPS592SUPP

Li, C., Lu, J., Park, J., Kim, K., Brinkley, P. and Peterson, J. (1999). Multivariate zero-inflated Poisson models and their applications. *Technometrics* **41**, 29–38.

Liu, Y. and Tian, G. L. (2015). Type I multivariate zero-inflated Poisson distribution with applications. *Computational Statistics & Data Analysis* **83**, 200–222. MR3281806 https://doi.org/10.1016/j.csda.2014.10.010

Key words and phrases. Bell distribution, Bell–Touchard distribution, Excess zeros, Poisson distribution.

Loperfido, N. (2015). Singular value decomposition of the third multivariate moment. *Linear Algebra and Its Applications* **473**, 202–216. MR3338332 https://doi.org/10.1016/j.laa.2014.05.043

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika* **36**, 519–530. MR0397994 https://doi.org/10.1093/biomet/57.3.519

Ministério da Saúde (2020). Ministério da Saúde: Coronavírus Brasil. Available at https://covid.saude.gov.br/.

Móri, T. F., Rohatgi, V. K. and Székely, G. J. (1993). On multivariate skewness and kurtosis. *Theory of Probability and Its Applications* **38**, 547–551. MR1404679 https://doi.org/10.1137/1138055

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

TAB-NET (2020). Ministério da Saúde: Departamento de Informática do SUS-DATASUS. Informações de Saúde, Epidemiológicas e Morbidade – banco de dados. Available at http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sih/cnv/niSP.def.

Tian, G. L., Liu, Y., Tang, M. L. and Jiang, X. (2018). Type I multivariate zero-truncated/adjusted Poisson distributions with applications. *Journal of Computational and Applied Mathematics* **344**, 132–153. MR3825508 https://doi.org/10.1016/j.cam.2018.05.014

Touchard, J. (1933). Propriétés arihtmétiques de certains nombres récurrents. *Annales de la Société Scientifique de Bruxelles* **53**, 21–31.

Zhang, C., Tian, G. L., Yuen, K. C., Wud, Q. and Li, T. (2020). Multivariate zero-and-one inflated Poisson model with applications. *Journal of Computational and Applied Mathematics* **365**, 112356. MR3987478 https://doi.org/10.1016/j.cam.2019.112356

AIMS AND SCOPE

The Brazilian Journal of Probability and Statistics aims to publish high quality research papers in applied probability, applied statistics, computational statistics, mathematical statistics, probability theory and stochastic processes.

More specifically, the following types of contributions will be considered:

(i) Original articles dealing with methodological developments, comparison of competing techniques or their computational aspects;
(ii) Original articles developing theoretical results;
(iii) Articles that contain novel applications of existing methodologies to practical problems. For these papers the focus is in the importance and originality of the applied problem, as well as, applications of the best available methodologies to solve it.
(iv) Survey articles containing a thorough coverage of topics of broad interest to probability and statistics. The journal will occasionally publish book reviews, invited papers and essays on the teaching of statistics.

GENERAL INFORMATION

Journal Homepage: https://www.imstat.org/journals-and-publications/brazilian-journal-of-probability-and-statistics/

Submissions: Manuscripts for *Brazilian Journal of Probability and Statistics* should be submitted online. Authors may access the Electronic Journals Management System (EJMS) at https://www.e-publications.org/ims/submission/.

Preparation of Manuscripts: https://imstat.org/journals-and-publications/brazilian-journal-of-probability-and-statistics/brazilian-journal-of-probability-and-statistics-manuscript-submission/

Correspondence. Mail concerning membership, subscriptions, nonreceipt claims, copyright permissions, advertising or back issues should be sent to the IMS Dues and Subscription Office, PO Box 729, Middletown, Maryland 21769, USA. Mail concerning submissions or editorial content should be sent to the Editor at nancy@ime.unicamp.br. Mail concerning the production of this journal should be sent to: Geri Mattson at bjps@mattsonpublishing.com.

Individual and Organizational Memberships:
https://www.imstat.org/individual-membership/

Individual and General Subscriptions:
https://imstat.org/journals-and-publications/subscriptions-and-orders/

Electronic Access to IMS Journals:
https://www.imstat.org/journals-and-publications/electronic-access/

The *Brazilian Journal of Probability and Statistics* is an IMS supported journal:
https://www.imstat.org/journals-and-publications/ims-supported-journals/