Pla de treball

Candidat: Víctor Peña

Categoria: Professor lector

Referència del concurs: LE-128/715

Resolució: núm. 103 SAiP PDI 2023-3859/5

Contents

1 Teaching statement

1.1 Introduction

At the moment of submitting this statement, I have over eight years of teaching experience. I started teaching when I was a Ph.D. student at Duke. Later, as an assistant professor at the City University of New York, I had to adapt to teaching business school students. During the last two years, I have been teaching at Universitat Politècnica de Catalunya as a Maria Zambrano fellow.

I have taught a wide array of courses at various levels. On the most basic end, I instructed an introductory statistics course for business students and a course on quantitative literacy for incoming freshmen. On the more advanced end, I taught generalized linear models and multivariate analysis for master's students in Statistics (a comprehensive list of the courses I have taught can be found in my CV). I also have experience developing courses from scratch: during my time at the City University of New York, I created two `R` courses, one for undergraduates and another for graduate students.

As an instructor, I strive to create an environment where students are engaged and participative. In my experience, achieving this goal is challenging unless students perceive that the instructor is enthusiastic about teaching and willing to work as hard as they do. For that reason, I find it useful to prepare my teaching materials (notes and problem sets) rather than relying on borrowed slides or closely following a textbook. Logistically, I proceed as follows: (1) I upload an incomplete handout or set of slides the day before the lecture so that students can have an initial idea of what is going to be covered, (2) I go through the handout/slides in the lecture, adding some examples that are not there, (3) after the lecture, I upload a complete handout with solutions and new exercises interspersed so that students can consolidate the concepts. In my experience, this systematic approach works well because students know exactly what to expect from the course.

1.2 Courses in job description

In the job description, it is stated that the successful candidate will be expected to teach the following courses: Bayesian methods (undergraduate and graduate), linear and generalized linear models (graduate), and Probability and Statistics 2 (undergraduate).

I have experience teaching Bayesian methods as a Ph.D. student at Duke. While I was there, most Statistics courses, especially the graduate courses, were taught from a Bayesian perspective. While at Duke, I was a teaching assistant for Bayesian decision theory, regression modeling, and predictive modeling. More recently, I have been teaching a course on Bayesian methods for undergraduate students on data science at the Universitat Oberta de Catalunya. Apart from my teaching experience, my research is focused on Bayesian methodology.

I am currently teaching linear and generalized linear models in the master's program in Statistics and Operations Research, and so I did last year. My experience so far has been positive and I would be happy to keep teaching the course. Similarly, I am currently teaching Probability and Statistics 2, which is an undergraduate course for data science students, jointly with Marta Pérez. The course covers basic statistical inference (confidence intervals and hypothesis tests), linear models, and generalized linear models. This is the first year the course is offered in this format. The students are highly intelligent and motivated, so it is a pleasure to teach this course.

2 Research statement

2.1 Background

My research background is in Bayesian theory and methods. I did my Ph.D. under the supervision of James O. Berger. With him, I worked on projects related variable selection (Peña and Berger, 2020), hypothesis testing (Mulder et al., 2021), and foundations of statistical science (more precisely, on the likelihood principle; Peña and Berger (2017)).

After graduating, I explored other topics, such as non-parametric inference for big data (Barrientos and Peña, 2020), multivariate time-series data (Peña and Irie, 2022), and non-parametric inference with shape constraints (Jauch et al., 2023).

With A.F. Barrientos (at Florida State University), I have written three articles on methods for analyzing confidential data. In our first article, we extended the randomized response mechanism to perform differentially private hypothesis testing (Peña and Barrientos, 2023a). In our second article, we developed methods for hypothesis testing, model-averaging, and selection for normal linear models under differential privacy constraints (Peña and Barrientos, 2023b). And recently, we submitted a paper on how to analyze confidential compositional data (Guo et al. (2023), also co-authored with Qi Guo, who is A.F. Barrientos' former student).

2.2 Active projects

2.2.1 New representations for the Generalized inverse Gaussian

The generalized inverse Gaussian, which we denote $\text{GIG}(p, a, b)$, is a three-parameter distribution that is useful for modeling continuous non-negative data. It is a rich family of distributions that encompasses the gamma distribution (if $b = 0$), the inverse gamma distribution (if $a = 0$), and the inverse Gaussian distribution (if $p = -1/2$).

The generalized inverse Gaussian was first proposed by Halphen in the 1940s for analyzing hydrological data (Halphen, 1941; Perreault et al., 1999). It regained popularity

in the mid to late 1970s, when a series of articles studied its properties in detail (see, for example, Barndorff-Nielsen and Halgreen (1977), Blæsild (1978), and Halgreen (1979)). These developments were compiled in the monograph Jørgensen (1982). The properties of the GIG(p, a, b) continued to be studied after the publication of Jørgensen (1982), a notable example being Letac and Seshadri (1983), which represented the GIG(p, a, b) in terms of continued fractions. For an up-to-date account of known properties and characterizations of the GIG(p, a, b), we refer the reader to Koudou and Ley (2014).

The generalized inverse Gaussian has been used in a wide array of applications. For example, it has been used for modeling circulatory transit times in pharmacokinetics (Weiss, 1984), neural activity (Iyengar and Liao, 1997), particle size distributions (Alexandrov and Lacis, 2000), and hydrological extremes (Chebana et al., 2010).

The generalized inverse Gaussian is often used as a latent model for rates in count data. In fact, one of its first applications is Good (1953), an article concerned with estimating population frequencies. A popular model for count data that uses the GIG(p, a, b) as a building block is the Sichel distribution (Sichel, 1971). The Sichel distribution arises by letting the data be distributed as Poisson(λ) with $\lambda \sim$ GIG(p, a, b), and then marginalizing out λ. This distribution has been used for analyzing sentence length (Sichel, 1974), word frequencies (Sichel, 1975), and literary style (Holmes, 1985), among other applications.

The parameters of the GIG(p, a, b) are $p \in \mathbb{R}$, $a > 0$, and $b > 0$, and its probability density function (PDF) is

$$f_X(x) = \frac{(a/b)^{p/2}}{2K_p(\sqrt{ab})}\, x^{p-1} \exp\left(-\frac{(ax + b/x)}{2}\right) \mathbf{1}(x > 0),$$

where $K_p(\cdot)$ is the modified Bessel function of the second kind (Abramowitz and Stegun, 1968). Its moments involve Bessel functions, but they are all known (see, for example, Jørgensen (1982)).

Random number generation and inference for the GIG(p, a, b) are challenging due to the presence of a Bessel function in the denominator of the PDF. Several articles have been written on generating random draws from the GIG(p, a, b), such as Dagpunar

(1989), Leydold and Hörmann (2011), Hörmann and Leydold (2014), Devroye (2014), and Zhang and Reiter (2022). Recently, Willmot and Woo (2022) derived expressions for the CDF of the generalized inverse Gaussian for half-integer p.

In our work, we describe two novel mixture representations for the generalized inverse Gaussian. The first representation expresses the GIG(p, a, b) as a continuous mixture of inverse Gaussians, whereas the second one expresses the GIG(p, a, b) as a continuous mixture of truncated exponentials.

We show that these representations are useful for random number generation. With the first representation, we found a geometrically ergodic Gibbs sampler whose stationary distribution is GIG(p, a, b). With the second representation, we derived a recursive algorithm to obtain exact independent draws for the GIG(p, a, b) for half-integer p (that is, $p \in \{..., -3/2, -1/2, 1/2, 3/2, ...\}$). The second representation also allowed us to write a recursive algorithm for evaluating the cumulative distribution function (CDF) of the GIG(p, a, b) for half-integer p. The algorithms are simple and can be easily implemented in standard programming languages.

This is joint work with Michael Jauch, at Florida State.

2.2.2 Bayesian analysis of fractional factorial designs

Fractional factorial designs have been widely studied and applied in practice (see, for example, Box et al. (2005)). Unfortunately, the existing Bayesian methods for the analysis of fractional factorial designs are not up-to-date with recent developments in the area of objective (and subjective) prior specification.

In our work, we propose new priors on the model space that respect hierarchy constraints and provide effective control for false positives, essentially extending the work of Scott et al. (2010). We provide a default prior that is fully automatic and a subjective prior that uses partial information on the number of expected active factors and the relative odds that each factor is active. After eliciting this partial subjective information, we propose finding the maximum entropy prior that satisfies the constraints, which is a convex optimization problem that can be handled by open-source solvers

such as SDPT3 (Toh et al., 2012) and SeDuMi (Sturm, 1999).

One of the key difference between our work and the existing literature is our treatment of models with aliased (identical) columns and models that have more variables than observations. These models are singular, in the sense that the design matrices are rank-deficient.

In the existing literature, models that contain more variables than observations are assigned probability zero a priori. We believe that this is not a sensible approach: the reason that these models are not included is not that those models are, in some sense, *impossible*. The exclusion of these models leads to an underestimation of posterior uncertainty, since the cardinality of the model space is made artificially smaller than it actually is.

We propose including all the singular models in the model space. Our priors on the regression coefficients are singular normal extensions of mixtures of g-priors. This extension does not lead to any complications analytically, algebraically, or computationally. The Bayes factors we propose satisfy three appealing properties:

1. Invariance with respect to the choice of generalized inverse. This is important because generalized inverses are not unique.

2. If the column spaces of the models φ and η are the same, the Bayes factor $B_{\varphi\eta}$ is equal to 1. This would not be satisfied if we put independent priors on the regression coefficients.

3. Predictive matching (Bayarri et al., 2012), which is a formalization of an idea first proposed in Jeffreys (1939). This property would not be satisfied by an independent prior specification for the regression coefficients.

This is joint work with Gonzalo Garcia-Donato, at the Universidad de Castilla la Mancha.

2.3 Future Work, Collaborations, and Grants

After I submit my ongoing work, I have a clear pipeline of projects I will pursue. Below, I list three projects in my pipeline:

- Michael Jauch (at Florida State) and I discovered a novel approach to deriving Stein's Unbiased Risk Estimate (SURE) for Bayesian models. Leveraging early work by Larry Brown and James O. Berger (which connect the divergence to functionals of the posterior distribution), we will be able to find SURE for complex models where estimation of marginal likelihoods is challenging.

- Kaoru Irie (at the University of Tokyo) and I will be working on dynamic models for matrix-variate data, leveraging recent computational developments for the multivariate generalized inverse Gaussian distribution (Hamura et al., 2023). In August 2024, I will be visiting the Economics department in the University of Tokyo to work on this project.

- Alexander Ly (at the University of Amsterdam) and I will be working on the asymptotic analysis of ANOVA models when the groups are highly unbalanced. The behavior of Bayes factors under common prior specifications is surprising (in a negative way).

I have also started collaborations with colleagues within the department:

- I am collaborating with Josep Ginebra and Xavier Puig (at UPC) on a project for modeling mortality rates in Catalonia using data provided by the "Departament de Salut de la Generalitat de Catalunya."

- After finishing the project on Bayesian analysis of factorial designs, I will be working with Pere Grima on a Bayesian approach to estimating effects in fractional factorial designs when there is partial prior information on the sign or magnitude of the effects.

If I stay within the department, I will make an effort to communicate with others and initiate further collaborations. I am mostly interested in theoretical and methodological development, and I could collaborate with professors like Josep Ginebra or Marta Pérez on topics like foundations of statistical inference or statistical theory for discrete models.

Currently, I am a collaborator on the grant "Métodos Bayesianos para la selección de variables en problemas de alta dimensionalidad y con datos perdidos," which was recently awarded by the Spanish government. In this project, I will be collaborating with Gonzalo Garcia-Donato (Universidad de Castilla y la Mancha), Maria Eugenia Castellanos (Universidad Rey Juan Carlos), Alicia Quirós (Universidad de León), Stefano Cabras (Universidad Carlos III), and Anabel Forte (Universitat de València). My contribution on this grant will be focused on theoretical development. I expect to collaborate with this group consistently throughout my career.

References

Abramowitz, M. and I. A. Stegun (1968). Handbook of mathematical functions with formulas, graphs and mathematical tables. *New York: Dover*.

Alexandrov, M. D. and A. A. Lacis (2000). A new three-parameter cloud/aerosol particle size distribution based on the generalized inverse Gaussian density function. *Applied Mathematics and Computation 116*(1-2), 153–165.

Barndorff-Nielsen, O. and C. Halgreen (1977). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 38*(4), 309–311.

Barrientos, A. F. and V. Peña (2020). Bayesian bootstraps for massive data. *Bayesian analysis*.

Bayarri, M., J. Berger, A. Forte, G. García-Donato, et al. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics 40*(3), 1550–1577.

Blæsild, P. (1978). *The shape of the generalized inverse Gaussian and hyperbolic distributions*. Department of Theoretical Statistics, Aarhus University.

Box, G. E., W. H. Hunter, S. Hunter, et al. (2005). *Statistics for experimenters (2nd Edition)*. Wiley.

Chebana, F., S. E. Adlouni, and B. Bobée (2010). Mixed estimation methods for Halphen distributions with applications in extreme hydrologic events. *Stochastic Environmental Research and Risk Assessment 24*, 359–376.

Dagpunar, J. (1989). An easily implemented generalised inverse Gaussian generator. *Communications in Statistics-Simulation and Computation 18*(2), 703–710.

Devroye, L. (2014). Random variate generation for the generalized inverse Gaussian distribution. *Statistics and Computing 24*(2), 239–246.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40*(3-4), 237–264.

Guo, Q., A. F. Barrientos, and V. Peña (2023). Differentially private methods for compositional data. *Submitted to JCGS*.

Halgreen, C. (1979). Self-decomposability of the generalized inverse Gaussian and hyperbolic distributions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47*(1), 13–17.

Halphen, E. (1941). Sur un nouveau type de courbe de fréquence. *Comptes Rendus de l'Académie des Sciences 213*, 633–635.

Hamura, Y., K. Irie, and S. Sugasawa (2023). Gibbs sampler for matrix generalized inverse gaussian distributions. *arXiv preprint arXiv:2302.09707*.

Holmes, D. I. (1985). The analysis of literary style—a review. *Journal of the Royal Statistical Society: Series A (General) 148*(4), 328–341.

Hörmann, W. and J. Leydold (2014). Generating generalized inverse Gaussian random variates. *Statistics and Computing 24*, 547–557.

Iyengar, S. and Q. Liao (1997). Modeling neural activity using the generalized inverse Gaussian distribution. *Biological cybernetics 77*(4), 289–295.

Jauch, M., A. F. Barrientos, V. Peña, and D. S. Matteson (2023). Mixture representations for likelihood ratio ordered distributions. *arXiv preprint arXiv:2110.04852 (Submitted to JASA)*.

Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.

Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*, Volume 9. Springer Science & Business Media.

Koudou, A. E. and C. Ley (2014). Characterizations of GIG laws: A survey. *Probability Surveys 11*, 161–176.

Letac, G. and V. Seshadri (1983). A characterization of the generalized inverse Gaussian distribution by continued fractions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 62*(4), 485–489.

Leydold, J. and W. Hörmann (2011). Generating generalized inverse Gaussian random variates by fast inversion. *Computational statistics & data analysis 55*(1), 213–217.

Mulder, J., J. O. Berger, V. Peña, and M. Bayarri (2021). On the prevalence of information inconsistency in normal linear models. *Test 30*, 103–132.

Peña, V. and J. O. Berger (2017). A note on recent criticisms to Birnbaum's theorem. *arXiv preprint arXiv:1711.08093*.

Peña, V. and J. O. Berger (2020). Constrained empirical Bayes priors on regression coefficients. *Bayesian Analysis*.

Peña, V. and A. F. Barrientos (2023a). Differentially private hypothesis testing with the subsampled and aggregated randomized response mechanism. *Statistica Sinica*.

Peña, V. and A. F. Barrientos (2023b). Differentially private methods for managing model uncertainty in linear regression models. *JMLR (accepted)*.

Peña, V. and K. Irie (2022). On the relationship between uhlig extended and beta-bartlett processes. *Journal of Time Series Analysis 43*(1), 147–153.

Perreault, L., B. Bobée, and P. F. Rasmussen (1999). Halphen distribution system. I: Mathematical and statistical properties. *Journal of Hydrologic Engineering 4*(3), 189–199.

Scott, J. G., J. O. Berger, et al. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics 38*(5), 2587–2619.

Sichel, H. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In *Proceedings of the third symposium on mathematical statistics*. SACSIR.

Sichel, H. S. (1974). On a distribution representing sentence-length in written prose. *Journal of the Royal Statistical Society: Series A (General) 137*(1), 25–34.

Sichel, H. S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association 70*(351a), 542–547.

Sturm, J. F. (1999). Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software 11*(1-4), 625–653.

Toh, K.-C., M. J. Todd, and R. H. Tütüncü (2012). On the implementation and usage of sdpt3–a matlab software package for semidefinite-quadratic-linear programming, version 4.0. In *Handbook on semidefinite, conic and polynomial optimization*, pp. 715–754. Springer.

Weiss, M. (1984). A note on the role of generalized inverse Gaussian distributions of circulatory transit times in pharmacokinetics. *Journal of mathematical biology 20*, 95–102.

Willmot, G. E. and J.-K. Woo (2022). Remarks on a generalized inverse Gaussian type integral with applications. *Applied Mathematics and Computation 430*.

Zhang, X. and J. P. Reiter (2022). A generator for generalized inverse Gaussian distributions. *arXiv preprint arXiv:2211.13049*.