The New England Journal of Statistics in Data Science Volume 0, 1–4 (2026)
XML]Editorial
DOI: https://doi.org/10.51387/26-NEJSDS41EDI

Editorial
Pushing the Boundaries of Data Science Through Statistical Modelling and Inference: Highlights from a Special Issue

Jyotishka DATTA and Feng GUO

1. INTRODUCTION

This special issue grew out of a conference held on July 13–14, 2023, at Virginia Tech in Blacksburg, Virginia, organized to celebrate Professor Dipak Kumar Dey's 70th birthday. The meeting brought together Professor Dey's former students, collaborators, friends, and mentors to honor his sustained and wide-ranging contributions to statistics. Reflecting the breadth of his influence, the program spanned modern Bayesian modeling and inference, decision-theoretic foundations, and emerging data-science methodology, together with a wide range of interdisciplinary applications.

Professor Dipak K. Dey: Brief Biography

Professor Dipak K. Dey is the Board of Trustees Distinguished Professor in the Department of Statistics at the University of Connecticut. He is widely recognised for his foundational and pioneering contributions to Bayesian analysis, decision science, and model selection. He received his bachelor's and master's degrees in Statistics from the Indian Statistical Institute and earned his Ph.D. in Statistics from Purdue University under the supervision of Professor Jim Berger. He has authored more than 260 research articles and has written or edited more than 10 books and volumes, with contributions that have shaped both statistical theory and practice. Before joining the University of Connecticut in 1985, he held academic appointments at Stanford University, the University of Kentucky, and Texas Tech University, and has also held visiting positions at universities and research institutions around the world.

Professor Dey is a Fellow of the American Association for the Advancement of Science, the American Statistical Association, the Institute of Mathematical Statistics, the International Society for Bayesian Analysis, and the International Statistical Institute, and he has received numerous honors recognizing his scholarship and service. His work has had a broad impact across a wide range of applied domains, including biometry, bioinformatics, genetics, econometrics, environmental science, and social sciences. He has supervised over 40 Ph.D. students, delivered more than 200 invited professional talks, and served the profession through editorial leadership, including as Associate Editor of the *Journal of the American Statistical Association* (1997–1999) and the *Journal of Statistical Planning and Inference* (2001–2003), and as Editor-in-Chief of *Sankhyā* since 2016.

At the 38th New England Statistics Symposium on June 3, 2025, the invited session "Pushing the Boundaries of Data Science through Statistical Modelling and Inference: Highlights from a Special Issue" showcased several of the contributions featured in this volume. Indranil Sahoo and Jyotishka Datta presented their respective work, and Dr. Colin Wu served as a discussant, offering a thoughtful synthesis of the themes and constructive commentary on both papers. Taken together, the articles in this special issue reflect the wide scope of contemporary statistical research, ranging from Bayesian computation, spatial modelling, and robust inference to extreme-value methods, network analysis, and applications in public health, environmental science, and biomedical studies.

We provide a brief summary of each of the papers in this special issue below. Many of the contributions reflect themes central to Professor Dey's own scholarship: Bayesian hierarchical modeling and computation remain an overarching theme in this issue; several papers take up questions of model selection and decision-theoretic inference under uncertainty; and the applications span the environmental, biomedical, and social-science domains.

2. RESEARCH ARTICLES

1. Chen Qian, Chenyang Tao, Jingbin Xu, Feng Guo. *Variational Inference of Extremes for Rare Event Modeling: Theory and Applications.*
   Rare events such as fraud, extreme weather, traffic accidents, and uncommon diseases pose a fundamental challenge to predictive modeling: the minority class is precisely what practitioners most need to model well, yet standard approaches are ill-suited to its scarcity. Rather than rebalancing training data through

sampling heuristics, Qian et al. [6] adopt a generative Bayesian perspective in which predictors serve as stochastic proxies for latent causes, and extremes arise from the exceedance of those latent quantities. By placing a Gumbel prior on the latent space and fitting the model within a variational inference framework, they obtain a disentangled additive monotonic neural architecture that simultaneously addresses tail behavior, uncertainty quantification, and interpretability. Empirical comparisons on several real-world benchmarks demonstrate that the method achieves improved generalization on minority classes relative to state-of-the-art alternatives.

2. Indranil Sahoo, Suman Majumder, Arnab Hazra, Ana Rappold, Dipankar Bandyopadhyay.
Computationally Scalable Bayesian SPDE Modeling for Censored Spatial Responses.
In this paper, Sahoo et al. [7] address the dual computational challenges of large-scale geostatistical data subject to left-censoring, as arises in groundwater pollutant monitoring. The authors approximate a Matern Gaussian process via a Gaussian Markov random field (GMRF) derived from a stochastic partial differential equation (SPDE) discretized on a triangulated mesh, yielding a sparse precision matrix that enables scalable MCMC. A GMRF-based nugget (measurement error) term is incorporated into the hierarchical model, which induces conditional independence across sites given the latent field and reduces censored-likelihood evaluation to a product of univariate truncated normal probabilities rather than an intractable high-dimensional integral. Applied to PFOS concentration data from 24,959 sites across California, nearly half of which are censored, the model produces full posterior predictive surfaces and uncertainty quantification in roughly an hour per chain, demonstrating its scalability for large-scale environmental monitoring data.

3. Matthias Reginald Fengler, Bruno Jäger, Ostap Okhrin.
Locally Adaptive Modeling of Unconditional Heteroskedasticity.
Fengler et al. [2] show that local change point detection in unconditional variance can be carried out using a generalized likelihood ratio test calibrated by a multiplier bootstrap. The authors introduce a multiplicative bias correction as an alternative to the existing additive correction, which yields closed-form bootstrap variance estimators (rather than requiring numerical root-finding), offers a clearer interpretation as a bias-corrected pooled variance, and substantially reduces computational cost. They develop tests for homogeneity in variance alone and for complete homogeneity in mean and variance, and embed these in a sequential local change point detection algorithm that mimics a data-driven rolling window estimator. Simulations show that the multiplicative correction performs comparably to the additive one in size and power, with some advantages in small samples and in detecting unconditional variance breaks under serial dependence and conditional heteroskedasticity (GARCH) settings, where the locally adaptive procedure is competitive with fixed-window rolling estimators. Applied to U.S. inflation and industrial production growth rates and to Bitcoin returns, the method identifies breaks in variance consistent with known macroeconomic episodes, such as the Great Moderation, as well as with periods of heightened cryptocurrency price volatility.

4. Brisilda Ndreka, Dipak K. Dey, Victor H. Lachos.
Bayesian Estimation of Contagion Effect: An Application of Friendship Networks and Alcohol Behavior.
Ndreka et al. [4] propose a fully Bayesian latent-space model that jointly estimates social contagion effects and homophily, addressing the difficulty of separating peer influence from selection in observational social network data. A longitudinal latent selection model captures tie formation as a function of time-invariant latent positions, and an associated dynamic linear influence model incorporates these estimated positions as covariates to control for latent homophily when estimating the network exposure (contagion) effect on behavior. Estimation is carried out via Hamiltonian Monte Carlo using `Stan`, and simulation studies show that omitting the latent trait biases the contagion effect estimate, while increasing network size, number of time points, or homophily level affects bias and coverage in predictable ways; the proposed approach also outperforms a two-step frequentist latent-space adjustment in both bias and coverage. Applied to three years of friendship network data from the Teenage Friends and Lifestyle Study, the model detects a statistically significant contagion effect for adolescent alcohol use alongside effects of prior own behavior, cannabis use, and sports participation.

5. Weiji Su, Xia Wang, Pedro Miranda-Afonso, Eleni-Rosalina Andrinopoulou, Rhonda D. Szczesniak.
Modeling Disease Progression in the Presence of an Outcome-Dependent Visiting Process with Application to Cystic Fibrosis Clinical Data.
Su et al. [8] propose a Bayesian shared-parameter joint model linking a mixed-effects model for a longitudinal marker with a Weibull proportional hazards model for the gap times between clinical visits, where a lognormal frailty term in the hazard submodel is shared with the longitudinal submodel through a hazard ratio of having a visit. Three random-effects structures (random intercept; random intercept and slope; random intercept with a Gaussian process over time) are compared using LPML, DIC, WAIC, and predictive accuracy, alongside single longitudinal models that instead adjust for visit frequency as a time-varying covariate. The methodology is then applied to lung function (FEV1) data from 197 pediatric cystic fibrosis patients, where the joint

model with a Gaussian process term fit best overall, and patients with lower lung function were estimated to visit more frequently, consistent with an outcome-dependent visiting process. Simulation studies further show that incorporating a time-dependent Gaussian process improves model fit more than adding the survival submodel via joint modeling, while a random-intercept-only model exhibits the largest bias when visiting is outcome-dependent.

6. Jyotishka Datta, Nicholas Polson.
Inverse Probability Weighting: From Survey Sampling to Evidence Estimation
The two popular inverse probability weighting estimators, Horvitz-Thompson and Hájek, are utilized in a remarkably broad range of statistical tasks, from survey sampling to causal inference and importance sampling in Bayesian computation. Yet these estimators are susceptible to paradoxical behavior in certain pathological cases identified by Basu (1988) and Wasserman (2004). Datta and Polson [1] revisit these 'weak paradoxes' and investigate two Bayesian remedies: a nonparametric 'Bayesian sieve' based on binning and smoothing, and a conjugate hierarchical model that borrows strength across observations via exchangeability. They also establish posterior consistency for both estimators under the missing-completely-at-random assumption. The paper draws connections to conditional Monte Carlo, trapezoidal integration, vertical likelihood, and average treatment effect estimation, demonstrating that improved inverse-probability weighting has unifying implications across a wide swath of statistical methodology.
7. Sai Ma, Jun Yan, Xuebin Zhang.
Extreme Value Modeling with Generalized Pareto Distributions for Rounded Data.
Rounding in environmental measurements such as precipitation records induces interval censoring of threshold exceedances, leading to biased parameter estimation and invalid inference for the generalized Pareto distribution (GPD). Furthermore, when rounding is ignored, goodness-of-fit tests suffer from size distortions. In this paper, Ma, Yan, and Zhang [3] construct the correct likelihood by treating each rounded observation as an interval-censored event, yielding a maximum likelihood estimator (MLE-IC) that is asymptotically unbiased and efficient for shape parameters, in contrast to the naive estimator (MLE-N) whose bias grows with the rounding level δ relative to the scale parameter σ. Classic goodness-of-fit statistics (Kolmogorov–Smirnov, Cramér–von Mises, and Anderson–Darling) are adapted to the resulting discrete distribution with p-values obtained via parametric bootstrap, with the Anderson–Darling test exhibiting the highest power across simulation settings. Applied to automated threshold selection for daily winter precipitation at 18 eastern Washington stations, MLE-IC selected a valid threshold at 9 stations compared to only 2 under MLE-N, yielding substantially narrower return-level confidence intervals where thresholds were identified by both methods. This paper fills an important gap in the literature: data rounding is pervasive in practice – instruments record to the nearest millimeter, degree, or kilogram – yet its effects on extreme-value inference beyond point estimation have received scant attention.
8. Eric Odoom, Xia Wang.
Consistent and Scalable Variable Selection with Robust Link Functions.
Odom and Wang [5] develop a hierarchical skinny Gibbs sampler for high-dimensional binary regression with a t-link function (HSGT), which generalizes the probit and logit links and offers robustness to outlying covariates and responses through the degrees-of-freedom parameter. Building on the skinny Gibbs framework and a continuous spike-and-slab prior, the authors establish posterior ratio consistency and strong selection consistency for the t-link model under standard high-dimensional conditions, extending prior results for the logit and probit links. Simulation studies show that the t-link models, particularly with small degrees of freedom, maintain variable selection accuracy and prediction performance under bad leverage and non-leverage outliers, whereas the probit link deteriorates noticeably, and the skinny Gibbs algorithm scales linearly in the number of covariates compared to the quadratic cost of the exact Gibbs sampler. The methodology developed here is applied to a mouse gene expression dataset, where the HSGT and exact Gibbs t-link models achieved the lowest cross-validated prediction error and the highest classification accuracy and AUC relative to logit and probit counterparts.

ACKNOWLEDGEMENTS

We thank the organizing committee of the conference held at Virginia Tech, and the participants of the invited session at the 38th New England Statistics Symposium for their contributions to the themes of this special issue. We also thank the Editors-in-Chief of *The New England Journal of Statistics in Data Science* for their support throughout the preparation of this issue, and the anonymous reviewers for their careful and constructive evaluations of the submitted manuscripts. Above all, we thank Professor Dipak K. Dey for his decades of mentorship and contributions to our field, which this special issue is honored to celebrate.

REFERENCES

[1] Datta, J. and Polson, N. (2026). Inverse Probability Weighting: From Survey Sampling to Evidence Estimation. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/26-NEJSDS100.

[2] Fengler, M. R., Jäger, B. and Okhrin, O. (2025). Locally Adaptive Modeling of Unconditional Heteroskedasticity. *The New England Journal of Statistics in Data Science* 1–26. https://doi.org/10.51387/25-NEJSDS91.

[3] Ma, S., Yan, J. and Zhang, X. (2026). Extreme Value Modeling with Generalized Pareto Distributions for Rounded Data. *The New England Journal of Statistics in Data Science* 1–12. https://doi.org/10.51387/26-NEJSDS101.

[4] Ndreka, B., Dey, D. K. and Lachos, V. H. (2026). Bayesian Estimation of Contagion Effect: An Application of Friendship Networks and Alcohol Behavior. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/26-NEJSDS94.

[5] Odoom, E. and Wang, X. (2026). Consistent and Scalable Variable Selection with Robust Link Functions. *The New England Journal of Statistics in Data Science* 1–15. https://doi.org/10.51387/26-NEJSDS102.

[6] Qian, C., Tao, C., Xu, J. and Guo, F. (2025). Variational Inference of Extremes for Rare Event Modeling: Theory and Applications. *The New England Journal of Statistics in Data Science*.

[7] Sahoo, I., Majumder, S., Hazra, A., Rappold, A. G. and Bandyopadhyay, D. (2025). Computationally Scalable Bayesian SPDE Modeling for Censored Spatial Responses. *The New England Journal of Statistics in Data Science* 1–15. https://doi.org/10.51387/25-NEJSDS78.

[8] Su, W., Wang, X., Miranda-Afonso, P., Andrinopoulou, E.-R. and Szczesniak, R. D. (2026). Modeling Disease Progression in the Presence of an Outcome-Dependent Visiting Process with Application to Cystic Fibrosis Clinical Data. *The New England Journal of Statistics in Data Science* 1–13. https://doi.org/10.51387/26-NEJSDS99.

Jyotishka Datta. Virginia Tech, United States.
E-mail address: jyotishka@vt.edu

Feng Guo. Virginia Tech, United States.
E-mail address: feguo@vt.edu