Objective Bayes Methodology Conference 2025

June 8-12, 2025

Stavros Niarchos Foundation Cultural Center (SNFCC), Athens-Greece

Abstracts of Poster Presentations

The conference is sponsored by the International Society for Bayesian Analysis (ISBA), OB section of ISBA, National Technical University of Athens (NTUA) and Athens University of Economics and Business (AUEB).

Contact: obayes2025@aueb.gr

Contents

Bayesian fMRI analysis in Fourier space

Konstantinos Bakas[1], John Kornak[2] and Hernando Ombao[1]

[1]King Abdullah University of Science and Technology, SA
[2]University of California, San Francisco, USA
konstantinos.bakas@kaust.edu.sa, john.kornak@ucsf.edu, hernando.ombao@kaust.edu.sa

Abstract

Functional magnetic resonance imaging (fMRI) is a powerful neuroimaging technique that indirectly measures neural activity by identifying changes in blood oxygenation levels. Many statistical models have been developed to analyze fMRI, with the massively univariate classical general linear model (GLM) being the most widely used. The current work proposes a spatiotemporal Bayesian method that leverages Fourier space properties to identify activated regions in the brain. Due to the high dimensionality in multi-subject fMRI studies, most Bayesian models are computationally expensive and require careful specification of prior distributions, usually through Gaussian Markov random fields (GMRF). Bayesian fMRI analysis in Fourier space (f-BFS) defines spatial prior distributions in Fourier space through parameter functions while simultaneously accounting for autocorrelation within the time series. Exploiting the independence assumption between Fourier space locations, the high-dimensional problem breaks down into multiple low-dimensional ones. The intuitive definition of the priors and the computational efficiency make the f-BFS approach a pivotal tool for analyzing fMRI data under the Bayesian perspective. The advantages of f-BFS are validated through a multi-subject simulation study and an fMRI study from the University of California, Riverside Center for Advanced Neuroimaging.

Multiphasic stochastic epidemic models

Petros Barmpounakis, Nikolaos Demiris

Objective Bayes Methodology, June 8-12 2025

Abstract

At the onset of the COVID-19 pandemic, a number of non-pharmaceutical interventions have been implemented in order to reduce the levels of infection, thus leading to multiple phases of transmission. The disease reproduction number Rt, a way of quantifying transmissibility, has been a key part of assessing such interventions' impact. In this paper, we discuss the distinct types of transmission models used and how they are linked. We consider a hierarchical stochastic epidemic model with piece-wise constant Rt, appropriate for modelling the distinct phases of the epidemic and quantifying the true disease magnitude. The points in time and scale of Rt changes are inferred directly from data while the number of transmissibility phases is allowed to vary. We learn the number of phases via appropriate Poisson point process and Dirichlet process-type modelling components, thus facilitating the determination of an important epidemic aspect. The models are evaluated using synthetic data sets and the methods are applied to freely available data from the United Kingdom and Greece as well as the states of California and New York. We estimate the true infected cases and the corresponding Rt, among other quantities, and independently validate the proposed approach using a large seroprevalence study. The results suggest that the key disease characteristics may be learned from publicly available data without imposing strong assumptions.

BMW: Inlier Prone Bayesian Models for Correlated Bivariate Data

Sumangal Bhattacharya,[1*] Ishapathik Das,[2]

[1*] Theoretical Statistics and Mathematics Unit, Indian Statistical Institute Delhi, New Delhi-110016, India

[2] Department of Mathematics and Statistics, Indian Institute of Technology Tirupati, Yerpedu - 517619, India

Abstract

In various fields of experiments, particularly prevalent in the lifetime of electronic components, clinical trials, rainfall data for specific locations, and insurance claim data, there are instances of zero or near-zero observations. These occurrences can be due to faulty construction, substandard quality, non-response to treatments, scant rain during the monsoon season, or a reduced accident rate. Similarly, certain bivariate correlated observations, such as twins' birth data, neonatal birth data, treatment responses for both eyes, and rainfall data from spatially correlated locations, exhibit a concentration of data points around a specific point (x0,y0), while the remaining observations follow a continuous distribution. In bivariate data, these concentrated responses are referred to as inliers. In nature, correlated bivariate data can include inliers and the region of these inliers is not restricted to a square shape. To address this, we utilize copulas to derive the Bivariate Modified Weibull (BMW) distribution, designed to capture correlated bivariate observations with inliers in either component. This approach uses the modified Weibull distribution as the marginal distribution, formulating it through a mixture model that combines a uniform distribution around a specific point with two parametric Weibull distributions. Moreover, we introduce a methodology for testing the presence of inliers within datasets using the Supremum of Bayes Factors (SBF). This methodology is validated through an extensive simulation study designed to assess the power of the test. The results, depicted using a confusion matrix, indicate that the accuracy of the test exceeds 98%. Estimation of the model parameters is carried out through a Bayesian framework, selecting suitable priors for each parameter. Given the model complexities and the dimensional breadth of the parameter space, we employ Hamiltonian Monte Carlo (HMC) techniques, specifically utilizing the No-U-Turn sampler (NUTS) for efficient posterior sampling of parameters. To demonstrate the applicability and assess the performance of our approach, we provide numerical examples and apply the method to analyze monsoon rainfall data from Assam, a popular tourist destination in India. In accordance with the prescribed testing methodology, our analysis indicates the presence of inliers within the rainfall dataset, as indicated by the SBF value exceeding 150. Our proposed model significantly fits the rainfall data better than the base model, as evidenced by the Deviance Information Criterion (DIC) and Bayesian Information Criterion (BIC).

Keywords: Bayes Factor, Bivariate Weibull Distribution, Copula, Hamiltonian Monte Carlo, Inliers, Instantaneous Failure, No-U-Turn Sampler.

Author Information:
Dr. Sumangal Bhattacharya (Male)
Current Position: Postdoctoral Fellow, Indian Statistical Institute, Delhi Centre, India
PhD: Awarded in 2024, Indian Institute of Technology Tirupati, India

Properly constrained reference prior decay rates: theory and practice

A. Van Biesbroeck^{†,1,2}, J. Garnier^{§,1}

†PhD student (presenting author). § PhD supervisor
PhD expected duration: Oct. 2022 – Sep. 2025

[1] CMAP, CNRS, ´Ecole polytechnique, Institut Polytechnique de Paris, France
antoine.van-biesbroeck@polytechnique.edu
[2] Universit´e Paris-Saclay, CEA, Service d'´Etudes M´ecaniques et Thermiques, France

Abstract

In Bayesian analysis, reference priors are widely recognized for their objective nature. Yet, they often lead to intractable and improper priors, which complicates their application. Besides, informed prior elicitation methods are penalized by the subjectivity of the choices they require. In this work, we aim at proposing a reconciliation of the aforementioned aspects. Leveraging the objective aspect of reference prior theory, we introduce two theoretical strategies of constraint incorporation to build tractable reference priors. One provides a simple and easy-to-compute solution when the improper aspect is not questioned, and the other introduces constraints to ensure the reference prior is proper, or it provides proper posterior. Our methodology emphasizes the central role of Jeffreys prior decay rates in this process.

Additionally, as the reference priors often yield a complex theoretical expression, we support the applicability of our results by suggesting a variational approximation of our reference priors. The latter consist in optimizing a neural network whose output should approach the target. We also conclude by applying the method to a case study in the field of seismic risk assessment.

References

[1] A. Van Biesbroeck, "Properly constrained reference priors decay rates for efficient and robust posterior inference", arXiv.2409.13041. 2024.

[2] N. Baillie, A. Van Biesbroeck, C. Gauchy, "Variational inference for approximate reference priors using neural networks", arXiv.2502.02364. 2025

Bayesian model discrimination for detecting heteroscedasticity in a regression model

Konstantinos Bourazas, Department of Economics, Athens University of Economics and Business

Laura Deldossi, Department of Statistical Sciences, Università Cattolica del Sacro Cuore

Guido Consonni, Department of Statistical Sciences, Università Cattolica del Sacro Cuore

Abstract

Identifying heterogeneity in regression models is an important topic, particularly when variability across individuals or groups is of interest, as is the case in many fields such as clinical trials, market research, and drug discovery. To optimize the process of data acquisition and maximize the amount of information gathered from our experiments, we adopt a Bayesian strategy to determine the optimal design ξ that enhances the ability to detect the presence of specified heteroscedasticity in Gaussian regression models.

The idea is to identify ξ so that the predicted responses, evaluated in ξ, vary as widely as possible across the rival models (with and without heterogeneity). Thus, by enhancing predictive model discrimination, we encourage stronger diversification of posterior model probabilities. By incorporating suitable priors on model space, as well as on the space of parameters (conditionally on each model), our approach uses Bayes Factors (BF) to discriminate between the two models, providing a structured approach for designing experiments that enhance the ability to detect heterogeneity, with potential applications across various fields.

PEPBVS: An R Package for Bayesian Variable Selection Using the Power–Expected–Posterior Prior

Konstantina Charmpi, Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Athens, Greece

Dimitris Fouskakis, Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Athens, Greece

Ioannis Ntzoufras, Department of Statistics, School of Information Sciences and Technology, Athens University of Economics and Business, Athens, Greece

Abstract

Bayesian variable selection in Gaussian linear models is a problem well–studied in statistical theory. When the prior on the model parameters is the power–expected–posterior or the intrinsic (a special case of the former), marginal likelihoods and posterior model probabilities can be derived in closed–form expressions. We have developed an R package called PEPBVS for this setup. Resolving numerical instabilities in the marginal likelihood computations is discussed. Model selection is performed by either implementing a full enumeration and evaluation of all possible models (for model spaces of small–to– moderate dimension) or using the Markov Chain Monte Carlo Composition (MC3) algorithm (for model spaces of large dimension). Selected models can be compared to those arising from other well–known priors on the model parameters. Additional functionalities such as making predictions and computing parameter estimates based on Bayesian model averaging, performing hypothesis testing, printing and visualizing the results in various ways are provided as well. How to use the package is illustrated with a real–life US crime dataset borrowed from the MASS package.

Bayesian causal inference for observational studies using factor models

Angelos Alexopoulos[1], Lea Anna Cozzucoli[2], Ioannis Ntzoufras[1] and Pantelis Samartsidis[3]

Abstract

Causal inference from observational panel data is frequently encountered in many scientific fields where randomized experiments are often infeasible. In fact, assessing the impact of an intervention by using time-series observational data across multiple units have to rely on statistical methods to estimate causal effects accounting for confounding factors and unobserved heterogeneity. Traditional statistical methods that fail to account for confounders which affect both the treatment and outcome often violate the assumption of parallel trends, which suggests that the treated and control units would have followed the same trajectory in the absence of the intervention, leading to biased estimates of causal effects.

Recently, there is a lot of work on the development of new methodologies which can more flexibly account for unobserved heterogeneity and confounders, providing more reliable causal estimates. Most of these methods build on the factor analysis model which naturally adjusts for unobserved confounders. However, they have several limitations. Notably, they can be applied to only a single outcome at a time while the temporal correlation between the multiple measurements of the outcome is not modeled. Moreover, when the total number of units is small, the new methods struggle to provide inference for the causal effects. Importantly, for most of the recent approaches it is hard to obtain uncertainty intervals for the causal effects of interest especially when the interest is on the estimation of unit-specific causal effects.

The factor analysis model is also our point of departure to construct a novel modeling and computational framework that addresses the limitations of recently developed causal inference methods. In particular, we employ dynamic factor models to account for unit-specific unobserved confounders and we assume that the latent factors evolve over time according to the autoregressive stochastic process of order one. The proposed modeling framework is designed to address staggered treatment adoption, where intervention timing varies between units. To provide uncertainty quantification for all quantities of interest we follow the Bayesian paradigm to estimate the parameters and latent states of the proposed model by using efficient Markov chain Monte Carlo methods.

After testing the proposed methods through simulations we applied them on a real data problem. In particular, by using unique administrative data from the Independent Authority for Public revenue in Greece we studied how tax and custom audits affect the compliance behavior of businesses in a particular, susceptible to fraud, sector of the Greek economy. The estimated individual treatment effects reveal heterogeneous intervention responses across units, with varying magnitude and pattern. Some units experience immediate and strong effects, while others show delayed or moderate response to the intervention.

[1] Athens University of Economics and Business, Greece,
[2] University of Trieste, Italy, leaanna.cozzucoli@phd.units.it
[3] University of Cambridge, UK,O'Bayes 2025 Posters

Bayesian Model Screening: On the way to scalable approximations

Argyro Damoulaki[1,2] and Ioannis Ntzoufras[1,3]

[1] Department of Statistics, Athens University of Economics and Business

[2] PhD student in Statistics, argyrodam@aueb.gr

[3] Professor in Statistics, ntzoufra@aueb.gr

Abstract

Bayesian model evaluation and selection is a fundamental challenge in statistical analysis, particularly in high-dimensional settings where the number of possible models grows exponentially with the number of covariates under consideration p. Posterior model probabilities and inclusion probabilities are typically computed using Bayes factors, which require evaluating 2p marginal likelihoods. This process often relies on Markov Chain Monte Carlo (MCMC) methods, which can be also computationally demanding.

In this work, we examine an alternative approach to approximate Bayes factors and variable inclusion probabilities without relying on MCMC or exploring the full model space in variable selection problems. Instead, we aim to develop an efficient screening approach that reduces the model space by setting limits/thresholds for inclusion probabilities and hopefully build computationally scalable techniques. Specifically, as a first approach, we explore cases where a conjugate prior structure is assumed, and the predictors are independent, allowing for computationally efficient approximations. Finally, our method aims to provide an initial screening of variables to reduce the number of models under consideration accounting for multicollinearity among predictors

The Interplay between Bayesian Inference and Conformal Prediction

Nina Deliu[1,2,*] and Brunero Liseo[1]

[1]MEMOTEF, Sapienza University of Rome, Italy

[2]MRC – Biostatistics Unit, University of Cambridge *nina.deliu@uniroma1.it

Abstract

Conformal prediction has emerged as a cutting-edge approach within statistics and machine learning, offering methods to derive uncertainty intervals with finite-sample frequentist coverage, or validity. Yet, its interplay with Bayesian statistics–often criticised for lacking frequentist guarantees–remains underexplored. Existing literature has hinted at the potential role of conformal prediction to "calibrate" (or validate in a frequentist sense) Bayesian procedures, calling for more attention on such frequentist-Bayesian methods (Wasserman, 2011). In this work, we further argue that Bayesian uncertainty quantification has the potential to enhance the distribution-free properties of conformal prediction. For example, it may optimise efficiency (that is, the interval's volume or Lebesgue measure) in a Bayes risk sense (Hoff, 2023). Therefore, the two frameworks may complement each other in achieving a more optimal balance between validity and efficiency. After surveying existing ideas for a Bayesian conformal inference direction, we formalize the full and split approaches of conformal prediction under a Bayesian framework. Special emphasis is placed on understanding the roles of exchangeability, statistical efficiency, and computational complexity. Parametric and non-parametric examples are presented and evaluated empirically. Finally, we discuss the advantages of Bayesian conformal inference in survey sampling, connecting to the "Calibrated Bayes" idea proposed by Little (2011) in a search for compromise between model-based and design-based solutions. In conclusion, this work aims to bridge existing gaps between competing statistical paradigms, providing new avenues for future theoretical and applied research.

References

Wasserman, L. (2011). Frasian Inference. *Statistical Science*, 26(3), 322-325

Hoff, P. (2023). Bayes-optimal prediction with frequentist coverage control. *Bernoulli*, 29(2), 901-928

Little, R. (2011). Calibrated Bayes, for Statistics in General, and Missing Data in Particular. *Statistical Science*, 26(2), 162-174

A Bayesian approach to extrapolation

Nikolaos Demiris

Department of Statistics, Athens University of Economics and Business

Abstract

This poster presentation is concerned with a Bayesian solution to the necessity for extrapolation in several application areas of statistics. Moving beyond the data is potentially dangerous and should not be made uncritically. However, such extrapolations are implicitly or explicitly needed in a number of diverse areas. Here we present a Bayesian approach to two such problems.

The first is concerned with survival extrapolation as needed for the estimation of the mean survival and other function(al)s of the survival curve. We perform this using a flexible family of parametric survival models termed poly-hazard models. We embed mortality projections in the sources of evidence that inform long-term survival and synthesize all the data using Hamiltonian Monte Carlo methods. The methods are illustrated using examples from cancer and cardiovascular applications.

The second extrapolation problem stems from the design of prediction algorithms. This often occurs in the presence of limited data and involves distinct types of algorithms. We present a framework based on learning curves represented as a suitable Gaussian process (GP). Then extrapolation is based on historical data acting as an anchor to the currently available limited amount of data. And within the mean structure of the GP one models the adjustments required for suitable extrapolation. This framework is applied to the accuracy of various prediction models and illustrated using multi-modal data from cancer predictions.

Theoretical guarantees for Bayesian inference with Missing Data

Stefan Franssen and Judith Rousseau

Abstract

Missing data is a common phenomenon that statisticians have to solve. This is already a widely studied problem in the frequentist literature, and efficient estimators point estimators are known. However, these estimators can be very complicated. Simpler methods are also applied in practice, but these can result in unwanted behaviour like biased estimates. Bayesian methods have been proposed as an alternative; however, no theoretical guarantees are known. We aim to rectify this. We have preliminary results for the linear regression model $Y = \beta T X + Z$, where X are covariates following some unknown distribution p_0, and Z normal noise. Because we do not always fully observe X, we also have to provide a prior on the distribution of X. To stop problems arising from misspecification, we have to put a nonparametric prior on p, and a parametric prior on β. Our goal is to provide a Bernstein-von Mises theorem for estimating β, considering p a nuisance parameter. This is an ongoing work.

Inference via mixture predictive distributions

Samuele Garelli, Department of Statistical Sciences “Paolo Fortunati” – University of Bologna (Italy)

Co-authors: Fabrizio Leisen (King's College London), Luca Pratelli (Accademia Navale Livorno), Pietro Rigo (University of Bologna)

Position: PhD Student

Supervisor: Prof. Pietro Rigo (University of Bologna)

Co-supervisor: Prof. Fabrizio Leisen (King's College London)

Personal webpage: https://www.unibo.it/sitoweb/samuele.garelli2/en

Abstract

A novel paradigm for Bayesian inference, which bypasses the usual likelihood-prior scheme, has been recently introduced by Fong, Holmes and Walker. The idea is to reconstruct the unobserved part of the population by sampling from a sequence of predictive distributions and then estimate the parameter of interest as a function of the observed and imputed data. Such mechanism is called predictive resampling. To guarantee reliable inference, predictives need to have a good fit on the initial dataset and converge to a limit distribution that preserves the information provided by the observed sample.

A new class of predictive distributions applicable in this context is proposed. We start by partitioning the observed data with model-based clustering and setting the first predictive as a mixture law whose components have the same mean and covariance matrix as the observed clusters. Then, the next predictives are obtained through a mechanism inspired by Polya urn models, which updates only the mean and covariance matrix of the component from which observations are drawn. The initial clustering ensures these predictives fit well the observed data. Moreover, we prove they converge almost surely in total variation to a mixture whose random parameters (i.e. weights, means and covariance matrices) have mean equal to the parameters of the initial predictive. This ensures predictive resampling generates uncertainty but no new information (which is desired, since the only available information about the true population is contained in the observed sample). Such approach is quite flexible, in that theoretical results hold for any mixture whose components have density with respect to Lebesgue measure.

The methodology is applied to simulated and real data and satisfying results are obtained in posterior estimation of parameters (e.g. mean, variance, skewness, kurtosis, quantiles and covariance) and in regression on non-linear and heteroscedastic data.

References:

Fong, Holmes, Walker (2023). Martingale posterior distributions (with discussion). JRSSB.

Garelli, Leisen, Pratelli, Rigo (2024). Asymptotics of predictive distributions driven by sample means and variances. https://arxiv.org/abs/2403.16828 .O'Bayes 2025 Posters

Bayesian nonparametric processes for clustering count data with potentially unobserved traits

Lorenzo Ghilotti, PhD student, Department of Economics, Management and Statistics, University of Milano–Bicocca, 20126 Milano, Italy. Email: l.ghilotti@campus.unimib.it

Co-authors: Federico Camerlenghi, Michele Guindani and Tommaso Rigon

Abstract

Count data arise in diverse domains such as criminology, where investigations track criminals at meetings and the information is encoded through binary data, and other fields like ecology. A key goal may often be to cluster a sample of subjects, e.g., criminals, based on their count records for the observed traits, e.g., meetings. While clustering the counts of the observed traits is intuitive, such models are often misspecified. Indeed, in a variety of settings, it is realistic to assume that a larger set of traits exists, of which only a subset is observed. In the criminal context, it is plausible that investigations missed some of the actual meetings. We propose a novel Bayesian nonparametric methodology for clustering that accounts for potentially unobserved traits. Our approach embeds trait allocation models into a mixture model. The random mixing measure is almost surely supported on a countable set of realizations from finite Completely Random Measures with shared common atoms. We analytically show that our methodology yields fewer clusters than its natural naïve competitor. Assuming a known clustering structure, we derive the closed-form expression for the marginal distribution of the data, we characterize the posterior distribution of the de Finetti vector of random measures and the predictive structure for new subjects. This fully-Bayesian analysis enables the development of a fast, efficient marginal algorithm and allows for trait-by-trait comparisons across clusters. Finally, we illustrate our methodology by analyzing a criminal dataset documenting the attendance of 'Ndrangheta affiliates at meetings.

A comparison of objective priors for Cronbach's coefficient alpha using a three-component hierarchical model

S. R. Izally[1,*], A. J. van der Merwe[2] and L.Raubenheimer[1]

[1]Department of Statistics, Rhodes Unviersity, Makhanda, South Africa

[2]Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa

Abstract

In this paper the reference and probability matching priors for Cronbach's alpha will be derived and the performance of these two priors will be compared to that of the well-known Jeffreys prior and a divergence prior. Cronbach's alpha is a measure used to assess the reliability of a set of test items. A simulation study will be considered to compare the performance of the priors, where the coverage rates, average interval lengths and standard deviations of the interval lengths will be computed. Two illustrative examples will also be considered. Throughout the paper, the three-component hierarchical model will be used.

Keywords: Coverage; Cronbach's alpha; Divergence prior; Probability matching prior; Reference prior.

Bayesian Approaches for Predicting Dementia Onset and Progression from Longitudinal Neuroimaging Data

Athanasia Kalogirou, Mathematician, Statistician, PhD Candidate, Statistics Department, Athens University of Economics and Business, akalogirou@aueb.gr

Ioannis Ntzoufras, Professor, Head of Statistics Department, Athens University of Economics and Business, Ntzoufras@aueb.gr

John Kornak, Professor in Residence, Head of Data Science Program and Director of Biostatistics Consulting Unit, Department of Epidemiology and Biostatistics, University of California, San Francisco, john.kornak@ucsf.edu

Abstract

Dementia is a devastating form of disease with high heterogeneity. This variation is due to different underlying disease types as well as sub-types and disease presentations, leading to variation in patients' experience and prognoses. The heterogeneity in dementia symptoms is further influenced by factors such as an individual's overall health, as well as the brain regions most affected. Even within the same type of dementia, there is inter-individual variability in the rate of disease progression and symptoms.

Understanding this heterogeneity is crucial for accurate diagnosis and disease course prediction. In particular, as potential new treatments are appearing on the market with many more in development, predicting cognitive impairment and decline is vital for optimizing patient selection in clinical trials and planning personalized treatment strategies.

Our research used brain imaging structural measures data (regional volumes) from dementia patients (the ALLFTD study) and at-risk individuals. It explored Bayesian Variable Selection methods for predicting the presence and severity of cognitive impairment. The frontal and temporal lobes were identified as key regions for degeneration onset. The most promising models combined predictive ability with clinical interpretability, which is essential for medical practice adoption.

Bayesian Analysis of Latent Underdispersion Using Discrete Order Statistics

Jimmy Lederman
Department of Statistics, University of Chicago
jlederman@uchicago.edu
Coauthor(s): Aaron Schein

Abstract

Researchers routinely analyze count data using models based on a Poisson likelihood, for which there exist many analytically convenient and computationally efficient strategies for posterior inference. A limitation of such models however is the equidsipersion constraint of the Poisson distribution. This restriction prevents the model’ s likelihood, and by extension its posterior predictive distribution, from concentrating around its mode. As a result, these models are parametrically bound to produce probabilistic predictions with high uncertainty, even in cases where low uncertainty is supported by the data. While count data often exhibits overdispersion marginally, such data may nevertheless be consistent with a likelihood that is underdispersed conditionally, given parameters and latent variables. Detecting conditional underdispersion, however, requires one to fit the “right” model and thus the ability to build, fit, and critique a variety of different models with underdispersed likelihoods. Towards this end, we introduce a novel family of models for conditionally underdispersed count data whose likelihoods are based on order statistics of Poisson random variables. More specifically, we assume that each observed count coincides with the j^{th} order-statistic of D latent i.i.d. Poisson random variables, where j and D are user-defined hyperparameters. To perform efficient MCMC-based posterior inference in this family of models, we derive a data-augmentation strategy which samples the other $D-1$ latent variables from their exact conditional, given the observed (j,D)-order statistic. By relying on the explicit construction of a Poisson order statistic, this data augmentation strategy can be modularly combined with the many existing inference strategies for Poisson-based models. We generalize this approach beyond the Poisson to any non-negative discrete parent distribution and, in particular, show that models based on negative binomial order statistics can flexibly capture both conditional under and overdispersion. To illustrate our approach empirically, we build and fit models to three real count data sets of flight times, COVID-19 cases counts, and RNA-sequence data, and we demonstrate how models with underdispersed likelihoods can leverage latent structure to make more precise probabilistic predictions. Although the possibility of conditional underdispersion is often overlooked in practice, we argue that this is at least in part due to the lack of tools for modeling underdispersion in settings where complex latent structure is present.

A uniform shrinkage prior in spatiotemporal Poisson models for count data

Krisada Lekdee*, Faculty of Science and Technology, Department of Mathematics and Statistics, Rajamangala University of Technology Phra Nakhon, Bangkok, Thailand;

Chao Yang*, Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA;

Lily Ingsrisawang, Faculty of Science, Department of Statistics, Kasetsart University, Bangkok, Thailand;

Yisheng Li ✉, Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

*Co-first authors; ✉Email: ysli@mdanderson.org

Abstract

We consider default Bayesian inference in a Poisson generalized linear mixed model for spatiotemporal data. Normal random effects are used to model the within-area correlation over time and spatial effects represented with a proper conditional autoregressive (CAR) model are used to model the between-area correlations. We develop a uniform shrinkage prior (USP) for the variance components of the spatiotemporal random effects. We prove that the proposed USP is proper, and the resulting posterior is proper under the proposed USP, an independent flat prior for each fixed effect, and a uniform prior for a spatial parameter, under suitable conditions. Posterior simulation is implemented and inference made using the OpenBUGS, R2OpenBUGS and RStan software packages. We illustrate the proposed method by applying it to a leptospirosis count dataset with observations from 17 northern provinces of Thailand across four quarters in 2011 to construct the disease maps. According to the deviance information criterion, the proposed USP for the variance components of the spatiotemporal effects yields a better performance than the conventional inverse gamma priors. A simulation study suggests that the estimated fixed- effect parameters are accurate, based on a relative bias criterion. The leptospirosis data analysis suggests the top 10 estimated leptospirosis morbidity rates (per 100,000 population), ranging from the highest to the lowest, are in Nan (Quarter (Q) 3, 9.8060), Chiang Rai (Q3, 3.1010), Nan (Q2, 3.0190), Nan (Q4, 2.2370), Nan (Q1, 2.2330), Phitsanulok (Q3, 2.2240), Lampang (Q3, 1.7890), Uttaradit (Q3, 1.5230), Phrae (Q3, 1.4720), and Phetchabun (Q3, 1.4380) provinces, respectively

Matching-assisted power prior for incorporating real-world data in randomized trials

Presenter: Bo Lu, The Ohio State University

Email: lu.232@osu.edu

Abstract

Leveraging external data information to supplement randomized clinical trials has been a popular topic in recent years, especially for medical device and drug discovery. In rare diseases, it is very challenging to recruit patients and run a large-scale randomized trial. To take advantage of real world data from historical trials on the same disease, we can run a small hybrid trial and borrow historical controls to increase the power. Bayesian power prior is a popular method in current practice. To better mitigate observed bias when incorporating external data, we propose a matching-assisted power prior approach. A subset of comparable external subjects is first selected by groups through template matching, and different weights are assigned to these groups based on their similarity to the current study population. Power priors are then implemented to incorporate the information into Bayesian inference. Unlike conventional power prior methods, which discount all control patients similarly, matching pre-selects good controls, hence improved the quality of external data being borrowed. We compare its performance with the existing power prior approach through simulation studies and illustrate the implementation using data from a real acupuncture clinical trial.

Mixture priors for replication studies

Roberto Macrì Demartino [1], Leonardo Egidi [1], Leonhard Held [2] and Samuel Pawel [2]

Abstract

The credibility of scientific research relies heavily on the replicability of its findings. However, in recent years, an increasing number of published results have failed to replicate, raising concerns about a "replication crisis" in several fields. As a result, the scientific community has increasingly emphasised the importance of replication studies, even though establishing replication success remains challenging.

Analysing replication studies essentially involves using historical data from the original study. Given the inherent nature of sequential information updating, Bayesian methods are a natural choice. In particular, an intuitive strategy is to incorporate historical information by using a prior distribution based on the data from the original study for the analysis of the replication data. We propose a novel Bayesian approach that employs mixture priors. Specifically, our method uses the mixture of the posterior distribution from the original study with a non-informative prior to evaluate the replication study. The mixture weight determines the extent to which the original and the replication data are pooled.

This mixture prior approach offers a flexible alternative to existing methods–such as hierarchical models and power priors–for assessing replication success. We explore two strategies for setting the mixture weight parameter. The first strategy fixes the weight at a specific value, for instance, based on expert knowledge or an empirical Bayes estimate, and then assesses the sensitivity of this choice using a tipping point analysis. The second strategy introduces uncertainty by assigning a prior distribution to the mixture weight parameter. Furthermore, within these frameworks, Bayes factors can be used for formal hypothesis testing, such as evaluating the presence or absence of an effect and to assess how closely the replications align with the original study. We analyse the asymptotic behaviour of the marginal posterior distribution for the weight parameter as the Bayes factor tends to zero or infinity. Moreover, we examine how the Bayes factor associated with the effect size behaves as the replication study's standard error approaches zero.

We demonstrate the applicability of our method using data from a communication science experiment titled "Labels", which was part of a large-scale replication project. In this experiment, the original study reported a positive effect of labelling on perceived attitudes, while three subsequent external replications yielded contrasting results.

Our findings suggest that mixture priors are a valuable and intuitive alternative to other Bayesian methods for analysing replication studies. Furthermore, we provide the free and open-source R package repmix, which implements the proposed methodology.

[1] University of Trieste, Department of Economics, Business, Mathematics, and Statistics "Bruno de Finetti", roberto.macridemartino@deams.units.it, https://orcid.org/0000-0002-5296-6566 ; legidi@units.it, https://orcid.org/0000-0003-3211-905X

[2] University of Zurich, Epidemiology, Biostatistics and Prevention Institute, Center for Reproducible Science, leonhard.held@uzh.ch, https://orcid.org/0000-0002-8686-5325; samuel.pawel@uzh.ch, https://orcid.org/0000-0003-2779-320X

A New Multivariate Logistic Distribution for Fast Objective Bayesian Inference

Paolo Onorati

Postdoctoral Researcher at Department of Statistical Sciences, University of Padova, Italy

paolo.onorati@unipd.it

Abstract

The Generalized Extreme Value (GEV) distribution is widely used for modeling extreme events. Despite its frequent application, there is no consensus on the choice of prior distributions for the parameters in the Bayesian framework. In this work, we propose the use of an objective prior based on a scoring rule. This leads to use a multivariate Lomax distribution for positive parameters and a double multivariate Lomax distribution for the general case. However, despite being well motivated by theoretical arguments, this choice introduces computational challenges, as the full conditional distributions do not have a closed form.

We provide practical solutions to these issues by exploiting a Generalized Elliptical Slice Sampling (GESS) method, which yields a self-contained algorithm that does not require tuning parameters and is rejection-free, ensuring that consecutive values in the generated Markov chain are distinct. This method is quite general and can be extended to many other parametric families. The approach utilizes a new version of the multivariate logistic distribution, represented as a scale mixture of Gaussian distributions. We show how to compute its probability density function and how to sample from the full conditional of the mixing distribution. The performance of the algorithm is evaluated through simulation studies and empirical analysis.

FBVS: a Fast Bayesian Variable Screening Method in Multiple Regression

Roberta Paroli, Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy (roberta.paroli@unicatt.it)

Dimitris Fouskakis, Department of Mathematics, National Technical University of Athens, Athens, Greece (fouskakis@math.ntua.gr)

Ioannis Ntzoufras, Department of Statistics, Athens University of Economics and Business, Athens, Greece (ntzoufras@aueb.gr)

Abstract

We propose a fast Bayesian variable screening method for Normal regression models using thresholds on Pearson and partial correlation coefficients. Although the proposed method is based on the computation of correlation coefficients, it is derived using purely Bayesian arguments obtained from thresholds on Bayes factors and posterior model odds.

The proposed method can be used to screen out the "non—important" covariates and reduce the size of the model space even in cases when the number of covariates is larger than the sample size. Then, on the reduced model space, obtained from the proposed approach, more accurate, traditional, computer-intensive, Bayesian variable selection methods can be implemented, if needed.

We focus on the use of g-priors where Bayes factors can be obtained analytically and the corresponding correlation threshold computations are exact. Nevertheless, the approach is general and can be easily extended to any prior setup.

The proposed method is illustrated using simulated examples

A General Framework for Bayesian Model Assessment via Case Deletion Weights

Wenxin Du (du.888@buckeymail.osu.edu) and Mario Peruggia (presenter, peruggia@stat.osu.edu)

Department of Statistics

The Ohio State University

Abstract

As Bayesian models become increasingly flexible and high-dimensional, precise diagnostic tools are essential for ensuring model validity. This work introduces a framework based on case deletion weights, which can reveal how individual observations influence the posterior and evaluate how well the model predicts new or held-out data. Under model misspecification, statistics derived from these weights exhibit different asymptotic behaviors than in correctly specified settings, leading to a general strategy for identifying unmodeled structure or incorrect assumptions. These results relate naturally to established methods for model assessment, such as the Conditional Predictive Ordinate (CPO), the Watanabe–Akaike Information Criterion (WAIC), and Leave-One-Out Cross-Validation (LOO-CV). Through simulated examples, our framework is shown to support model comparison, selection, and evaluation within a unified approach. Ongoing work extends these ideas beyond current techniques, offering new approaches to detect and locate the source of misspecification in Bayesian models.

Improving variable selection properties by integrating data

Paul Rognon-Vael (U. Pompeu Fabra, U. Polit`ecnica de Catalunya),

David Rossell (U. Pompeu Fabra), Piotr Zwiernik (U. Pompeu Fabra, U. of Toronto)

Abstract

We study the variable selection properties of ℓ_0 penalties depending on external information in the Gaussian sequence model and high-dimensional linear regression. We consider in particular the case where external information helps partition variables into blocks with potentially distinct characteristics and study non-exchangeable block-dependent penalties. We show that such block informed penalization achieves variable selection consistency under milder conditions than standard penalization and that oracle block penalties converge at faster rates. We propose an empirical Bayes method for the calibration of block penalties and demonstrate it realizes the benefits of the oracles.

Priors for second-order unbiased Bayes estimator

Mana Sakai, Graduate School of Economics, The University of Tokyo, High-Dimensional Causal Analysis Team, RIKEN Center for Advanced Intelligence Project, sakai-mana@g.ecc.u-tokyo.ac.jp

Takeru Matsuda, Graduate School of Information Science and Technology, The University of Tokyo, Statistical Mathematics Unit, RIKEN Center for Brain Science, matsuda@mist.i.u-tokyo.ac.jp

Tatsuya Kubokawa, Graduate School of Economics, The University of Tokyo, tatsuya@e.u-tokyo.ac.jp

Abstract

Asymptotically unbiased priors, introduced by Hartigan (1965), are designed to achieve second-order unbiasedness of Bayes estimators. Our research extends Hartigan's framework to non-i.i.d. models by deriving a system of partial differential equations that characterizes asymptotically unbiased priors. Furthermore, we establish a necessary and sufficient condition for the existence of such priors and propose a simple procedure for constructing them. The proposed method is applied to several examples, including the linear regression model and the nested error regression (NER) model (also known as the random effects model). Simulation studies evaluate the frequentist properties of the Bayes estimator under the asymptotically unbiased prior for the NER model, highlighting its effectiveness in small-sample settings.

Reference

[1] Hartigan, J. A. (1965). The asymptotically unbiased prior distribution. The Annals of Mathematical Statistics 36(4), 1137–1152

VARIABLE SELECTION AND ESTIMATION USING NONLOCAL PRIOR MIXTURES FOR DATA WITH WIDELY VARYING EFFECT SIZES

Nilotpal Sanyal
Department of Mathematical Sciences, University of Texas at El Paso,
500 W. University Ave., El Paso, TX 79968, USA
Email: nsanyal@utep.edu

Abstract

Different nonlocal priors offer distinct advantages for supporting smaller and larger effect sizes. Building on this observation, we explore mixtures of nonlocal priors, which provide enhanced adaptability for sparse datasets with widely varying effect sizes—common in public health data integrating biological, clinical, and environmental features. Specifically, we investigate a mixture prior comprising a point mass and two nonlocal prior components. Beyond evaluating existing alternatives, we introduce a novel, objective method for setting the tuning parameters of nonlocal priors to maximize support for diverse effect sizes. Adopting a full Bayesian framework, we develop a joint selection and estimation procedure using Gibbs sampling and Laplace approximation. Preliminary analyses with sensible hyperpriors reveal that our method delivers competitive performance compared to established approaches like LASSO and the horseshoe prior. The utility of the proposed method is further demonstrated through real-world data applications.

Keywords: Nonlocal prior; mixture prior; selection and estimation; Gibb's sampling; Laplace approximation

A general framework for probabilistic model uncertainty

Vik Shirvaikar, Stephen G. Walker, Chris Holmes

Abstract

Existing approaches to model uncertainty typically either compare models using a quantitative model selection criterion or evaluate posterior model probabilities having set a prior. In this paper, we propose an alternative strategy which views missing observations as the source of model uncertainty, where the true model would be identified with the complete data. To quantify model uncertainty, it is then necessary to provide a probability distribution for the missing observations conditional on what has been observed. This can be set sequentially using one-step-ahead predictive densities, which recursively sample from the best model according to some consistent model selection criterion. Repeated predictive sampling of the missing data, to give a complete dataset and hence a best model each time, provides our measure of model uncertainty. This approach bypasses the need for subjective prior specification or integration over parameter spaces, addressing issues with standard methods such as the Bayes factor. Predictive resampling also suggests an alternative view of hypothesis testing as a decision problem based on a population statistic, where we directly index the probabilities of competing models. In addition to hypothesis testing, we provide illustrations from density estimation and variable selection, demonstrating our approach on a range of standard problems.

On the Statistical Capacity of Deep Generative Models

Edric Tam (presenter, edrictam@stanford.edu), David Dunson (dunson@duke.edu)

Abstract

Deep generative models are routinely used in generating samples from complex, high dimensional distributions. Despite their apparent successes, their statistical properties are not well understood. A common assumption is that with enough training data and sufficiently large neural networks, deep generative model samples will have arbitrarily small errors in sampling from any continuous target distribution. We set up a unifying framework that debunks this belief. We demonstrate that broad classes of deep generative models, including variational autoencoders and generative adversarial networks, are not universal generators. Under the predominant case of Gaussian latent variables, these models can only generate concentrated samples that exhibit light tails. Using tools from concentration of measure and convex geometry, we give analogous results for more general log-concave and strongly log-concave latent variable distributions. We extend our results to diffusion models via a reduction argument. We use the Gromov–Levy inequality to give similar guarantees when the latent variables lie on manifolds with positive Ricci curvature. These results shed light on the limited capacity of common deep generative models to handle heavy tails, and the importance of the choice of latent variable distributions. We illustrate the empirical relevance of our work with simulations and financial data.

Bayesian Transfer Learning to Boost Facebook Prophet's Performance on Short Time-Series *

Jovan Krajevski and Biljana Tojtovska

Faculty of Computer Science and Engineering
University "Ss. Cyril and Methodius"
Skopje, R.N. Macedonia
jovan.krajevski@gmail.com
biljana.tojtovska@finki.ukim.mk

Abstract

Seasonality and trend are common characteristics of many time-series, particularly in business contexts. Traditional statistical and machine learning approaches often assert that modeling seasonality requires a sample size at least twice the seasonal period. Additionally, introducing non-linear trend components with change points risks overfitting on small samples. However, recent advancements have challenged numerous long-held assumptions. In particular, transfer learning has demonstrated that high-performing models with a high number of parameters can be developed with limited data by fine-tuning models that are pretrained on related tasks and datasets. Inspired by this paradigm shift, we explore multiple effective transfer learning techniques within a fully Bayesian framework. By building on the Facebook Prophet, we utilize Variational Inference to efficiently approximate the posterior distribution of the seasonal and trend parameters learned from the long time-series. We complete the transfer learning paradigm by setting priors for the short time-series that are dependent on these posterior distribution approximations. We are developing a Python package that reimagines the user experience compared to Facebook Prophet's design by offering a more intuitive API that closely resembles the mathematical notation used to define models. This tool seamlessly enables training on long time-series and fine-tuning on short ones, making advanced forecasting accessible to all. We validate our approach on daily stock data, using just three months of observations to predict an entire year. The preliminary results based on these methods show significant improvement across standard forecasting metrics, capturing long-term seasonalities and trend shifts that traditional models miss. The next step in our research is to compare the results to commonly used models for short time-series like ETS, ARIMA, SARIMA and State space models.

Keywords: Bayesian inference · Bayesian transfer learning · Facebook Prophet · seasonality · trend · short time-series · stock data.

Bayesian Modelling and Inference for the Degrees of Freedom of the Wishart Distribution

Lampis Tzai
School of Criminal Justice, University of Lausanne, Switzerland and Department of Statistics Athens University of Economics and Business, Greece
Email: lampis.tzai@unil.ch

Ioannis Ntzoufras
Department of Statistics Athens University of Economics and Business, Greece
E-mail: ntzoufras@aueb.gr

Silvia Bozza
Department of Economics, Ca' Foscari University of Venice, Italy and School of Criminal Justice, University of Lausanne, Switzerland
Email: silvia.bozza@unive.it

Abstract

This study aims to enhance Bayesian modeling for covariance matrices in multivariate statistics by investigating various Markov Chain Monte Carlo (MCMC) methods and prior distributions for the degrees of freedom in the Wishart distribution, with the goal of drawing more efficient posterior samples. Specifically, the No-U-Turn Sampler (NUTS) implemented in Stan, along with hybrid MCMC methods such as Random Walk Metropolis within Gibbs, Slice Sampling within Gibbs, and Hamiltonian Monte Carlo within Gibbs, are rigorously evaluated. These hybrid methods efficiently leverage the conjugate form of the conditional distribution of the scale matrix, provided by the inverse Wishart distribution. The evaluation employs both simulated datasets, which provide a controlled environment for assessing performance, and real datasets, which offer practical validation. Each method exhibits distinct strengths and weaknesses in terms of robustness, accuracy, and sampling efficiency. For low-dimensional covariance matrices, NUTS with a Log-Normal prior demonstrated superior performance, whereas Slice Sampling within Gibbs proved more effective for high-dimensional data, where NUTS is constrained by time limitations. The results from real datasets are consistent with those from simulated datasets, confirming the reliability of the methods. Bayesian inference for the parameters of the Wishart distribution is conducted for three real datasets, considering the distribution of the degrees of freedom and the eigenvalues of the scale matrix.

Keywords: 1. Wishart distribution 2. Degrees of freedom 3. Hybrid MCMC 4. No-U-Turn Sampler 5. Bayesian inference 6. High-dimensional covariance matrices

Shrinkage Priors via Random Imaginary Data

G. Tzoumerkas, Department of Mathematics, Department of Mathematics, Greece,
tzoumg@mail.ntua.gr

D. Fouskakis , National Technical University National Technical University of Athens, Greece of
Athens, fouskakis@math.ntua.gr

Abstract

In this work, focus is given in the Bayesian variable selection problem in high–dimensional linear regression models. The use of shrinkage priors, when the number of available observations n is less than the number of explanatory variables p, is a well-established method, which shares great theoretical and empirical properties. The Power-Expected-Posterior (PEP) prior methodology, uses random imaginary (training) data in order to create objective and compatible priors, with appealing interpretation. We explore the idea of the utilization of an augmented imaginary design matrix with orthogonal columns, under the PEP methodology, in order to produce independent PEP-Shrinkage priors, using random imaginary data. Under this setup, properly chosen hyperpriors are placed on the power parameters of the PEP methodology, in order to produce mixtures of independent priors suitable for the variable selection problem when n << p. We check the theoretical properties of our proposed method and we explore its behavior via simulated studies and a real life example.

On Some Composite Estimators for Domain Estimation under Systematic Sampling Design

Garima Yadav* [1] and Piyush Kant Rai†[2]

[1,2] Dept. of Statistics, Banaras Hindu University, Varanasi

Abstract

Domain Estimation in Survey Sampling refers to the process of estimating population parameters (such as totals, means, or proportions) for specific subgroups, or ”domains,” within the larger population using data collected from a sample. In many instances, the focus is not on the entire population but on specific subgroups that are of particular relevance. This study presents a family of composite estimators for domain estimation using auxiliary information under a systematic sampling scheme. The proposed estimators are based on modifications of the Bahl and Tuteja (1991) estimator in survey sampling, leading to a generalized class of composite estimators. These estimators integrate exponential-type direct and synthetic ratio estimators, along with different forms of the proposed domain estimators. Theoretical properties of the proposed estimator are also discussed, including expressions for the weights used. Furthermore, the study demonstrates the application of these composite estimators for domain estimation and compares their relative performance to the individual constituent estimators through a simulation study.

Keywords-Composite estimator, Domain estimation. Auxiliary variable, Exponential type estimator, Systematic sampling scheme.

---

*garima.yadav@bhu.ac.in

†raipiyush5@gmail.com

Principled priors for Bayesian inference of circular models

Xiang Ye
Department of Statistics, King Abdullah University of science and technology
and
Janet Van Niekerk
Department of Statistics, King Abdullah University of science and technology
and
H˚avard Rue
Department of Statistics, King Abdullah University of science and technology

Abstract

Advancements in computational power and methodologies have enabled research on massive datasets. However, tools for analyzing data with directional or periodic characteristics, such as wind directions and customers' arrival time in 24-hour clock, remain underdeveloped. While statisticians have proposed circular distributions for such analyses, significant challenges persist in constructing circular statistical models, particularly in the context of Bayesian methods. These challenges stem from limited theoretical development and a lack of historical studies on prior selection for circular distribution parameters.

In this article, we propose a practical and systematic framework for selecting priors that effectively prevents overfitting in circular scenarios, especially when there is insufficient information to guide prior selection. We introduce well-examined Penalized Complexity (PC) priors for the most widely used circular distributions. Comprehensive comparisons with existing priors in the literature are conducted through simulation studies and a practical case study. Finally, we discuss the contributions and implications of our work, providing a foundation for further advancements in constructing Bayesian circular statistical models

Bayesian Mediation Analysis for Time-to-Event Outcome: Investigating Racial Disparity in Breast Cancer Survival

Qingzhao Yu [1] and Bin Li [2]

Abstract

Mediation analysis investigates the effects of mediators intervening in the pathways between exposure and outcome variables. Bayesian methods are naturally suited for mediation analysis due to the hierarchical structure of Bayesian models. This paper introduces an innovative adaptive Bayesian mediation analysis method incorporating adaptive Laplace priors into the predictive model to account for high-dimensional mediators. This approach applies a penalization function to the estimated direct and indirect effects rather than solely on the coefficients of predictive models. Consequently, estimated effects that lack statistical significance may shrink to zero, facilitating a more robust analysis.

We demonstrate the efficacy of our adaptive mediation analysis method through simulations and analysis of a Louisiana triple-negative breast cancer (TNBC) dataset to examine racial disparity in diagnosed stages among TNBC patients diagnosed between 2010 and 2017. The dataset is linked to the 2017 hazardous air pollutant emissions burden estimation database using patients' residential addresses. Our analysis effectively explains a portion of the disparity using currently collected variables. It identifies crucial mediators and confounders, highlighting the significance of variables such as age at diagnosis, insurance status, tumor grades, and the concentration of Naphtha in the air.

[1] Qingzhao Yu, Professor, Biostatistics and Data Science, Louisiana State University Health-New Orleans, USA, qyu@lsuhsc.edu

[2] Bin Li, Professor, Department of Experimental Statistics, Louisiana State University, Baton Rouge, USA, bli@lsu.edu

Bayesian Analysis for Over-parameterized Linear Model via Effective Spectra

Tomoya Wakayama

Abstract

In the field of high-dimensional Bayesian statistics, a variety of methodologies have been developed, including various prior distributions that result in parameter sparsity. However, such priors exhibit limitations in handling the spectral structure of data, rendering estimation less effective for analyzing high-dimensional linear models that lack sparsity. This chapter introduces a Bayesian approach that employs a prior distribution dependent on the eigenvectors of data covariance matrices. We also provide contraction rates of the derived posterior distribution and develop a truncated Gaussian approximation of the posterior distribution. The former demonstrates the efficiency of posterior estimation, whereas the latter facilitates the uncertainty quantification of parameters using a Bernstein–von Mises-type approach. These findings suggest that Bayesian methods capable of handling data spectra and estimating non-sparse high-dimensional parameters are feasible.

Soft labeling in semi-supervised hidden Markov models

Filip Wichrowski[1], Katarzyna Kaczmarek-Majer[1]

[1]Systems Research Insistute, Polish Academy of Sciences
01-447 Newelska 6
Warsaw, Poland

Abstract

Hidden Markov Models (HMMs) are widely used latent-variable models that provide flexible approaches to learning from sequential data. However, standard HMM training typically relies on either fully supervised or fully unsupervised methods, both of which can be suboptimal in many real-world settings where data labeling is often expensive or imperfect. A solution proposed by [1] integrates partial supervision into HMMs by modifying the forward–backward recursions, effectively constraining certain paths in the state lattice when partial labels are known. While this approach elegantly handles hard labeling constraints, it does not extend naturally to scenarios with soft or uncertain labels because the forward–backward definition must remain consistent across all time steps. In this work, we propose a novel semi-supervised HMM training strategy that relocates the burden of partial labeling from the forward–backward variables to the emission probabilities. By incorporating partial supervision directly into the emission densities, we preserve the standard HMM recursion while leveraging soft labeling information as a prior distribution. Consequently, our method naturally accommodates cases in which labels are not entirely reliable: whether due to subjective human annotation or automated labeling processes prone to error. In controlled experiments with HMM-simulated data, we examine how varying structural assumptions, particularly the extent of overlap among emission distributions, interact with soft labeling to produce different outcomes. Our empirical findings demonstrate that shifting partial supervision into the emission probabilities not only recovers the original hard-labeling results under strict constraints but also broadens HMM applicability to settings where label uncertainty must be carefully accounted for in model inference.

References

[1] Huggins-Daines D., Rudnicky A. I., A constrained baum-welch algorithm for improved phoneme segmentation and efficient training, Proc. Interspeech, 2016, ss.1205-1208.

[2] X.D. Huang, M.A. Jack, Semi-continuous hidden Markov models for speech signals, Computer Speech & Language, 1989, ss. 239-251.

[3] Rabiner L., Juang B., An introduction to hidden Markov models, IEEE ASSP Magazine, 1986, ss. 4-16.

The project *ExplainMe: Explainable Artificial Intelligence for Monitoring Acoustic Features extracted from Speech (FENG.02.02-IP.05-0302/23)* is carried out within the First Team programme of the Founda- tion for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG).

Efficient Bayesian Variable Selection in Gaussian Process Regression Models

Yueheng Wu, Joran Jongerling, and Joris Mulder. Department of Methodology and Statistics, Tilburg University.

Abstract

Gaussian processes (GPs) are a flexible and mathematically elegant methodology for modeling nonlinear phenomena. When building such nonlinear models, a central question is for which predictors nonlinear GP components are necessary. This can be viewed as a nonlinear variable selection problem. Besides avoiding statistical overfitting, omitting negligible nonlinear effects is especially important due to the computational complexity of GP components. This poster aims to address this problem using an efficient Bayes factor testing approach. To facilitate the specification of a meaningful prior, a reparameterization of the GP parameters is proposed. The transformed parameters have a clear interpretation regarding the 'size' of the potential nonlinear effects (if present). Moreover, to avoid the need to compute marginal likelihoods for all possible models where GP components are included/excluded, a Savage-Dickey density approach is proposed. This enables the simultaneous computation of Bayes factors between all candidate models through a single full-model fit, dramatically reducing computational overhead. Through simulations and empirical examples, we demonstrate the performance and the behavior of the proposed model selection framework. We hope the methodology will provide researchers with a rigorous yet user-friendly and efficient tool for nonlinear variable selection problems, supporting broader applicability across various research disciplines.