Generalized additive latent variable modeling

Øystein Sørensen

Center for Lifespan Change in Brain and Cognition, Department of Psychology, University of Oslo, Norway, oystein.sorensen@psykologi.uio.no

Generalized linear mixed models (GLMMs) are an essential tool when analyzing clustered data. However, both GLMMs and their nonlinear extensions require the parametric form of nonlinear effects to be specified a priori, and this is often not practical. For example, cognitive abilities across various domains follow distinctive trajectories across the lifespan, which are not easily parametrized. Generalized additive mixed models (GAMMs) are an excellent alternative in these cases, able to flexibly adapt to the underlying nonlinear shape. With multivariate response data, however, GLMMs and GAMMs are not able to project the responses onto a lower-dimensional space of latent variables. Structural equation models are ideal for this type of modeling, but also have important limitations in terms of incorporating explanatory variables, modeling multilevel data, and analyzing repeated measures data with irregular time intervals. Generalized linear latent and mixed models (GLLAMMs)[2] combine the best of both worlds by allowing latent variable modeling combined with the flexibility GLMMs. While GLLAMMs allow nonlinear effects in the explanatory variables, they also require the parametric form of the relationships to be explicitly formulated. We have therefore developed generalized additive latent and mixed models (GALAMMs), an extension of GLLAMMs in which both the linear predictor and latent variables may depend smoothly on observed explanatory variables, as in GAMMs. Utilizing the mixed model view of smoothing, we show that any GALAMM can be represented as a GLLAMM, with smoothing parameters estimated by maximum likelihood. This allows fitting GALAMMs using a profile likelihood approach developed for GLLAMMs[1]. We further show how standard errors and confidence bands of the estimated smooth functions can be computed, extending upon existing methods for GAMMs. The application motivating the development concerns how level and change of human cognitive function is correlated across cognitive domains, and how environmental and genetic factors affect the lifespan trajectories, and we show results of analyses using GALAMMs on a dataset with repeated high-dimensional measures relating to various cognitive domains in a dataset with 1850 participants between 6 and 93 years. The methods are implemented in the R package 'galamm', which will also briefly be demonstrated.

References

[1] Jeon, M. and Rabe-Hesketh, S. (2012). Profile-Likelihood Approach for Estimating Generalized Linear Mixed Models With Factor Structures. *Journal of Educational and Behavioral Statistics*, **37**(4):518–542.

[2] Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, **69**(2):167–190.

Empirical likelihood clustering for noisy data

Dace Pētersone[1]

[1] *University of Latvia, Latvia, petersondace@gmail.com*

Clustering methods are extensively applied in various fields including finance, medicine, biology etc. Presence of noise in data is inevitable in real life applications, it can be introduced from several sources like sensor or measurement tool error or by human mistake. In general, clustering techniques can be split into five groups: hierarchical, centroids-based, distribution-based, density-based and fuzzy clustering. Common assumption of widely used clustering algorithms is noise - free input and violation of this assumption can make it hard for the algorithm to create reliable results. We use empirical likelihood ratio based clustering approach introduced in [1] to assess its performance on real-life and generated noisy data and compare it with most popular cluster analysis methods.

References

[1] Melnykov, V., Shen, G. (2013). Clustering through empirical likelihood ratio. *Computational Statistics & Data Analysis*, **62**, 1-10.

Testing many restrictions under heteroskedasticity

Stanislav Anatolyev[1] **and Mikkel Sølvsten**[2]

[1] *CERGE-EI, Czech Republic, stanislav.anatolyev@cerge-ei.cz*
[2] *University of Wisconsin, USA, soelvsten@wisc.edu*

In many models, one is willing to test hundreds or thousands of restrictions on regression coefficients, for example, implied by the absence of a particular dimension of heterogeneity. The present paper provides a tool to conduct a test of such hypotheses.

We propose a test that allows for many tested restrictions in a heteroskedastic linear regression model. The test compares the conventional F statistic to a critical value that corrects for many restrictions and conditional heteroskedasticity. The correction utilizes leave-one-out estimation to correctly center the critical value, and a novel tool – leave-*three*-out estimation – to appropriately scale the recentered critical value. Large sample properties of the test are established in an asymptotic framework where the number of tested restrictions may be fixed or may grow with the sample size and can even be proportional to the number of observations.

We show that the test is asymptotically valid and has non-trivial asymptotic power against the same local alternatives as the exact F test when the latter is valid. Simulations corroborate the relevance of these theoretical findings and suggest excellent size control in moderately small samples also under strong heteroskedasticity.

Large-scale inference of correlation among mixed-type biological traits with Phylogenetic multivariate probit models

Zhenyu Zhang[1], Akihiko Nishimura[2], Paul Bastide[3,4], Xiang Ji[5], Rebecca P. Payne[6], Philip Goulder[7,8,9], Philippe Lemey[3], and Marc A. Suchard[1]

[1] *University of California, Los Angeles, USA, zyz606@ucla.edu*
[2] *Johns Hopkins University, USA*
[3] *KU Leuven, Belgium*
[4] *Université de Montpellier, France*
[5] *Tulane University, USA*
[6] *Newcastle University, UK*
[7] *University of Oxford, UK*
[8] *University of KwaZulu-Natal, South Africa*
[9] *Ragon Institute of MGH, MIT, and Harvard, USA*

Inferring concerted changes among biological traits along an evolutionary history remains an important yet challenging problem. Besides adjusting for spurious correlation induced from the shared history, the task also requires sufficient flexibility and computational efficiency to incorporate multiple continuous and discrete traits as data size increases. To accomplish this, we jointly model mixed-type traits by assuming latent parameters for binary outcome dimensions at the tips of an unknown tree informed by molecular sequences. This gives rise to a phylogenetic multivariate probit model. With large sample sizes, posterior computation under this model is problematic, as it requires repeated sampling from a high-dimensional truncated normal distribution. Current best practices employ multiple-try rejection sampling that suffers from slow-mixing and a computational cost that scales quadratically in sample size. We develop a new inference approach that exploits 1) the bouncy particle sampler (BPS) based on piecewise deterministic Markov processes to simultaneously sample all truncated normal dimensions, and 2) novel dynamic programming that reduces the cost of likelihood and gradient evaluations for BPS to linear in sample size. In an application with 535 HIV viruses and 24 traits that necessitates sampling from a 12,840-dimensional truncated normal, our method makes it possible to estimate the across-trait correlation and detect factors that affect the pathogen's capacity to cause disease. This inference framework is also applicable to a broader class of covariance structures beyond comparative biology.