Abstracts of the Ninth Workshop on Simulation (2018) 1-2

Comparing non-parametric bootstrap and subsampling batch means methods for confidence assessment of ranked list latent variable estimates

Vendula Švendová[1], Sereina A. Herzog[1], Michael G. Schimek[1]

1 Introduction

Data pooling, also called data fusion or meta analysis, is a standard approach for merging data in order to obtain robust results. A large number of methods use ranking as a means of data representation, as ranks have the advantage of being independent of data type, scale, normalisation or other transformations. One such pooling method which estimates the underlying signals in multiple ranked lists is proposed in [1]. This method takes lists of rankings and recovers the latent variables responsible for the observed ranks. The parameter optimisation is performed with a Markov chain Monte Carlo (MCMC) algorithm, using 10 independent chains. Naturally, there is a need for a certainty assessment of such estimated values. In [1], the standard errors (SEs) of the estimates are calculated via non-parametric bootstrap. Each bootstrap sample $b = 1, \ldots, B$ of the observed ranked lists requires an initialisation of 10 additional Markov chains, which has a dramatic impact on the algorithm's computational demand, with runtime being in the order of hours even for small datasets. In this work, the objective is to estimate the SEs directly from the 10 original chains available from the latent variable estimation, and hence speed up the SE calculation B times. We use a subsampling approach called batch means method, as explained in [3]. The bootstrap method proposed in [1] and the batch means method proposed in this work are compared in terms of speed and quality of the SE estimates.

2 Methods

In order to compare the behavior of the bootstrap method and the batch means method for SE estimation, we simulated two scenarios, each with $n = 20$ lists and $p = 10$ objects. In the first scenario, the lists were in high agreement (Kendall's τ correlation ~ 0.7), while in the second scenario they were in mild agreement (Kendall's τ correlation ~ 0.4). We simulated the datasets and their respective true latent variables $\theta \in \mathbf{R}^p$ in the same way as in [1]. The estimates $\hat{\theta} \in \mathbf{R}^p$ were obtained using Metropolis MCMC optimisation, 10 independent chains, each with 20 000 steps. Bootstrap SE estimates for each object were calculated from

[1]Institute for Medical Informatics, Statistics and Documentation – Medical University of Graz, Austria, E-mail: vendula.svendova@medunigraz.at

$B = 50$ bootstrap samples. For every sample, we initiated 10 chains with 20 000 steps each. The details of the data simulation and bootstrap SE estimation are explained in [1]. Batch means SEs were calculated using the R package `mcmcse` [4]. We applied the batch means method on a window of w multivariate points centred around the estimate $\hat{\theta}$. The final certainty intervals were determined as $\hat{\theta}\pm$ 2*SE for both methods. Each scenario was simulated multiple times in order to confirm the results.

We compared the two methods based on the execution time and width of the certainty intervals.

3 Results

Both methods provided reliable SE estimates, succesfully covering the true signal values. With the number of bootstrap samples set to $B = 50$, the execution time of the batch means method was 50 times shorter. In our simulated scenarios with 20 lists and 10 objects, the runtime was reduced to 20.7 minutes using the batch means method, compared to 17.25 hours using the bootstrap method (benchmarked on a 4-core Intel(R) Core(TM) i7-6700 CPU@3.4GHz, running Win64bit). The width of the 2*SE intervals when using the batch means method depended heavily on the chosen window size w. Using a window size of $w = 500$, we obtained 2*SE intervals comparable to the bootstrap method, independently of the level of agreement between the lists. We conclude that the batch means method can be used as a much faster alternative to the bootstrap method.

References

[1] Švendová V., Schimek M.G. *A novel method for estimating the common signals for consensus across multiple ranked lists* / Computational Statistics & Data Analysis, 2017, v. 115, p. 122–135.

[2] Politis D.N., Romano J.P., Wolf M. *Subsampling* / Springer-Verlag New York, 1999.

[3] Flegal J.M., and Galin L.J. *Batch means and spectral variance estimators in Markov chain Monte Carlo* / The Annals of Statistics, 2010, 38.2, p. 1034–1070.

[4] Flegal J.M., Hughes J., Dootika V., Dai N. *mcmcse: Monte Carlo Standard Errors for MCMC* / R package version 1.3-2, 2017.