Random projections for Bayesian regression

Leo N. Geppert, Katja Ickstadt, Alexander Munteanu, Christian Sohler

Massive data sets with a very large number of observations n (“Big data”) are becoming more and more frequent. When conducting Bayesian regression on such data sets, the required computing time using methods like MCMC becomes infeasible. We aim to overcome this restriction by reducing the number of observations in the data set while retaining the information about the regression model.

To reduce the data set, we propose using subspace embeddings for the likelihood or both the likelihood and the prior distribution. We employ three subspace embedding techniques based on the Johnson-Lindenstrauss-Transform. The reduced data set contains $k \ll n$ observations, which consist of linear combinations of the original data set. k only depends on the number of variables d and the desired approximation error, but is independent of n, making the method especially useful for very large data sets. Bayesian regression is then conducted on the reduced data set. A variety of methods can be used to this end. We have concentrated on MCMC-methods due to their reliability.

We investigate the resulting approximation error both theoretically and empirically. On the theoretical side, the difference between the posterior distribution based on the reduced data set and the posterior distribution based on the original data set is a fraction of the latter’s parameters. This guarantees that the results of both Bayesian regressions models only differ by a small margin. We conduct a simulation study to evaluate the approximation error empirically. In accordance with the theoretical results we find only small differences and a small amount of additional variance introduced due to the reduction.

The method reduces the total running time for the analysis. Embedding the data set only takes a short amount of time (in the range of minutes even for very large data sets). As the size of the reduced data set is independent of n, the subsequent Bayesian regression takes far less running time than the same analysis on the full data set. The subspace embeddings can handle very large data sets, including data sets that do not fit into the computer’s memory.