Report on

Calibrated Data Augmentation for Scalable Markov Chain Monte Carlo

The paper points out difficulties when using Gibbs sampling in big data situations and proposes a Gibbs sampling within MCMC simulation scheme in these situations. An efficient data augmentation is developed and used as a proposal mechanism within the Metropolis-Hastings algorithm. The proposed method has some merit, however I have several concerns as follows.

- The proposed method aims to scale up applications of MCMC in big data (big n). Because of the nature of most data augmentation methods which require a loop over n data points to sample the latent z_i, it seems to me that such a framework cannot offer a truly solution for big data. The recent big-data MCMC literature has focused on using data subsampling to avoid computation of big sums over n terms. In order to make the CDA truly scalable, more needs to be done to avoid big loops over n data points in each MCMC iteration.

- As the CDA doesn't sample directly from the target posterior but an alternative $\pi_{r,b}(\theta)$, a M-H correction step is needed. It's not clear how the proposal $\pi_{r,b}(\theta)$ is guaranteed to be an efficient proposal in the M-H step.

- It seems very challenging to tune the high dimensional parameters r and b. The authors do describe a method for doing so, but the method is too problem-specific and based on quantities such as the inverse Fisher information matrix which is not always easy to obtain.