MCMC for big data

Geir Storvik

BigInsight lunch - May 2 2018

Outline

- Why ordinary MCMC is not scalable
- Different approaches for making MCMC scalable
- Summary/status

Big data and statistics

- Jordan et al. (2013), On statistics, computation and scalability:

  gatherers of large-scale data are often forced to turn to ad hoc procedures that perhaps do provide algorithmic guarantees but which may provide no statistical guarantees and which in fact may have poor or even disastrous statistical properties

- Statistical "solutions":
  - Better algorithms for "standard" optimal solutions
    - Embarrassingly parallel methods
      - Bootstrapping
      - Bagging/random forrest
    - Divide-and-conquer methods (`biglm` package in R)
    - Dynamic updating (Kalman filtering, particle filters)
    - Sub-sampling (stochastic approximation)
  - Alternative procedures that are both computationally and statistically efficient
    - Bags of Little Bootstraps (Kleiner et al., 2014)
- Bayesian (MCMC-based) methods: Left-behind!

Computing expectations

- Complicated expectations needed in many statistical inference settings
  - ML estimation with latent variables

$$L(\theta) = p(\mathbf{y}|\theta) = \int_{\mathbf{z}} p(\mathbf{y}|\mathbf{z};\theta)p(\mathbf{z}|\theta)d\mathbf{z} = E^{p(\mathbf{z}|\theta)}[p(\mathbf{y}|\mathbf{Z};\theta)]$$

  - Bayesian statistics

$$\hat{\theta} = E^{p(\theta|\mathbf{y})}[\theta|\mathbf{y}]$$

- Markov chain Monte Carlo (Bayesian setting):
  - Simulate Markov chain $\theta^1, \theta^2, \ldots$
    - Exact MCMC

$$\theta^m \overset{D}{\Rightarrow} p(\theta|\mathbf{y}) \text{ as } m \to \infty$$

$$M^{-1}\sum_{m=1}^{M} \theta^m \to E^{p(\theta|\mathbf{y})}[\theta|\mathbf{y}] \text{ as } M \to \infty$$

    - Approximate MCMC

$$\theta^m \overset{D}{\Rightarrow} \tilde{p}(\theta|\mathbf{y}) \text{ as } m \to \infty$$

$$|p(\theta|\mathbf{y}) - \tilde{p}(\theta|\mathbf{y})| \leq \varepsilon$$

Metropolis-Hastings

- Algorithm:
  - Generate $\theta^* \sim q(\cdot|\theta^{i-1})$
  - Calculate $\alpha = \min\left\{1, \frac{\pi(\theta^*|\mathbf{y})q(\theta^{i-1}|\theta^*)}{\pi(\theta^{i-1}|\mathbf{y})q(\theta^*|\theta^{i-1})}\right\}$
  - Put

$$\theta^i = \begin{cases} \theta^* & \text{with probability } \alpha; \\ \theta^{i-1} & \text{otherwise.} \end{cases}$$

- For transition density $P(\theta^*|\theta)$:

$$p(\theta|\mathbf{y})P(\theta^*|\theta) = p(\theta^*|\mathbf{y})P(\theta|\theta^*) \quad \text{Detailed balance}$$

- Calculation of α (independent case):

$$\frac{\pi(\theta^*|\mathbf{y})}{\pi(\theta^{i-1}|\mathbf{y})} = \frac{\pi(\theta^*)p(\mathbf{y}|\theta^*)}{\pi(\theta^{i-1})p(\mathbf{y}|\theta^{i-1})} \overset{ind}{=} \frac{\pi(\theta^*)\prod_{i=1}^n p(y_i|\theta^*)}{\pi(\theta^{i-1})\prod_{i=1}^n p(y_i|\theta^{i-1})}$$

- For big data: Product too time/memory-consuming

Alternatives

- Change estimator
  - Approximate Bayesian Computation (ABC)
  - Variational Bayes
- Alternative MCMC methods (Bardenet et al., 2017)
  - Divide-and-conquer methods
  - Exact sub-sampling methods
  - Approximate sub-sampling methods
  - (Methods dynamically including more data)

Divide-and-conquer metods

- Procedure
  - Split the data into a large number of smaller (possibly overlapping) data sets
  - Perform inference on each smaller data set
  - Combine the results
  - Neiswanger et al. (2013); Scott et al. (2016); Wang and Dunson (2013); Li et al. (2017); Minsker et al. (2014)
- Properties
  - Computation only on smaller datasets
  - Easy to run in parallel.
  - Separation lead to inexact results
    - Some asymptotic results available, but in the limit simple Laplace approximations better and easier (?)

Consensus Monte Carlo (Scott et al., 2016)

- Assume independent blocks $\mathbf{y}_1, \ldots, \mathbf{y}_S$:

$$p(\theta|\mathbf{y}) = \prod_{s=1}^{S} p_s(\theta|\mathbf{y}_s) \propto \prod_{s=1}^{S} p(\mathbf{y}_s|\theta)p(\theta)^{1/S}$$

- Simulate $\theta_{s1}, \ldots, \theta_{sG}$ from $p_s(\theta|\mathbf{y}) \propto p(\mathbf{y}_s|\theta)p(\theta)^{1/S}$
- Combine $\theta_g = \left(\sum_s \mathbf{W}_s\right)^{-1} \sum_s \mathbf{W}_s \theta_{sg}$
- Properties:
  - Exact if $p_s(\theta|\mathbf{y})$, $s = 1, \ldots, S$ are Gaussian ($\mathbf{W}_s = \text{Var}^{p_s(\theta|\mathbf{y})}[\theta]$)
  - Approximate in general
  - When choice of model complexity involved:
    - How does prior $p(\theta)^{1/S}$ and subset of data influence complexity?
- Alternative: $p_s(\theta|\mathbf{y}) \propto p(\mathbf{y}_s|\theta)^S p(\theta)$

Delayed acceptance (Banterle et al., 2015)

- We have $\alpha(\theta, \theta^*) = \min\{1, \rho(\theta, \theta^*)\}$ where

$$\begin{aligned}\rho(\theta, \theta^*) &= \frac{p(\theta^*)}{p(\theta)} \prod_{i=1}^{n} \frac{p(y_i|\theta^*)}{p(y_i|\theta)} \\ &= \frac{p(\theta^*)}{p(\theta)} \prod_{s=1}^{S} \frac{p(\mathbf{y}_s|\theta^*)}{p(\mathbf{y}_s|\theta)}\end{aligned}$$

- Delayed acceptance: Accept with probability

$$\prod_{s=1}^{S} \min\left\{1, \left[\frac{p(\theta^*)}{p(\theta)}\right]^{1/S} \frac{p(\mathbf{y}_s|\theta^*)}{p(\mathbf{y}_s|\theta)}\right\}$$

- Sequential procedure with only evaluating $\frac{p(\mathbf{y}_s|\theta^*)}{p(\mathbf{y}_s|\theta)}$ at each step
- Possible gain when rejecting, not when accepting

Subsampling-based

- Independent data:

$$\log p(\mathbf{y}|\theta) = \sum_{i=1}^{n} \log p(y_i|\theta)$$

- ML/Gradient methods

$$\hat{\theta}^{s+1} = \hat{\theta}^s + \gamma \nabla [\log p(\hat{\theta}^s)) + \log p(\mathbf{y}|\hat{\theta}^s)]$$

- Big data:

$$\hat{\theta}^{s+1} = \hat{\theta}^s + \gamma_s [\nabla \log p(\hat{\theta}^s)) + \widehat{\log p}(\mathbf{y}|\hat{\theta}^s)]$$

$$\widehat{\log p}(\mathbf{y}|\theta) = \frac{n}{m} \sum_{j=1}^{m} p(y_{i_j}|\theta) \qquad i_1, \ldots, i_m \text{ random subsample of } \{1, \ldots, n\}$$

- Utilising an unbiased estimate of $\log p(\mathbf{y}|\theta)$.
- Stochastic gradient descent, convergence if

$$\sum_{s=1}^{\infty} \gamma_s = \infty, \quad \sum_{s=1}^{\infty} \gamma_s < \infty^2$$

Pseudo-likelihood approach

- Idea: Replace α by

$$\hat{\alpha} = \min\left\{1, \frac{\hat{\pi}(\theta^*|\mathbf{y})q(\theta^{i-1}|\theta^*)}{\hat{\pi}(\theta^{i-1}|\mathbf{y})q(\theta^*|\theta^{i-1})}\right\}$$

- If $E[\hat{\pi}(\theta|\mathbf{y})] = \pi(\theta|\mathbf{y})$ and positive:
  - convergence properties are preserved! (Beaumont, 2003; Andrieu et al., 2009)
- Question: How to construct $\hat{\pi}(\theta|\mathbf{y})$?
- Problem with subsampling $\widehat{p}(\mathbf{y}|\theta) = \exp(\widehat{\log p}(\mathbf{y}|\theta))$ is a biased estimate of $p(\mathbf{y}|\theta)$.
- Jacob et al. (2015): Without additional knowledge on $\widehat{\log p}(\mathbf{y}|\theta)$ we cannot obtain positive, unbiased estimates $p(\mathbf{y}|\theta)$!

Firefly MCMC (Maclaurin and Adams, 2014)

- $p(\theta|\mathbf{y}) \propto p(\theta) \prod_i p(y_i|\theta)$ can be extended to

$$p(\theta, \mathbf{z}|\mathbf{y}) \propto p(\theta) \prod_i p(y_i|\theta) \prod_i \left[\frac{p(y_i|\theta) - B_i(\theta)}{p(y_i|\theta)}\right]^{z_i} \left[\frac{B_i(\theta)}{p(y_i|\theta)}\right]^{1-z_i}$$

  where $z_i \in \{0, 1\}$ and $0 < B_i(\theta) \leq p(y_i|\theta)$.

- $p(\theta, \mathbf{z}|\mathbf{y})$ has $p(\theta|\mathbf{y})$ as marginal
- Simulation:

$$p(\theta|\mathbf{y}, \mathbf{z}) \propto \prod_i [p(y_i|\theta) - B_i(\theta)]^{z_i} [B_i(\theta)]^{1-z_i}$$

  Only require evaluation of $p(y_i|\theta)$ for $z_i = 1$!

$$p(\mathbf{z}|\mathbf{y}, \theta) \propto \prod_i \left[\frac{p(y_i|\theta) - B_i(\theta)}{p(y_i|\theta)}\right]^{z_i} \left[\frac{B_i(\theta)}{p(y_i|\theta)}\right]^{1-z_i}$$

  Simple binomial sampling

- Main benefit if $B_i(\theta) \approx p(y_i|\theta)$ and simple to calculate
- Enough to resample a (small) fraction of z_i's at each iteration.

Stochastic Gradient Langevin Dynamics (Welling and Teh, 2001)

- Stochastic optimisation (convergence towards mode):

$$\theta_{t+1} = \theta_t + \frac{\varepsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{n}{m}\sum_{i=1}^{m} \nabla \log p(y_{ti}|\theta)\right)$$

  Require $\sum_{t=1}^{\infty} \varepsilon_t = \infty, \sum_{t=1}^{\infty} \varepsilon_t^2 < \infty$

- Langevin dynamics (convergence towards posterior distribution)

$$\theta_{t+1} = \theta_t + \frac{\varepsilon}{2}\left(\nabla \log p(\theta_t) + \sum_{i=1}^{n} \nabla \log p(y_i|\theta)\right) + \eta_t, \quad \eta_t \sim N(0, \varepsilon)$$

- Stochastic gradient Langevin:

$$\theta_{t+1} = \theta_t + \frac{\varepsilon_t}{2}\left(\nabla \log p(\theta_t) + \frac{n}{m}\sum_{i=1}^{m} \nabla \log p(y_{ti}|\theta)\right) + \eta_t, \quad \eta_t \sim N(0, \varepsilon)$$

  Require $\sum_{t=1}^{\infty} \varepsilon_t = \infty, \sum_{t=1}^{\infty} \varepsilon_t^2 < \infty$

Noisy MCMC

- Simplifying notation: $\pi(\theta|\mathbf{y}) = \pi(\theta)$
- "Standard" MCMC: $\sup_{\theta_0} ||\delta_{\theta_0} P^s - \pi||_{TV} \leq C\rho^s$
- What if we use $\widehat{P}$ instead of P where $\pi\widehat{P} \neq \pi$?
- Mitrophanov (2005); Alquier et al. (2016):

$$||\delta_{\theta_0} P^s - \delta_{\theta_0} \widehat{P}^s||_{TV} \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho}\right) ||P - \widehat{P}||_{TV}$$

where

$$\lambda = \lceil \frac{\log(1/C)}{\log(\rho)} \rceil$$

Note: Independent of s!

Noisy Metropolis Hastings

- Draw $\mathbf{U} \sim F(\mathbf{u}|\theta')$
- Apply acceptance rate $\hat{\alpha}(\theta, \theta', \mathbf{u}) \approx \alpha(\theta, \theta')$ within M-H.
- Properties:
  - Assume

    $$E^{F(\mathbf{u}'|\theta')}|\hat{\alpha}(\theta, \theta', \mathbf{u}') - \alpha(\theta, \theta')| \leq \delta(\theta, \theta')$$

    Then

    $$|\delta_{\theta_0} P^s - \delta_{\theta_0} \widehat{P}^s| \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho}\right) \sum_\theta \int q(\theta'|\theta)\delta(\theta, \theta')d\theta'$$

- Examples:
  - Ignoring discretisation error in Langevin Dynamics
  - Using pseudo-likelihoods within Gibbs random fields

Summary

- Many interesting approaches
  - Some are exact but can have slow convergence
  - Some are approximate, difficult to evaluate performance
  - (Most success perhaps gained through use of sequential Monte Carlo as well!)
- "Standard" accelerating MCMC approaches also useful:
  - Simulated tempering
  - Adaptive MCMC
  - Multiple-try MCMC
  - Rao-Blackwellisation
- More research needed!

References

P. Alquier, N. Friel, R. Everitt, and A. Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47, 2016.

C. Andrieu, G. O. Roberts, et al. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.

M. Banterle, C. Grazian, A. Lee, and C. P. Robert. Accelerating metropolis-hastings algorithms by delayed acceptance. *arXiv preprint arXiv:1503.00996*, 2015.

R. Bardenet, A. Doucet, and C. Holmes. On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

M. A. Beaumont. Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164(3):1139–1160, 2003.

P. E. Jacob, A. H. Thiery, et al. On nonnegative unbiased estimators. *The Annals of Statistics*, 43(2):769–784, 2015.

M. I. Jordan et al. On statistics, computation and scalability. *Bernoulli*, 19(4):1378–1390, 2013.

A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816, 2014.

C. Li, S. Srivastava, and D. B. Dunson. Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104(3):665–680, 2017.

D. Maclaurin and R. P. Adams. Firefly Monte Carlo: Exact MCMC with Subsets of Data. In *UAI*, pages 543–552, 2014.

S. Minsker, S. Srivastava, L. Lin, and D. Dunson. Scalable and robust Bayesian inference via the median posterior. In *International Conference on Machine Learning*, pages 1656–1664, 2014.

A. Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

W. Neiswanger, C. Wang, and E. Xing. Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*, 2013.

S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch. Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

X. Wang and D. B. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

M. Welling and Y. W. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 554–561. Morgan Kaufmann Publishers Inc., 2001.