A One-Pass Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets

Suhrid Balakrishnan[†]

David Madigan[‡]

[†]Department of Computer Science

[‡]Department of Statistics

Rutgers University

Piscataway, NJ 08854

August 28, 2004

Practicalities of Bayesian Analysis

- In all but trivial cases, analytical posterior unavailable.
- Sequential setup is appealing, but most priors are not conjugate.
- Approximations (Normal/Laplace) may not be feasible.

Practicalities of Bayesian Analysis

- In all but trivial cases, analytical posterior unavailable.
- Sequential setup is appealing, but most priors are not conjugate.
- Approximations (Normal/Laplace) may not be feasible.
- MCMC is typically employed. However, MCMC needs to lap repeatedly through the dataset (#laps $\geq$ length of the chain).

Practicalities of Bayesian Analysis

- In all but trivial cases, analytical posterior unavailable.
- Sequential setup is appealing, but most priors are not conjugate.
- Approximations (Normal/Laplace) may not be feasible.
- MCMC is typically employed. However, MCMC needs to lap repeatedly through the dataset (#laps $\geq$ length of the chain).
- What if your dataset is too large for this to be feasible?

Problem formulation

- Goal: to compute the expected value of $h(\theta)$

$$E(h(\theta)|x_1, \ldots, x_N) = \int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta \qquad (1)$$

- $f(\theta|\mathbf{x})$ is the posterior density of the parameters given the observed data $\mathbf{x} = x_1, \ldots, x_N$.

The Monte Carlo approximation for this expected value, based on M samples from the posterior, $\theta_1, \ldots, \theta_M$ would be $\frac{1}{M} \sum_{i=1}^{M} h(\theta_i)$.

Problem formulation

- Goal: to compute the expected value of $h(\theta)$

$$E(h(\theta)|x_1, \ldots, x_N) = \int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta \qquad (1)$$

- $f(\theta|\mathbf{x})$ is the posterior density of the parameters given the observed data $\mathbf{x} = x_1, \ldots, x_N$.

The Monte Carlo approximation for this expected value, based on M samples from the posterior, $\theta_1, \ldots, \theta_M$ would be $\frac{1}{M} \sum_{i=1}^{M} h(\theta_i)$.

- However massive data (and model complexity) make it hard to sample from $f(\theta|\mathbf{x})$.

Problem formulation

- Goal: to compute the expected value of $h(\theta)$

$$E(h(\theta)|x_1, \ldots, x_N) = \int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta \qquad (1)$$

- $f(\theta|\mathbf{x})$ is the posterior density of the parameters given the observed data $\mathbf{x} = x_1, \ldots, x_N$.

The Monte Carlo approximation for this expected value, based on M samples from the posterior, $\theta_1, \ldots, \theta_M$ would be $\frac{1}{M}\sum_{i=1}^{M} h(\theta_i)$.

- However massive data (and model complexity) make it hard to sample from $f(\theta|\mathbf{x})$.
- Main Ideas: Use importance sampling, set up problem in a data sequential manner, i.e. particle filtering.
  [Ridgeway and Madigan, 2002, Chopin, 2002a]

Importance sampling

Cannot sample from "target density," $f(\theta|\mathbf{x})$, but can from a "sampling density," $g(\theta)$. Then :

$$\int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta = \int h(\theta) \frac{f(\theta|\mathbf{x})}{g(\theta)} g(\theta) d\theta \quad (2)$$

$$= \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} w_i h(\theta_i) \quad (3)$$

θ_i is a draw from $g(\theta)$ and $w_i = f(\theta_i|\mathbf{x})/g(\theta_i)$. Since the expected value of w_i under $g(\theta)$ is 1, we need only compute weights up to a constant of proportionality and then normalize:

$$\int h(\theta) f(\theta|x_1, \ldots, x_N) d\theta = \lim_{M \to \infty} \frac{\sum_{i=1}^{M} w_i h(\theta_i)}{\sum_{i=1}^{M} w_i} \quad (4)$$

Sequential formulation

Let the sampling distribution $g(\theta) = f(\theta|x_1, \ldots, x_n)$ where $n \ll N$.

Note: We are partitioning the dataset into two disjoint pieces, a manageable portion $D_{1:n} = x_1, \ldots, x_n$ and the remainder of the data, $D_{n+1:N} = x_{n+1}, \ldots, x_N$.

The importance weights simplify (to the likelihood of the observations evaluated at each particle):

$$
\begin{aligned}
w_i = f(\theta_i|\mathbf{x})/g(\theta_i) &= f(\theta_i|D_{1:N})/f(\theta_i|D_{1:n}) && (5)\\
&= \frac{f(D_{1:N}|\theta_i)f(\theta_i)}{f(D_{1:N})}\frac{f(D_{1:n})}{f(D_{1:n}|\theta_i)f(\theta_i)} && (6)\\
&\propto f(D_{n+1:N}|\theta_i) = \prod_{x_j \in D_{n+1:N}} f(x_j|\theta_i) && (7)
\end{aligned}
$$

Formulation contd.

1. Load as much data into memory as possible to form $D_{1:n}$
2. Draw M times from $f(\theta|D_{1:n})$ via Monte Carlo or Markov chain Monte Carlo
3. Iterate through the remaining observations (those that comprise $D_{n+1:N}$). For each observation, x_j, update the log-weights on all of the draws from $f(\theta|D_{1:n})$. Set $j = n + 1$.
   While $j < N$

   for i in $1, \ldots, M$ do $w_i \leftarrow w_i \times f(x_j|\theta_i)$

Formulation contd.

1. Load as much data into memory as possible to form $D_{1:n}$
2. Draw M times from $f(\theta|D_{1:n})$ via Monte Carlo or Markov chain Monte Carlo
3. Iterate through the remaining observations (those that comprise $D_{n+1:N}$). For each observation, x_j, update the log-weights on all of the draws from $f(\theta|D_{1:n})$. Set $j = n + 1$.
   While $j < N$

   for i in $1, \ldots, M$ do $w_i \leftarrow w_i \times f(x_j|\theta_i)$

- Are we done?

Degeneracy

- If all we do is re-weight existing particles, sample degeneracy quickly becomes an issue.

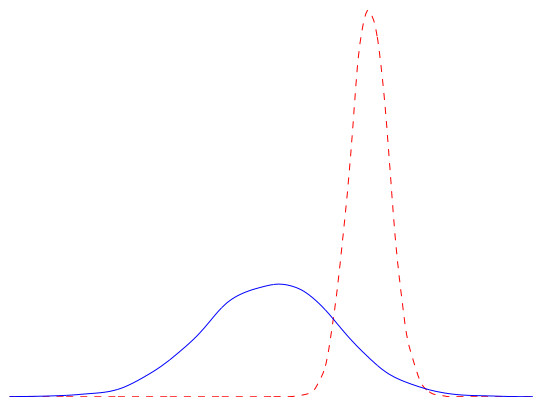
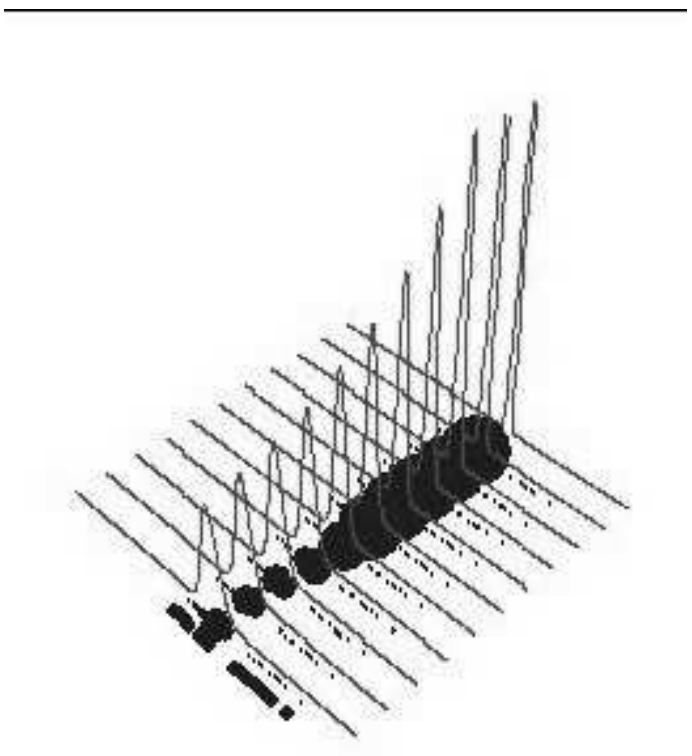


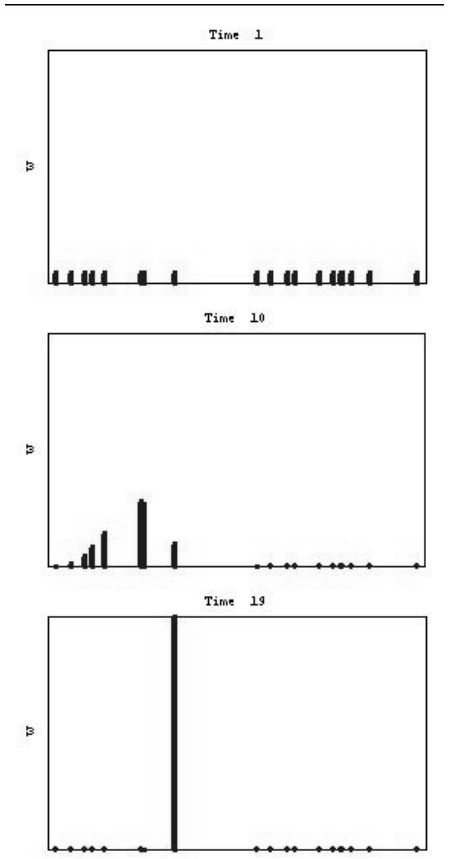
Figure 1: Comparison of $f(\theta|D_{1:n}, D_{n+1:N})$ (dashed) and $f(\theta|D_{1:n})$ (solid)

$$Var(\theta|D_{1:n}) = E(Var(\theta|D_{1:n}, D_{n+1:N})) + Var(E(\theta|D_{1:n}, D_{n+1:N}))$$

Further illustrating degeneracy

- Images from "Tutorial on Particle filters" by Keith Copsey





Sequential Monte Carlo Ideas

- Fight sample degeneracy by resampling [Gordon et al., 1993, Kitagawa, 1996] and rejuvenating particles (we will make precise exactly when a little later...) using a move step [Gilks and Berzuini, 2001]. This is a single Metropolis-Hastings step, **conditioned on all the data seen thus far**, $\mathbf{x}'$.

1. Draw a proposal θ' from $q(\theta|\theta^{i-1})$,
2. Compute the acceptance probability

$$\alpha(\theta', \theta^{i-1}) = \min\left(1, \frac{f(\theta'|\mathbf{x}')q(\theta^{i-1}|\theta')}{f(\theta^{i-1}|\mathbf{x}')q(\theta'|\theta^{i-1})}\right) \tag{8}$$

3. With probability $\alpha(\theta', \theta^{i-1})$ set $\theta^i = \theta'$.
   Otherwise set $\theta^i = \theta^{i-1}$

Remaining issues

1. When to apply this resample-move step
2. This is a very expensive step!

Remaining issues

1. When to apply this resample-move step
2. This is a very expensive step!

- 1. Monitor the Effective Sample Size (ESS) [Kong et al., 1994, Liu, 2001].

The ESS is the number of observations from a simple random sample needed to obtain an estimate with Monte Carlo variation equal to the Monte Carlo variation obtained with the M weighted draws of θ_i.

More on the ESS

Although computing the ESS depends on the quantity we are trying to estimate, the $h(\theta)$ in our case, it can be approximated as

$$\text{ESS} \approx \frac{M}{1 + \text{Var}(w)} = \frac{\left(\sum w_i\right)^2}{\sum w_i^2}. \tag{9}$$

- So to counter degeneracy all we do is monitor the approximate ESS. Whenever it falls below a certain fraction, p of M (and note that it always must be $\leq M$) say, it's time to resample-move.

More on the ESS

Although computing the ESS depends on the quantity we are trying to estimate, the $h(\theta)$ in our case, it can be approximated as

$$\text{ESS} \approx \frac{M}{1 + \text{Var}(w)} = \frac{\left(\sum w_i\right)^2}{\sum w_i^2}. \tag{9}$$

- So to counter degeneracy all we do is monitor the approximate ESS. Whenever it falls below a certain fraction, p of M (and note that it always must be $\leq M$) say, it's time to resample-move.
- Now we focus on the second issue: the computational expense of the resample-move step.

The resample-move step

- While the MH step does exactly what we want (rejuvenate the particles from the correct distribution), it necessarily needs to look at the entire dataset seen till that point.
- Intuition: In some sense, the MH step does excessive work. Our current posterior sample (albeit impoverished) is drawn from the correct distribution. If all we want is a new and diverse set of particles, drawing from a smoothed approximation to the current posterior distribution should do the trick as well.

  Work on these lines include
  [Liu and West, 2000, Stavropoulos and Titterington, 2001]

1 Pass Filtering with Shrinkage

When it is time to rejuvenate the particles, resample as usual and then sample from the approximate importance sampling posterior density $f(\theta|D_{1:n+n_1})$ given by:

$$\widehat{f}(\theta|D_{1:n+n_1}) \quad = \quad \sum_{i=1}^{M} K(\theta; \widetilde{\theta}_i, b^2 V) \tag{10}$$

where $K(\theta; s, T)$: value at θ of the kernel function (e.g. Gaussian) with mean s and variance matrix T. $\widetilde{\theta}_i$ and V are the ***shifted*** sample/particle values and the sample Monte Carlo variance respectively with b being the kernel bandwidth.

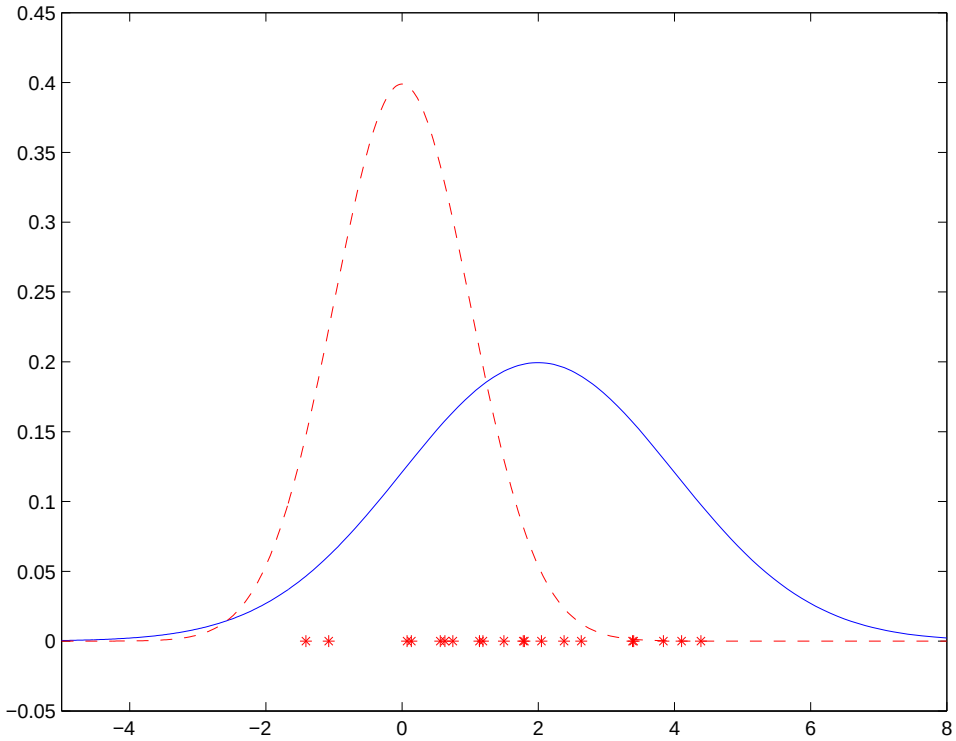
Shrinkage

- The shrinkage rule [Liu and West, 2000] specifies the shifted sample locations as:

$$\widetilde{\theta}_i = a\theta_i + (1-a)\overline{\theta} \tag{11}$$

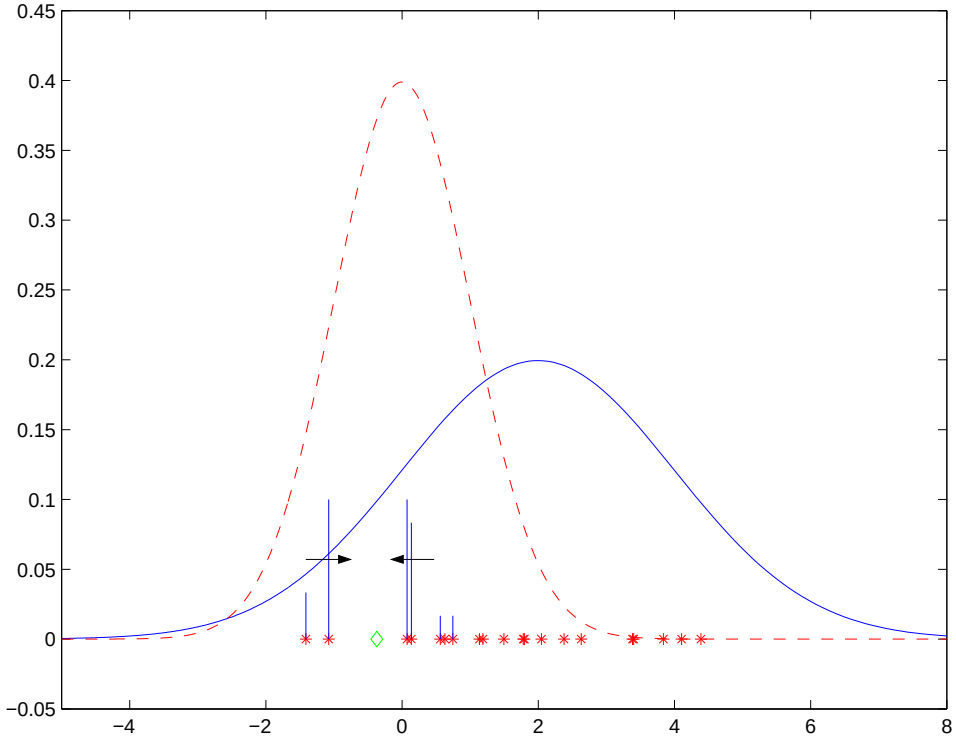
where $a = \sqrt{1-b^2}$ and $\overline{\theta}$ is the current Monte Carlo mean θ_i value. The sample drawn from the kernels placed at the shrinkage locations will now have both the correct mean and variance (the sample mean $\overline{\theta}$, and the sample variance V).

1PFS walkthrough I



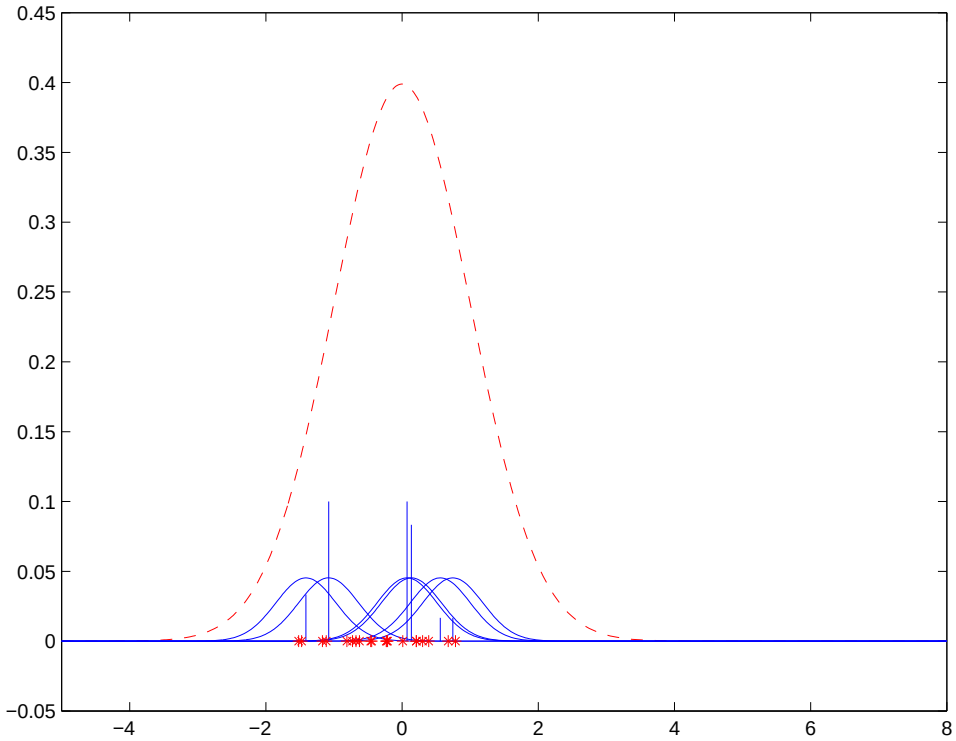
- The new resample-move step. Generate an initial sample from $f(\theta|D_{1:n})$ (the solid curve). The stars mark the particles, the sampled θ_i.

1PFS walkthrough II



- Weight based on $f(\theta|D_{1:n}, D_{n+1:N})$ (the dashed density) and resample, the length of the vertical lines indicate the number of times resampled. Shrink these locations towards $\overline{\theta}$ (the open diamond).

1PFS walkthrough III



- For each θ_i sample from the now shifted kernel density distribution and thus diversify and obtain the new sample (the stars mark these locations).

Convergence

- There exist established asymptotic Central Limit Theorems for Sequential Monte Carlo methods -
  [Del Moral and Guionnet, 1999], [Gilks and Berzuini, 2001], and [Chopin, 2002b].
- These results hold for the more general version of the problem involving unseen state variables in addition to static model parameters.
- The static parameter only case is better behaved and more tractable than the general problem [Chopin, 2002b].
- Finally, we are concerned solely with the convergence properties of the final posterior distribution estimate that our algorithm returns (not all the intermediate distributions).

Convergence contd.

- Thus, convergence is guaranteed if we can prove (asymptotically) that the samples returned by the kernel smoothing approximation to the importance sampling posterior distribution $\widehat{f}(\theta|x)$, resemble random samples from the target distribution $f(\theta|x)$.

- [Stavropoulos and Titterington, 2001] prove a restricted version of the above. Their theorem states:

  Theorem 1 *Under mild conditions, for univariate θ and the Normal kernel K, the cumulative distribution function of the values generated by the kernel approximation to the posterior distribution $\widehat{f}(\theta|x)$, converges to that of the target density, $f(\theta|x)$.*

Other convergence related observations

- The previous theorem has a multivariate generalization and can be adapted for non-Normal K as well.
- The assumptions under which it holds are essentially the same as those required by [Geweke, 1989] for the importance sample estimates to converge, with the additional requirement that the kernel functions variance should shrink to zero as the number of particles tends to infinity.

Bandwidth selection

- For Normal kernels (where $K(s,T) = \varphi(s,T)$, the Gaussian density function), kernel density estimation literature [Silverman, 1986] suggests a choice of $T = V{b_M}^2$, with

$$b_M = \left(\frac{4}{(d+2)M}\right)^{\frac{1}{d+4}} \tag{12}$$

  where d is the dimensionality of the samples.

- This choice of bandwidth is asymptotically optimal if the density being approximated is multivariate-Normal, and the samples had been obtained from this distribution.

Case Study I - Fully Bayes Logistic Regression

- The training data comprise vectors $\mathbf{x_i} = [x_{i_1}, \ldots, x_{i_d}]^T$ in $\mathbf{R^d}$ and $y_i \in \{0, 1\}$, $i = 1, \ldots, n$. We consider a model of the form:

$$p(y = 1|\mathbf{x}) = \psi(\boldsymbol{\beta}^T \mathbf{x}) \tag{13}$$

  where $\boldsymbol{\beta}$ is a vector of regression co-efficients and $\psi(\cdot)$ is the logistic link function.

- Sparse model [Tibshirani, 1995, Figueiredo, 2001]; we use an independent Laplace prior for each component of $\boldsymbol{\beta}$:

$$\pi(\beta_i|\gamma) = \frac{1}{2}\sqrt{\gamma}e^{-\gamma|\beta_i|}, \lambda > 0, n = 1, \ldots, d.$$

  which results in posterior modes of zero for many parameters (simultaneous variable selection, we set $\gamma = 5$).

- Our interest here however is not in obtaining the posterior mode but rather in fully Bayesian inference for arbitrary characteristics of the posterior distribution of $\boldsymbol{\beta}$.
- "outpic data" comprising $N = 744,963$ customer records (57 Mb in double precision). Telecommunications company data. The binary response variable identifies customers who have switched to a competitor. 7 predictor variables. 5 continuous and two 3-level categorical variables (total 10 parameters for regression).
- The dataset is small enough that regular MCMC to compute $f(\boldsymbol{\beta}|D_{1:N})$, while cumbersome, is still feasible. We also used MCMC to generate the initial particles from $f(\boldsymbol{\beta}|D_{1:n})$. In both cases we used a straightforward Metropolis-within-Gibbs sampler.

IPFS implementation details

- 1PFS implemented using Gaussian kernel function. The optimal Gaussian bandwidth formula (formula 12) defines the kernel bandwidth.
- Conditioning on the first 10,000 observations (i.e., $n = 10,000$), we generated 25,000 initial particles using the MCMC algorithm, dropping the first 5,000.
- Thus we accessed each of the first 10,000 obervations 25,000 times. 1PFS executed a rejuvenation step whenever ESS dropped below 10,000 (occurred 51 times).

IPFS implementation details

- 1PFS implemented using Gaussian kernel function. The optimal Gaussian bandwidth formula (formula 12) defines the kernel bandwidth.
- Conditioning on the first 10,000 observations (i.e., $n = 10,000$), we generated 25,000 initial particles using the MCMC algorithm, dropping the first 5,000.
- Thus we accessed each of the first 10,000 obervations 25,000 times. 1PFS executed a rejuvenation step whenever ESS dropped below 10,000 (occurred 51 times).
- **Results**

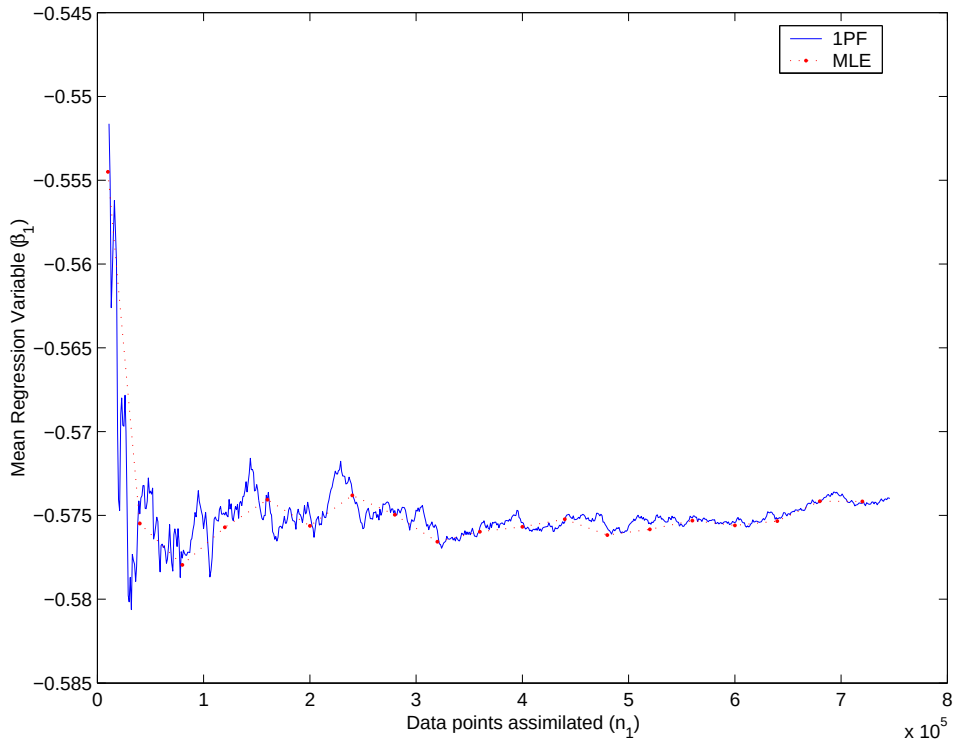
MLE: the logistic regression model parameters fit using maximum likelihood and MCMC: MCMC run on the entire dataset.

MLE	-0.574	0.155	0.056	0.220	-0.087	0.361	-0.358	-0.204	0.079	0.079
MCMC	-0.574	0.155	0.056	0.220	-0.087	0.360	-0.358	-0.204	0.080	0.078
1PFS	-0.574	0.156	0.056	0.221	-0.087	0.360	-0.357	-0.204	0.079	0.079

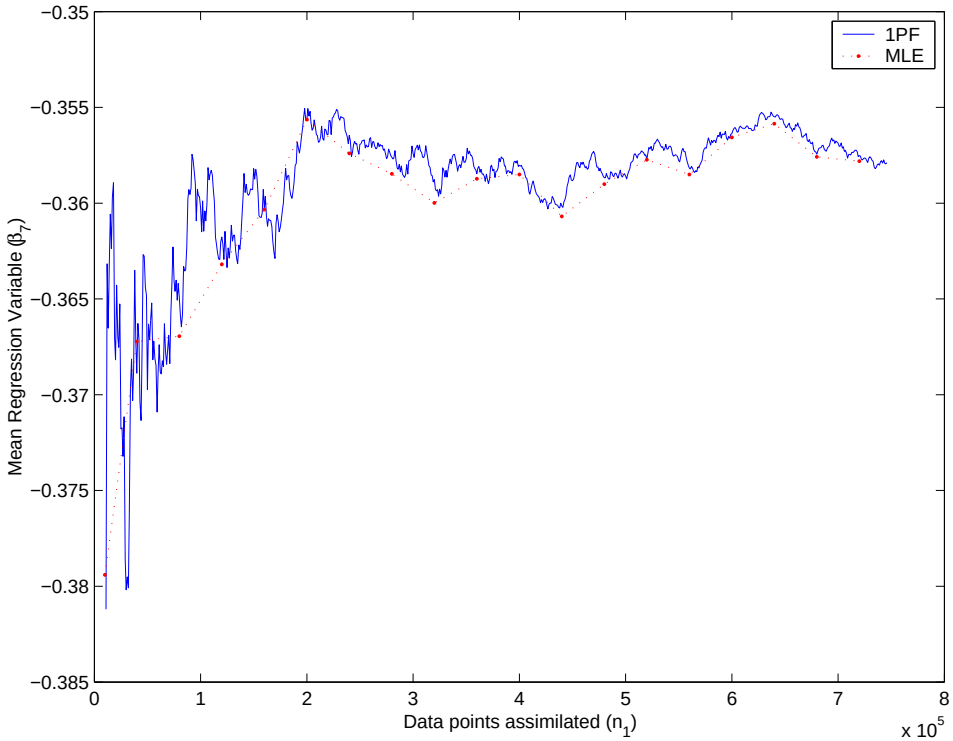
Table 1: Mean $\boldsymbol{\beta}$ estimates obtained from Bayesian logistic regression analysis of the outpic data.

Algorithm	first 10,000	next 734,963
MCMC	2.5×10^8	1.8×10^{10}
R&M	2.5×10^8	2.4×10^6
1PFS	2.5×10^8	7.3×10^5

Table 2: Total number of data accesses for MCMC, R&M (Ridgeway and Madigan's Particle Filter), and 1PFS.



- Plot showing the representative mean posterior parameter values of $\beta(1)$ determined via 1PFS as a function of the amount of data processed. Also shown on the plot is the corresponding MLE for the same amount of data.



- Plot showing the representative mean posterior parameter values of $\beta(7)$ determined via 1PFS as a function of the amount of data processed. Also shown on the plot is the corresponding MLE for the same amount of data.

Case Study II - Mixtures of Transition Models

- Mixtures of first-order transition model:
  [Cadez et al., 2000, Ridgeway, 1997, Ramoni et al., 2002].
- Data comprise N state sequences of random length generated by one of C $S \times S$ transition matrices.
- Unknowns: the C transition matrices, the mixing vector of length C, $P_1, \ldots, P_C$ and the N cluster assignments, $z_j \in \{1, \ldots, C\}, j = 1, \ldots, N$. We assume that both C and S are fixed.
- Ridgeway describe a Gibbs sampler that generates draws from this posterior distribution. Two scans of the entire dataset are required per iteration.

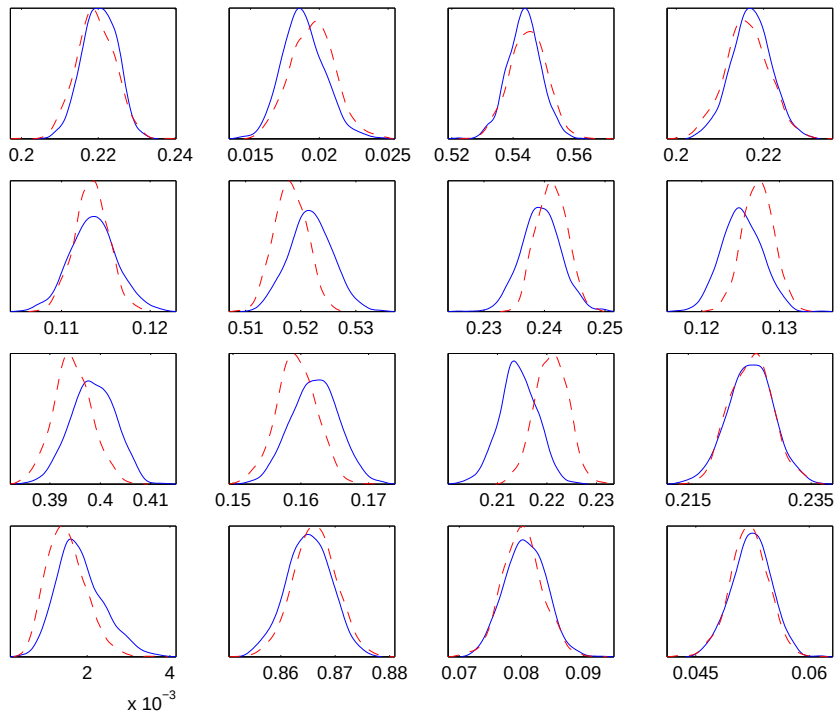
Dataset; 1PFS implementation

- We generate 1 million sequences of length between 5 and 20 from two 4×4 transition matrices. We used the first $n = 1000$ sequences to obtain the initial sample of $M = 1000$ particles. We execute a rejunevation step each time ESS drops below 100.
- The rows of the transition matrices and the vector of mixing proportions must sum to one, therefore we chose a Dirichlet kernel function $K(.,.)$. Following [Aitchison and Lauder, 1985], we choose the bandwidth b that maximizes the pseudo-likelihood (the average leave-one-out cross validation approximated likelihood).
- One extra detail. The shrinkage rule requires parametrization of the kernel $K(\widetilde{\theta}_i, b^2 V)$ in terms of its mean $\widetilde{\theta}_i$ and variance $b^2 V$. Unfortunately, starting from a mean and variance, a

closed-form expression for the corresponding Dirichlet distribution Dirichlet(α) does not exist.

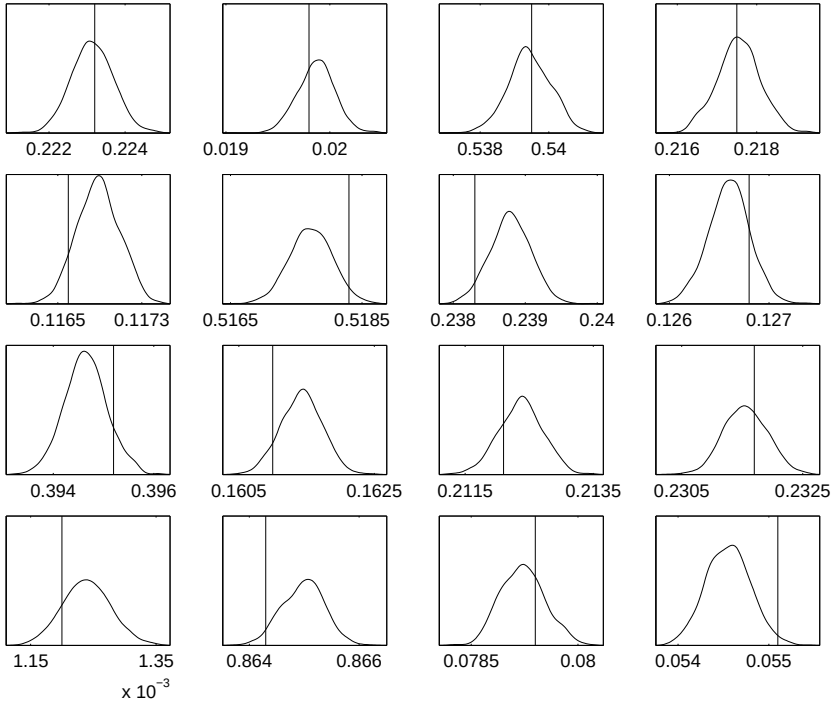
- As per [Ronning, 1989] we compute an approximation to α by matching first and second moments.
- Aside from the first 1000 observations, 1PFS accesses each of the remaining observations once. Represents a substantial computational savings as compared to the Ridgeway scheme. A Gibbs sampler, conditioned on the entire dataset, would need to access each observation 2000 times.

Comparison to Gibbs sampler



- The posterior distribution of the transition probabilities for one of the transition matrices for the first 10,000 observations. MCMC posterior: blue solid line; 1PFS: red dashed line.

Comparison to ground truth



- The posterior distribution of the transition probabilities for one of the transition matrices. 1PFS generated these densities. The vertical line marks the true value used to simulate the dataset.

Concluding Remarks

- The number of data accesses are reduced but high-dimensional data may be problematic.
- The fully Bayes approach is only one way to approach this problem; posterior mode estimation/approximation represents a viable alternative.
- Future work: application to more complex examples - streaming financial data or to text. Perhaps explore issues with state-space models.

References

[Aitchison, 1986] Aitchison, J.:1986, The Statistical Analysis of Compositional Data. New York: Chapman Hall.

[Aitchison and Lauder, 1985] Aitchison, J. and Lauder I.J.: 1985, Kernel Density Estimation for Compositional Data. *Applied Statistics*, **34** No 2, 129-137.

[Andrieu et al., 2003] Andrieu, C., N. de Freitas, A. Doucet, and M. I. Jordan: 2003, An Introduction to MCMC for Machine Learning . *Machine Learning*, **50(1/2)**.

[Besag et al., 1995] Besag, J., P. Green, D. Higdon, and K. Mengersen: 1995, Bayesian computation and stochastic systems (with Discussion) . *Statistical Science*, **10**, 3-41.

[Cadez et al., 2000] Cadez, I. and Heckerman, D. and Meek C. and Smyth, P. and White, S.: 2000, Visualization of navigation patterns

on a Web site using model-based clustering. Technical Report MSR-TR-00-18, Microsoft Research.

[Carlin et al., 2000] Carlin, B. and T. Louis: 2000, Bayes and Empirical Bayes Methods for Data Analysis. Boca Raton, FL: Chapman and Hall, 2nd edition.

[Chopin, 2002a] Chopin, N.: 2002, A sequential particle filter method for static models . *Biometrika*, **89(3)**, 539-552.

[Chopin, 2002b] Chopin, N.: 2002, Central Limit Theorem for Sequential Monte Carlo Methods and its Applications to Bayesian Inference, Technical Report 2002-44, CREST, Available at `http://www.crest.fr/doctravail/document/2002-44.pdf \verb.`

[DeGroot, 1970] DeGroot, M.: 1970, Optimal Statistical Decisions. New York: McGraw-Hill.

[Del Moral and Guionnet, 1999] Del Moral, P. and Guionnet, A.: 1999, A central limit theorem for nonlinear filtering using interacting particle systems.*Annals of Applied Probability*, **9**, 275-297.

[Doucet et al., 2001] Doucet, A., N. de Freitas, and N. Gordon: 2001, Sequential Monte Carlo Methods in Practice. Springer-Verlag.

[DuMouchel, 1999] DuMouchel, W.: 1999, Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system (with discussion) . *The American Statistician*, **53(3)**, 177-190.

[Figueiredo, 2001] Figueiredo, M.: 2001, Adaptive sparseness using Jeffreys prior . In: Neural Information Processing Systems - NIPS 2001.

[Figueiredo and Jain, 2001] Figueiredo, M. and A. K. Jain: 2001, Bayesian Learning of Sparse Classifiers . In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - CVPR 2001.

[Friedman et al., 1999] Friedman, N., I. Nachman, and D. Peer: 1999, Learning Bayesian Network Structures from Massive Datasets: The Sparse Candidate Algorithm . In: Proceedings of the Fifteenth Conference on Uncertainty in Articial Intelligence (UAI99). pp. 206-215.

[Gelman et al., 1995] Gelman, A., J. Carlin, H. Stern, and D. Rubin: 1995, Bayesian Data Analysis. New York: Chapman Hall.

[Geman and Geman, 1984] Geman, S. and D. Geman: 1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images . *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-741.

[Genkin et al., 2004] Genkin, A. and D.D. Lewis, and D. Madigan: 2004, Large-Scale Bayesian Logistic Regression for Text Categorization. In preparation.

[Geweke, 1989] Geweke, J.: Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, **24**, 1317-1399.

[Gilks and Berzuini, 2001] Gilks, W. and C. Berzuini: 2001, Following a moving target - Monte Carlo inference for dynamic Bayesian models . *Journal of the Royal Statistical Society B*, **63(1)**, 127-146.

[Gilks et al., 1996] Gilks, W., S. Richardson, and D. J. Spiegelhalter (eds.): 1996, Markov Chain Monte Carlo in Practice. Chapman and Hall.

[Girosi, 1998] Girosi, F.: 1998, An Equivalence Between Sparse Approximation And Support Vector Machines . *Neural Computation*, **10**, 1455-1480.

[Gordon et al., 1993] Gordon N.J., Salmond D.J., and A.F.M. Smith: 1993, Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation . *IEE-Proceedings-F*, **140**, 107-113.

[Hastings, 1970] Hastings, W. K.: 1970, Monte Carlo Sampling Methods Using Markov Chains and Their Applications . *Biometrika*, **57**, 97-109.

[Kitagawa, 1996] Kitagawa G.: 1996, Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational Graphics and Statistics*, **5**, 1-25.

[Kong et al., 1994] Kong, A., J. Liu, and W. Wong: 1994, Sequential imputation and Bayesian missing data problems . *Journal of the American Statistical Association*, **89**, 278-288.

[Le Cam and Yang, 1990] Le Cam, L. and G. Yang: 1990, Asymptotics in Statistics: Some Basic Concepts. New York: Springer-Verlag.

[Liu and West, 2000] Liu, J. and West, M. (2000). Combined parameter and state estimation in simulation-based filtering. In A. Doucet, J. F. G. De Freitas and N. J. Gordon (eds.), Sequential Monte Carlo Methods in Practice. New York: Springer-Verlag.

[Liu, 2001] Le Cam, L. and G. Yang: 1990, Monte Carlo strategies in scientific computing. New York: Springer-Verlag.

[Metropolis et al., 1953] Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller: 1953, Equa-tions of state calculations by fast computing machine . *Journal of Chemical Physics*, **21**, 1087-1091.

[Posse, 2001] Posse, C.: 2001, Hierarchical Model-based Clustering For Large Datasets . *Journal of Computational and Graphical Statistics*, **10(3)**, 464-486.

[Ridgeway, 1997] Ridgeway, G.: 1997, Finite discrete Markov process clustering. Technical Report MSR-TR-97-24, Microsoft Research.

[Ridgeway and Madigan, 2002] Ridgeway, G. and Madigan, D.: 2002, A Sequential Monte Carlo Method for Bayesian Analysis of Massive Datasets, *Journal of Knowledge Discovery and Data Mining*,**7**, 301-319.

[Ramoni et al., 2002] Ramoni, M. and Sebastiani, P. and Cohen, P.: 2002, Bayesian Clustering by Dynamics, *Machine Learning*, **47**(1), 91121.

[Ronning, 1989] Ronning, G.: 1989, Maximum likelihood estimation of dirichlet distributions, *Journal of Statistical Computation and Simulation*, **32**(4), 215-221.

[Ross, 1993] Ross, S. M.: 1993, Probability Models. Academic Press, 5th edition.

[Silverman, 1986] Silverman, B.W.: 1986, Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability. New York: Chapman Hall.

[Stavropoulos and Titterington, 2001] Stavropoulos, P. and Titterington, D.M. : 2001, Improved particle filters and smoothing, In A. Doucet, J. F. G. De Freitas and N. J. Gordon (eds.), Sequential Monte Carlo Methods in Practice. New York: Springer-Verlag.

[Tibshirani, 1995] Tibshirani, R.: 1995, Regression selection and shrinkage via the lasso . *Journal of the Royal Statistical Society, Series B*, **57**, 267-288.

[Tipping, 2001] Tipping, M. E.: 2001, Sparse Bayesian Learning and the Relevance Vector Machine . *Journal of Machine Learning Research* **1**, 211-244.

[Zhang and Oles, 2001] Zhang, T. and F. J. Oles: 2001, Text categorization based on regularized linear classification methods . *Information Retrieval* **4**, 5-31.

[Ronning, 1989] details

Specifically, the parameter values α_{isc} for the ith row of the transition matrix P_c are:

$$\alpha_{isc} = \widetilde{\theta}_{isc} \sum_s \alpha_{isc} \tag{14}$$

$$\log \sum_s \alpha_{isc} = \frac{1}{S-1} \sum_{s=1}^{S-1} \log \left(\frac{\widetilde{\theta}_{isc}(1-\widetilde{\theta}_{isc})}{b^2 V_{isc}} - 1 \right) \tag{15}$$

We model each row independently. A similar set of equations exists for the mixing vector's parameters.