Bayesian Analysis (2010) **5**, Number 3, pp. 429–450

Selection Sampling from Large Data Sets for Targeted Inference in Mixture Modeling

Ioanna Manolopoulou*, Cliburn Chan† and Mike West‡

Abstract. One of the challenges in using Markov chain Monte Carlo for model analysis in studies with very large datasets is the need to scan through the whole data at each iteration of the sampler, which can be computationally prohibitive. Several approaches have been developed to address this, typically drawing computationally manageable subsamples of the data. Here we consider the specific case where most of the data from a mixture model provides little or no information about the parameters of interest, and we aim to select subsamples such that the information extracted is most relevant. The motivating application arises in flow cytometry, where several measurements from a vast number of cells are available. Interest lies in identifying specific rare cell subtypes and characterizing them according to their corresponding markers. We present a Markov chain Monte Carlo approach where an initial subsample of the full dataset is used to guide selection sampling of a further set of observations *targeted* at a scientifically interesting, low probability region. We define a Sequential Monte Carlo strategy in which the targeted subsample is augmented sequentially as estimates improve, and introduce a stopping rule for determining the size of the targeted subsample. An example from flow cytometry illustrates the ability of the approach to increase the resolution of inferences for rare cell subtypes.

Keywords: Flow cytometry, large data sets, mixture models, rare events, resampling, selection sampling, sequential Monte Carlo

1 Introduction

Advances in technology in biological research, as in other fields, are challenging our ability to routinely analyse increasing large data sets. In the motivating application area of flow cytometry, routine assays generate multiple measurements on cell surface markers on each of tens of thousands to millions of individual cells (Chan et al. 2008). Mixture models are applied for cell subtype classification and discrimination, and specific interests often relate to characteristics of rather rare subtypes. For example, polyfunctional lymphocyte subsets that are of interest in predicting vaccine efficacy (Seder et al. 2008) may have frequencies of 0.01% or less of the total blood cell population. Markov chain Monte Carlo is a powerful tool for drawing inferences in mixture models, but its use requires computations on the full dataset at each iteration. This is a drawback in cases of very large datasets, so some approaches have been developed to focus on randomly drawn subsamples of data. For example, Ridgeway and Madigan (2003) proposed a

*Department of Statistical Science, Duke University, Durham, NC, mailto:im30@stat.duke.edu
†Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, mailto:cliburn.chan@duke.edu
‡Department of Statistical Science, Duke University, Durham, NC, mailto:mw@stat.duke.edu

 DOI:10.1214/10-BA517

two-step algorithm of drawing subsamples via Sequential Monte Carlo, and this was improved upon by Balakrishnan and Madigan (2006) by introducing a rejuvenation step based on a kernel smoothing approximation similar to Liu and West (2000). Here, the observations of interest are rare, so that random subsamples typically contain very few observations of the rare subtype, and new approaches are required.

Generally, we are interested in drawing inferences about low probability regions of sample space in mixture model analyses of large datasets. We use traditional Bayesian mixtures admitting uncertainty about the number of components (e.g. Müller et al. 1996; MacEachern 1998; Ishwaran and James 2002; Suchard et al. 2010). We focus on inferences about a low probability mixture component, or a group of several low probability components that together represent a scientifically relevant subpopulation. Our central idea is to use an initial random subsample of data in order to construct a weight function directed around the region of interest, and use this to subsequently draw a *targeted subsample* of data preferentially selected from that region of interest. This builds on traditional ideas of selection and weighted sampling (e.g. Heckman 1979; Bayarri and Berger 1998) and their application in discovery sampling (West 1994, 1996). Here the use of non-parametric Bayesian mixture models allows us to link regions in sample space with specific components of the model and naturally identify subsets of observations which are relevant to the scientific question at hand through a component-driven weight function. We implement a two-step Markov chain Monte Carlo approach that first uses the random subsample to obtain an initial posterior, then adds the targeted subsample to draw component-specific inferences. We extend the method to a Sequential Monte Carlo algorithm whereby the targeted subsample is augmented sequentially, guided by a stopping rule, to successively refine inferences on the rare subpopulation, to the extent feasible.

2 Modelling and posterior distributions

In contexts such as our motivating flow cytometry applications, Gaussian mixtures are used as flexible overall models and scientifically relevant subpopulations are identified by (typically, small) *subsets* of Gaussian components that can reflect non-Gaussianity within subpopulations (Chan et al. 2008). Hence, with no loss of generality here, we consider a Gaussian mixture for $p-$variate data, with N samples $x_i, (i = 1, \ldots, N)$, defining a data set X. The density of the mixture is

$$f(x_i|\mu, \Sigma) = \sum_{k=1}^{K} \pi_k N(x_i \,|\, \mu_k, \Sigma_k),$$

and we adopt a standard truncated Dirichlet process model to define the prior over the mixing probabilities based on some (large) upper bound K (Ishwaran and James 2002). Let

$$\theta \quad = \quad \{\alpha, \pi_{1:K}, \phi_{1:K}\}, \quad \phi_k = \{\mu_k, \Sigma_k\}. \tag{1}$$

The mixture model can be realized through the configuration indicators z_i for each observation x_i with prior $p(z_i = k \,|\, \pi) = \pi_k$, so that we obtain the standard hierarchical

model

$$(x_i \,|\, z_i = k, \phi_k) \sim N(x_i \,|\, \phi_k), \quad (\phi_k \,|\, G) \sim G, \quad (G \,|\, \alpha, G_0) \sim DP(\alpha, G_0), \tag{2}$$

where $G(\cdot)$ is an uncertain distribution function, $G_0(\cdot)$ is the prior mean of $G(\cdot)$ and $\alpha > 0$ the total mass, or precision of the DP. From the Pólya urn scheme,

$$\phi_k \,|\, \phi_1, \ldots, \phi_{k-1} \sim \frac{\alpha}{k-1+\alpha} G_0(\cdot) + \frac{1}{k-1+\alpha} \sum_{i=1}^{k-1} \delta_{\phi_i}(\cdot). \tag{3}$$

The truncated Dirichlet process prior is such that $\pi_1 = V_1$ and $\pi_k = V_k \times \prod_{i=1}^{k-1}(1 - V_i)$, $k > 1$, where $V_i \sim Be(1, \alpha)$, $i < K$ independently over i and $V_K = 1$. Prior specification for each component k is completed with a traditional normal-inverse Wishart form,

$$G_0(\mu_k, \Sigma_k) = N(\mu_k \,|\, \mu_0, t_0 \Sigma_k) IW(\Sigma_k \,|\, s_0, S_0) \tag{4}$$

and with a Gamma prior for the Dirichlet concentration parameter $\alpha \sim Ga(\eta_1, \eta_2)$. Placing a prior on α (see Ishwaran and James 2002) allows us to draw inferences about the number of mixture components through the role of α of the Pólya urn scheme as the prior number of observations in each component. Finally, we label components in decreasing weight π as an identifiability criterion to address the label-switching problem.

In general, a scientific question may define focus on a specific region of sample space. Here we take one key example, that of identifiable regions of low probability but of high scientific importance. For specificity, we assume this is defined by a low probability component of the mixture. In more general versions, this could be defined by a small set of mixture components, or other devices. Hence, we focus on targeting inferences about the parameters $\phi_{k^*} = (\mu_{k^*}, \Sigma_{k^*})$ of a low-probability component k^*.

The objective is to identify and analyze subsamples of the data which contain information about the specific subset of the parameters of interest. The key idea is to obtain a rough estimate of the low-probability component k^* based on a random subsample, which is subsequently used to draw weighted subsamples of the data that are more likely to be relevant to our analysis, providing us with higher resolution about the structure of the distribution in the region of interest.

The direct approach is to follow a two-step procedure of Markov chain Monte Carlo samplers. We use an initial, randomly drawn subsample from the data in order to obtain an estimate of the parameters, and use this estimate to draw a more informative subsample. The two subsamples are then combined in a joint Markov chain Monte Carlo sampler to provide us with a posterior distribution of ϕ_{k^*} which will be an improved estimate with respect to the total posterior based on the whole sample. Although interest specifically lies in estimating the parameters of component k^* given by μ_{k^*}, Σ_{k^*}, inference on the full set of μ, Σ is required in order to carry out the analysis.

We denote the two subsamples (random and targeted) X^R and X^T of size n^R and n^T respectively, where $n^T \ll n^R$ throughout this paper. The first is drawn randomly

from the data, whereas the second is drawn according to weights $w_i \;\; 1 \leq i \leq N$. We aim to choose the weights so that the targeted subsample is expected to be enriched in observations from component k^*; one specific choice is to take

$$w_i = N(x_i \,|\, m, S_\tau),$$

the density of the normal distribution in which m, S are estimates of μ_{k^*}, Σ_{k^*} from the initial analysis based on an initial random subsample, and $S_\tau = TST$ where $T = \text{diag}(\tau_1^{1/2}, \ldots, \tau_p^{1/2})$ is a $p \times p$ diagonal matrix based on a set of positive *variance multipliers* τ_j, $(j = 1, \ldots, p)$. These allow us to concentrate (or expand) targeted resampling differentially in different dimensions with due regard for correlation structure, if desired. We use the notations $w_i = w(x_i)$ and $w_i = w(x_i \,|\, m, \tau, S)$, the latter when the explicit dependence of the weight function on m, τ, S is to be high-lighted.

The likelihood of the data (X^R, X^T) then takes the following form. For observations in the random subsample:

$$\begin{aligned} f(x_i^R \,|\, \pi, \mu, \Sigma) &= \sum_{k=1}^{K} \pi_k N(x_i^R \,|\, \mu_k, \Sigma_k), \quad i = 1, \ldots, n^R. \\ f(x_i^R \,|\, z_i^R = k, \mu, \Sigma) &= N(x_i^R \,|\, \mu_k, \Sigma_k), \quad i = 1, \ldots, n^R, \end{aligned} \tag{5}$$

The first expression provides the likelihood of the standard mixture model, whereas the second expression represents the likelihood *conditionally* on the configuration indicator z_i^R, the component where observation x_i^R belongs.

Similarly, for observations in the targeted subsample:

$$f(x_i^T \,|\, \pi, \mu, \Sigma, m, \tau, S) = \sum_{k=1}^{K} \tilde{\pi}_k(\theta, m, \tau, S) N(x_i^T \,|\, \tilde{\mu}_k, \tilde{\Sigma}_k), \quad i = 1, \ldots, n^T, \tag{6}$$

with

$$\begin{aligned} \tilde{\pi}_k(\theta, m, \tau, S) &= \frac{\pi_k N\left(\mu_k \,|\, m, (S_\tau + \Sigma_k)\right)}{\sum_{k=1}^{K} \pi_k N\left(\mu_k \,|\, m, (S_\tau + \Sigma_k)\right)}, \\ \tilde{\Sigma}_k &= (\Sigma_k^{-1} + S_\tau^{-1})^{-1}, \\ \tilde{\mu}_k &= \tilde{\Sigma}_k(\Sigma_k^{-1}\mu_k + S_\tau^{-1} m), \end{aligned} \tag{7}$$

and

$$\begin{aligned} f(x_i^T \,|\, z_i^T = k, \mu, \Sigma, m, \tau, S) &\propto w(x_i^T \,|\, m, \tau, S) N(x_i^T \,|\, \mu_k, \Sigma_k) \\ &\propto N(x_i^T \,|\, \tilde{\mu}_k, \tilde{\Sigma}_k), \quad i = 1, \ldots, n^T. \end{aligned} \tag{8}$$

Note here that, conditional on all the parameters, targeted observations are independent, using the infinite population assumption for the weights $w(x_i)$. This means that we assume a very large number of observations within the region of non-negligible

support of the weight function $w(x)$, creating an inherent trade-off between the validity of the infinite population assumption and the focus of the targeted subsample. These can be monitored through the normalizing constant of the sampling weights $\sum_{i=1}^{N} w(x_i)$, and tuned through the parameter vector τ (see Appendix for a more detailed analysis of the role of τ).

The first Markov chain Monte Carlo sampler is a standard blocked Gibbs sampler (Ishwaran and James 2002) to simulate $p(\alpha, \pi, \mu, \Sigma \,|\, X^R)$. In order to carry out the second Markov chain Monte Carlo sampling based on the random and targeted subsamples combined, the posterior distributions of the parameters of α, π, μ, Σ have to be re-calculated to define efficient proposals. Although conjugacy for μ, Σ, π is lost, we keep the standard normal-inverse Wishart and Dirichlet priors for the mixture model.

The posterior for z for both subsamples is multinomial with probabilities

$$p(z_i = k \,|\, X^R, X^T, \pi, \mu, \Sigma) \propto \pi_k N(x_i \,|\, \mu_k, \Sigma_k). \tag{9}$$

The posterior for π has density

$$\begin{aligned} p(\pi|X^R, X^T, z, \mu, \Sigma, m, \tau, S) =& p(\pi|X^R, z^R, \mu, \Sigma) \prod_{k=1}^{K} \tilde{\pi}_k^{n_k^T} \\ =& \, p(\pi|X^R, z^R, \mu, \Sigma) \prod_{k=1}^{K} \left(\frac{\pi_k N\left(\mu_k|m, \left(S_\tau^{-1} + \Sigma_k^{-1}\right)^{-1}\right)}{\sum_{j=1}^{K} \pi_j N(\mu_j|m, \left(S_\tau^{-1} + \Sigma_j^{-1}\right)^{-1})} \right)^{n_k^T}. \end{aligned}$$

The contribution of the targeted subsample to the posterior distribution for π provides little additional information about the distribution of π when the elements S_τ are small, and becomes more significant as the elements of S_τ increase, allowing observations in the targeted subsample to belong to components other than k^*.

The posterior for α only depends on the data through V and thus will have the usual posterior distribution (see Ishwaran and James 2002)

$$\alpha \;\sim\; Ga\left(\eta_1 + K - 1, \eta_2 - \sum_{i=1}^{K-1} \log(1 - V_i)\right). \tag{10}$$

The posterior for μ_k can be calculated exactly as

$$\mu_k \,|\, X, z, \Sigma_k, m, \tau, S \sim N(m_k^\mu, S_k^\mu), \tag{11}$$

where S_k^μ, m_k^μ are given by

$$\begin{aligned} S_k^\mu &= \Sigma_k \left(1/t_0 + n_k^R + n_k^T \left(S_\tau^{-1}\Sigma_k + I\right)^{-1}\right)^{-1}, \\ m_k^\mu &= S_k^\mu \left(n_k \Sigma_k^{-1} \bar{x}_k - n_k^T \left(S_\tau^{-1}\Sigma_k + I\right)^{-1} S_\tau^{-1} m + \mu_0\right), \end{aligned} \tag{12}$$

where n_k is the total number of data points in component k and n_k^T is the number of data points in that component coming from the targeted subsample. Notice that the contribution of the targeted subsample to the posterior variance of μ_k is $n_k^T(S_\tau^{-1}\Sigma_k + I)^{-1}$, and since S is an estimate of Σ_k, this quantity has diagonal elements of the order of $n_k^T\tau_j(1+\tau_j)^{-1}$ for $j = 1, \ldots, p$; a more concentrated weight function reduces the information about μ_k.

The posterior for $\boldsymbol{\Sigma_k}$ has density

$$\begin{aligned} p(\Sigma_k \,|\, X, z, \mu_k, m, \tau, S) &\propto |\,\Sigma_k\,|^{-s_0}\,|\,\Sigma_k\,|^{-n_k^R/2} \times |\,S_\tau^{-1} + \Sigma_k^{-1}\,|^{n_k^T/2} \\ &\times \exp\Big\{ -\frac{S_0\Sigma_k^{-1}}{2} - \sum_{i=1}^{n_k} \frac{x_i^T\Sigma_k^{-1}x_i}{2} + n^R\mu_k^T\Sigma_k^{-1}\bar{x}^R - \frac{n_k^R}{2}\mu_k^T\Sigma_k^{-1}\mu_k \\ &- \frac{n_k^T}{2}\mu_k^T\left(S_\tau^{-1}\Sigma_k + I\right)^{-1}\Sigma_k^{-1}\mu_k - n_k^T\mu_k^T\left(S_\tau^{-1}\Sigma_k + I\right)^{-1}S_\tau^{-1}m \\ &- \frac{n_k^T}{2}m(\Sigma_k^{-1}S_\tau + I)^{-1}S_\tau m\Big\}. \end{aligned} \tag{13}$$

The non-standard posteriors (10) and (13) lead to the need for creative approximations to define efficient MCMC proposals, as now developed.

3 Markov chain Monte Carlo approach

We construct a MCMC sampler with stationary distribution $p(\mu, \Sigma, z, V, \alpha \,|\, X^R, X^T)$. The chain is initialized by drawing $\mu, \Sigma, z, V, \alpha$ from their priors, then iterates through the following steps.

1. Update z by generating from the posterior given in Equation (9).
2. Update V through a Metropolis-Hastings step by generating from the posterior based only on the initial random subsample,

$$p(V_k \,|\, X^R) \sim Be(1 + n_k^R, \alpha + \sum_{l=k+1}^{K} n_k^R), \tag{14}$$

with $V_K = 1$. Set $\pi_k = V_k\prod_{j=1}^{k-1}(1-V_j)$ and accept the proposed move with probability

$$\min\left(1, \prod_{i=1}^{K}\left(\frac{\tilde{\pi}_i'(\theta, m, \tau, S)}{\tilde{\pi}_i(\theta, m, \tau, S)}\right)^{n_i^T}\right).$$

Recall that $\tilde{\pi}(\theta, m, \tau, S)$ given in Equation (7) corresponds to the component probability weights in the targeted subsample.

If the targeted subsample is indeed drawn such that most of its points belong to component k^*, the acceptance probability will be ≈ 1.

3. Update α from its posterior given V given in Equation (10).

4. Update each μ_k through a Gibbs step using

$$\mu_k \,|\, X, z, \Sigma_k, m, S \sim N(m_k^{\mu}, S_k^{\mu}) \tag{15}$$

given in Equation (11).

5. For each Σ_k, we construct a proposal $q(\Sigma_k' \,|\, X^R, X^T, z, \mu)$ for a Metropolis-Hastings step using the fact that

$$f(x_i^T \,|\, X^R, z_i = k, \mu, \Sigma, m, \tau, S) = N(x_i^T \,|\, \tilde{\mu}_k, \tilde{\Sigma}_k).$$

We use the inverse transformation to obtain

$$\tilde{x}_i \,|\, z_i = k, \mu, \Sigma, m, \tau, S \sim N(\mu_k, \Sigma_k), \tag{16}$$

where

$$\begin{aligned} \tilde{x}_i^R &= x_i^R, \\ \tilde{x}_i^T &= \Sigma_k \left(\tilde{\Sigma}_k^{-1} x_i^T - S_\tau^{-1} m \right). \end{aligned} \tag{17}$$

In practice, of course, Σ_k is not known. A similar transformation of X^T can be obtained using an estimate of Σ_k, providing a proposal distribution

$$q(\Sigma_k' \,|\, X^R, X^T, z, \mu) \sim IW(W_k + S_0, n_k + s_0 + p - 1), \tag{18}$$

where

$$W_k = \sum_{z_i = k} \tilde{x_i} \tilde{x_i}^T - \frac{1}{n_k} \sum_{z_i = k} \tilde{x_i} \sum_{z_i = k} \tilde{x_i}^T. \tag{19}$$

In order to increase the variance of the proposal kernel, a discount factor may also be used.

The Markov chain Monte Carlo sampler sweeps through the updates described above, yielding estimates for the posterior distribution of the parameters of interest. However, due to the high number of parameters to be estimated and the difficulty in defining efficient proposals, the acceptance rate quickly drops to zero for targeted subsamples of moderate size.

4 Focusing on the low-probability component

The dimensionality of the problem, combined with the difficulty to construct efficient proposals, results in Markov chain Monte Carlo samplers which require very long running times in order to reach stationarity with respect to the posterior distribution. At the same time, the approach described above does not exploit the results from the initial run based on the random sample, except for extracting the estimates of μ_{k^*}, Σ_{k^*}.

We describe how the dimensionality of the problem can be greatly reduced using the posterior distribution estimates obtained from the initial Markov chain Monte Carlo simulation.

Notice that the objective is to draw inferences about a region in the sample space which has very low probability. Consequently, very few points in the initial random sample will belong to that region. On the other hand, the targeted sample will, generally, contain observations from the low-probability region. This implies that the posterior distribution of the parameters based on both the random and targeted samples (X^R, X^T)

$$p(\mu, \Sigma \,|\, X^R, X^T) \quad = \quad \sum_{z^R} p(\mu, \Sigma \,|\, X^R, X^T, z^R) \times p(z^R \,|\, X^R, X^T),$$

can be approximated as

$$p(\pi, \mu, \Sigma \,|\, X^R, X^T) \quad = \quad \sum_{z^R} \underbrace{p(\pi, \mu, \Sigma \,|\, X^R, X^T, z^R)}_{(a)} \times \underbrace{p(z^R \,|\, X^R)}_{(b)}, \tag{20}$$

using $p(z^R \,|\, X^R, X^T) \approx p(z^R \,|\, X^R)$. The approximation is reasonable since X^T is centred around a specific region, implying that the component structure of most of the sample space remains unchanged after introducing X^T, while the number of observations in z^R in the region of interest is far outnumbered by the z^Ts in that region. The approximation (a) requires integrating over a much smaller set of parameters z^T and can be calculated much more efficiently, and (b) is known from the first Markov chain Monte Carlo run, allowing us to update only z^T and draw z^R from the existing posterior samples. This de-couples the z-dependence of the random and the targeted sample, greatly reducing the dimensionality of the second analysis, since $n^T \ll n^R$. By fixing the configuration indicators of the random subsample, the component sums of X^R and updates of z^R remain unchanged.

The second Markov chain Monte Carlo is then adapted to a set of chains run for a set of samples. Each chain will provide posterior estimates for the parameters *conditionally* on a fixed draw of z^R, so that, marginally, we obtain a set of samples from the (approximate) posterior distribution shown in Equation (20). Specifically, for chains $l = 1 : L$, draw a sample from $(z, \pi, \mu, \Sigma)_l \,|\, X^R$ and apply the second sampler for each chain only on $\pi, z^T, \mu, \Sigma, \alpha \,|\, X^R, X^T, (z^R)_l$ keeping Z^R fixed, combining samples at the end. In effect, the algorithm amounts to an Importance Sampler (Doucet et al. 2001). This approach greatly reduces both the complexity of the calculations per sweep, as well as the total number of samples required in order to obtain a good approximation of the posterior distribution. However, because the posteriors $\mu_{k^*}, \Sigma_{k^*} \,|\, X^R$ and $\mu_{k^*}, \Sigma_{k^*} \,|\, X^R, X^T$ may differ substantially, the sampler still suffers from very low acceptance rates and with a moderately sized targeted subsample can fail to reach the region in parameter space of high posterior probability.

5 Sequential Monte Carlo approach

The focused approach drastically reduces the dimensionality of the algorithm, and as a result the computational complexity. However, Metropolis-Hastings updates still show low acceptance rates, since the posteriors conditional on X^R and (X^R, X^T), respectively, can be very different in the subspace of parameters related to the targeted component. In addition, the size of the targeted subsample is chosen manually rather than through an automated procedure. Both drawbacks may be addressed drawing the targeted sample through a Sequential Monte Carlo simulation rather than using a two-step procedure. A large number of random samples (particles) is used to approximate the targeted sequence of distributions, so that asymptotically this converges to the true target distribution; see Doucet et al. (2001), Chopin (2002), Carvalho et al. (2010). We consider a sequential scheme such that the targeted sample is selected B data points at a time, at each draw updating parameter estimates for a set of particles.

For each of a set of particles $j = 1 : J$, draw a sample of $(z, \pi, \mu, \Sigma) \,|\, X^R$ from the posterior distribution estimates obtained in the Markov chain Monte Carlo sampler. Then repeatedly augment the targeted subsample and mutate the parameter draws through the following steps.

For $j = 1 : J$ and for a fixed sequence of variance scaling vectors $\tau = t_1, \ldots, t_J$:

1. Select a particle u uniformly at random and set $m_j = \{\mu^j_{k^*}\}_u$ and $S_j = \{\Sigma^j_{k^*}\}_u$ where $\{\phi^j_{k^*}\}_u$ is the sample of the uth particle at step j for component k^*.
2. Draw another batch of B targeted observations X^T_j *without replacement* according to weights $w_i \propto w(x_i \,|\, m_j, t_j, S_j)$.
3. Using a fixed number of Metropolis-Hastings steps following the iterates described in the Markov chain Monte Carlo approach above, update the configuration indicators z^T, the weights π, the concentration parameter α and the component-specific parameters μ, Σ by repeating the following updates:
   (a) Update z^T through update (9), π through update (14) and α through (10).
   (b) Similar to the posterior distribution for μ given in (15), the posterior distribution of μ_k now becomes

$$\mu_k \,|\, X^R, X^T_{1:j}, z, \Sigma_k \sim N(m^\mu_k, S^\mu_k), \tag{21}$$

   where

$$\begin{aligned} S^\mu_k &= (\Sigma_k^{-1}/t_0 + n^R_k \Sigma_k^{-1} + B \sum_{i=1}^{j} \left(U_i^{-1}\Sigma_k + I\right)^{-1} \Sigma_k^{-1})^{-1} \\ m^\mu_k &= S^\mu_k \left(n_k \Sigma_k^{-1} \bar{x}_k - B \sum_{i=1}^{j} \left(U_i^{-1}\Sigma_k + I\right)^{-1} U_i^{-1} m_i + \mu_0 \right). \end{aligned} \tag{22}$$

   with $U_i = S_\tau$ at the values $S = S_i$ and $\tau = t_i$.

(c) Update each Σ_k using a proposal distribution similar to (18), replacing m, S in Equation (17) by

$$\begin{aligned} \tilde{x}_i^R &= x_i^R, \\ \tilde{x}_i^T &= \Sigma_k \left(\tilde{\Sigma}_k^{-1} x_i^T - U_{batch(i)}^{-1} m_{batch(i)} \right), \end{aligned} \tag{23}$$

where $batch(i)$ represents the batch of targeted observations that x_i^T was sampled in.

The sequential Monte Carlo algorithm described does not include a particle resampling step; since each particle has a fixed z^R, re-sampling would result in poor coverage of the z^R space. The algorithm can be classified as a Sequential Important Sampler (Gordon et al. 1993). At each draw j, the samples of μ and Σ are obtained through a Metropolis-Hastings kernel targeting the posterior $p(\mu, \Sigma \,|\, X^R, X_{1:j}^T, m_{1:j}, S_{1:j})$. In this Sequential Importance Sampling setting, asymptotically (as the number of particles becomes sufficiently large), the approximation will converge to the true posterior, since each particle (provided a reasonable number of Metropolis-Hastings updates) is indeed sampling from the posterior. A larger number of Metropolis-Hastings updates will be needed when the component structure changes through the targeted sample in order to fit the emerging components. The tuning parameter vector τ allows monitoring of dispersal of the targeted sample and also the adequacy of the infinite population sampling assumption (see Appendix 6).

Owing to the method in which the parameters m, S of the weight function are fixed at each step of the re-sampling, weight functions located around different regions of sample space may be chosen. When the low-probability component follows a mixture distribution between different regions of sample space, this will be reflected in the estimates obtained from each particle, resulting in each particle potentially corresponding to a different component. Through our adaptive algorithm, the sample space is explored flexibly and posterior estimates of the parameters are updated incrementally as the targeted subsample is augmented, allowing more efficient inferences.

This approach immediately poses the question of when to stop drawing observations for the targeted subsample. Ideally, we would like the targeted sample to contain all data points of component k^*. In order to address this, we introduce a decision rule such that the targeted sample stops being augmented when no more data points in the remaining original data show a high probability of belonging to component k^*. A natural approach to use is the Bayes Factor for that component, using for example a threshold of the order e^2; see West and Harrison (1997). In other words, we introduce an extra decision step, as follows:

5a. For each unsampled observation x_i, define the Bayes Factor

$$BF_{k^*}(x_i) = \frac{\pi_{k^*}(x_i)/(1 - \pi_{k^*}(x_i))}{\pi_{k^*}/(1 - \pi_{k^*})}$$

where $\pi_{k^*}(x_i) \propto \pi_{k^*} N(x_i \,|\, \mu_{k^*}, \Sigma_{k^*})$. Stop sampling if there are no observations such that $BF_{k^*}(x_i) > BF_{threshold}$, a specified threshold.

The calculation of the Bayes Factor is computationally demanding; as an alternative, the stopping rule may be expressed purely as a function of the weights. In other words:

5b. If there are fewer than $N_{threshold}$ unsampled observations within a $c_{threshold}$ contour of the weight function, stop.

The threshold values are important for monitoring the impact of the infinite population assumption on our inferences, representing the 'number of points carrying significant weight'. The values chosen will depend upon the number B of batched targeted observations drawn at each iteration of the sampler and the dimension of the data; for example, one may use $N_{threshold} = 3B$ and $c_{threshold} = \exp(-p/4)$, where p is the number of markers.

The Sequential Monte Carlo approach provides an efficient method of drawing inferences about parameters relevant to a low probability region of sample space, at the same time allowing the algorithm to automatically monitor the number of observations in the region of interest.

5.1 Example: Synthetic dataset

We implement our methods on a 2-dimensional synthetic dataset with 5,000 observations from a Gaussian mixture model with 5 components. This serves to illustrate the approach in a synthetic context where the relatively small sample size allows comparison with posterior computations based on the full data set. The model structure, shown in Figure 1, was chosen so that the component of interest can be visually separated from the rest, at the same time having significant overlap with the remaining components. The component k^* is defined as centered within a certain region, and closest to its estimate of the mean after the initial samples from the posterior distribution based on a random subsample of size 700. If no such component exists, the weight function is initialized with $m = m_0$, the centre of the pre-specified region, and $S = S_0/s_0$, the prior mode of Σ. Using our sequential Monte Carlo approach, we draw $B = 10$ observations at a time, fixing $\tau_j = 1$, $j = 1, \ldots, p$. We define the stopping rule using $N_{threshold} = 20$ and $c_{threshold} = \exp(-0.5)$, resulting in 200 targeted obervations. We carry out a standard Gibbs sampler on the full dataset, and compare the results between the two methods.

The simulation results show that the posterior distribution conditional on a targeted subsample of 900 observations, in addition to the random sample of 700 initial observations, shows an improvement of the posterior estimates comparable to those obtained using the complete dataset, but using less than 20% of the total observations; this by far exceeds the expected improvement using a random subsample of the same size, providing a much more efficient algorithm. For example, the boxplot shown in Figure 1 representing posterior inferences on the first dimension of the mean vector of component k^* is centered closer to the true mean with a much tighter posterior variance. As expected, the posterior variance is still lower using the full data set: this is both because there are more data in component k^* (the stopping rule will always leave some data

out in order to monitor the infinite popultion assumption), and because it carries more information about the component weights π.

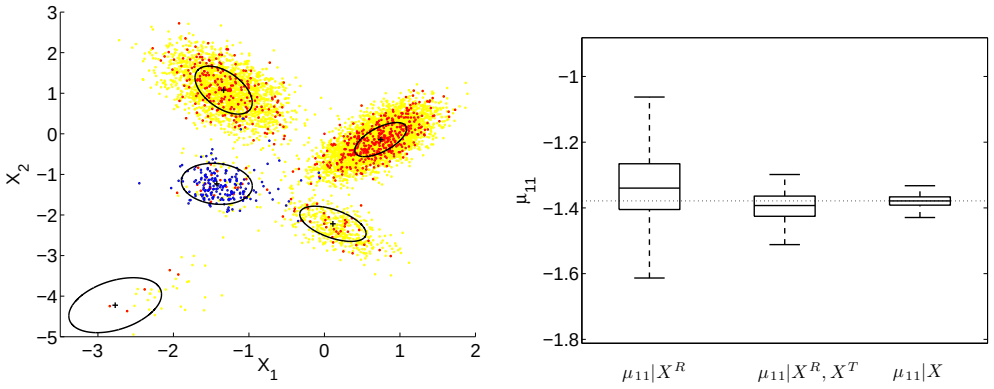

Figure 1: On the left, a scatter plot of the 2 dimensional data. The full data are shown in yellow, the random subsample in red, and the targeted subsample after the SMC algorithm in blue. The contours show the component structure of a single draw from the total posterior, with crosses at the mean and ellipses at one standard deviation in each direction. On the right, three boxplots for the first coordinate of the posterior mean of interest: conditional on the random subsample (700 observations), the random plus targeted subsample (900 observations), and the full data (5000 observations). The horizontal dotted line shows the true value of the component mean.

5.2 Example: Flow cytometry

The motivating example for this study is a problem arising in flow cytometry, where cellular subtypes may be associated with one (or more) components of a Gaussian mixture model (see Chan et al. 2008). Flow cytometers detect fluorescent reporter markers that typically correspond to specific cell surface or intracellular proteins on individual cells, and can assay millions of such cells in a fluid stream in minutes. Datasets are typically very large, and as a result inference on the full data is computationally prohibitive. Interest lies in identifying and characterizing rare cell subtypes using a mixture model fitted on those markers. The ability to identify such rare cell subsets plays important roles in many medical contexts - for example, the detection of antigen-specific cells with MHC class I or class II markers, identification of polyfunctional T lymphocytes that correlate with vaccine efficacy or host resistance to pathogens, or in resolving variants of already low frequency cell types, e.g. subtypes of conventional dendritic cells.

We use a dataset of 50,000 data points from human peripheral blood cells, with 6 marker measurements each: Forward Scatter (measure of cell size), Side Scatter (measure of cell granularity), CD4 (marker for helper T cells), IFNg+IL-2 (effector

cytokines), CD8 (a marker for cytotoxic T cells), CD3 (marker for all T cells)[1]. The objective is to provide higher resolution on the structure and patterns of covariation of cells of a specific cell subtype, specifically cells high in CD3 and/or CD8 secreting IL-2/IFN-g when challenged with a specific viral antigen. In other words, we are particularly interested in observations with large values in the 4th dimension together with the 5th and/or 6th. The data show a clear component structure for some of the markers (see Figure 2), whereas in others the rare cell subtypes of interest are not separated. To illustrate our methods and for ease of exposition, we adapt our algorithm by targeting inferences towards the component with CD4 and IFNg+IL-2 (3rd and 4th dimension) centred closest to a specific point, and set each τ_j to be very large, namely $\tau_j = 1000$ for all but the 4th dimension, focusing on the secretion of IFNg+IL-2. Here we set $K = 200$, as we expect the maximum number of components to be far fewer than 200. Owing to the structure of the model, using an upper bound much larger than the number of components necessary does not affect the accuracy of the posterior estimates.

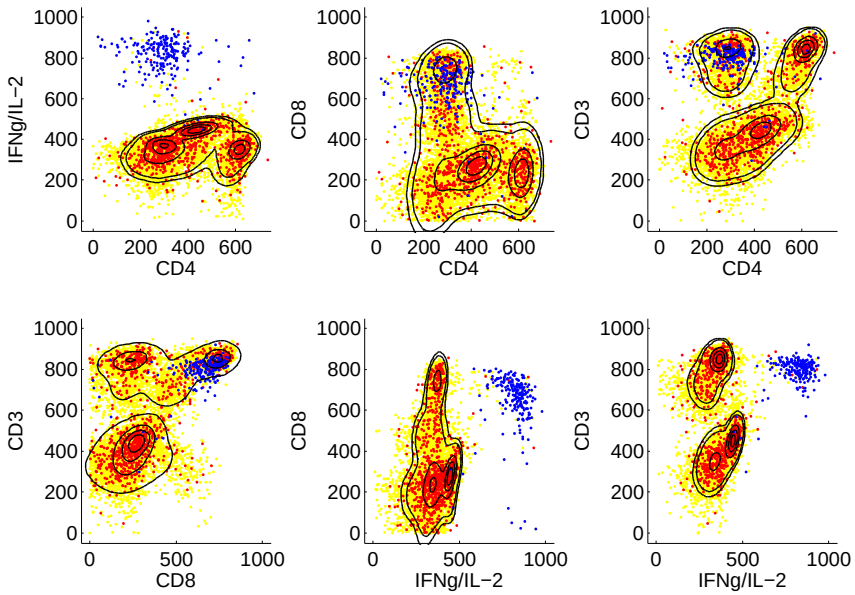

Figure 2: Scatter plots of the data for the last 4 markers: CD4, IFNg, CD8 and CD3. The complete data set is shown in yellow. We aim to use the random subsample (shown in red) in order to obtain samples from the initial posterior $p(\mu, \Sigma, \pi, \alpha \mid X^R)$ and draw the targeted subsample (shown in blue) using estimates of the distribution of the data (superimposed as a contour plot).

[1]Data from an NIAID/BD IntraCellular Staining Quality Assurance Panel (ICS QAP) kindly provided by the Duke Center for AIDS Research (CFAR) Immune Monitoring Core

An initial random subsample of size 5,000 is drawn, providing us with initial estimates m, S for the mean and covariance of the component closest to the high CD4+ region. Due to the strong covariation between the markers, several components are needed (see Figure 3) in order to capture the inhomogeneity of the data. Using initial weights $w(x) = N(x \mid m, S_\tau)$, we apply our sequential Monte Carlo algorithm to obtain a complete targeted subsample in terms of the stopping rule as well as posterior samples for all our parameters. We draw 30 observations at a time so that $B = 30$, with $c_{threshold} = 0.2$ and $N_{threshold} = 30$, resulting in a total of 390 targeted observations.

Figure 3 shows estimated posterior distributions for the total number of components based on the initial random subsample, and subsequently after the SMC sampler given both the random and targeted subsamples. We observe the efficacy of the targeted approach, reflected in part through improved identification of the structure of the model density in the targeted region (Figure 4). More specifically, observing samples from the mixture model (see Figure 3) in the CD4 and IFNg/IL-2 markers before and after the targeted subsample, we see that the targeted approach has led to finer resolution on the mixture structure in the region of the targeted rare cell subtypes, providing higher resolution about the structure and covariation of their markers. Importantly, this revealed components in the low probability subregion which emerge due to the covariation with the remaining markers. This is confirmed by the analysis on the full dataset, shown in Appendix 6, where several more components are inferred. The improvement of the estimates obtained using our selection sampling approach on a total subsample (random and targeted) of size 5390 by far exceeds the expected improvement using a random subsample of the same size. The findings agree with the biologists' expectation that cell subtypes often have a non-Gaussian structure, (see Chan et al. 2008; Pyne et al. 2009), and provide an efficient method of detecting and drawing inferences about rare populations in the presence of very large datasets.

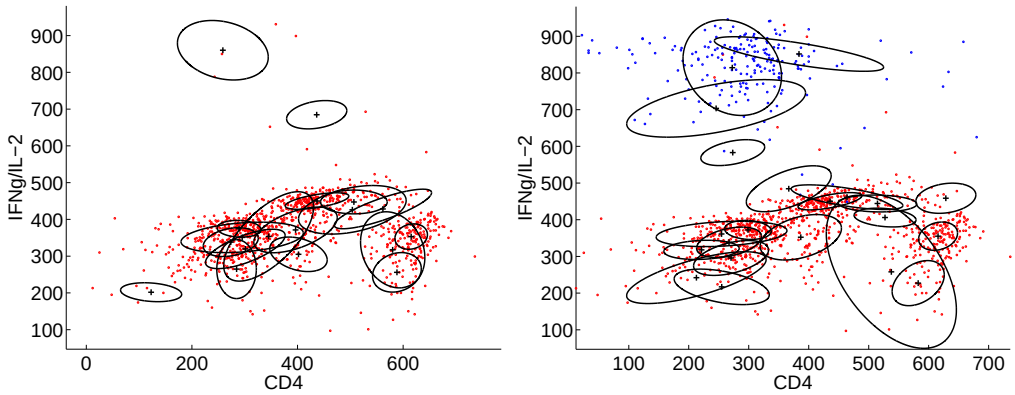

Figure 3: Flow cytometry data analysis via using the Sequential Monte Carlo targeted re-sampling algorithm: sample realization of the mixture model (a) based on the random subsample and (b) based on both the random subsample and the targeted subsample. Crosses are shown at the mean of each component, with 50% contours drawn.

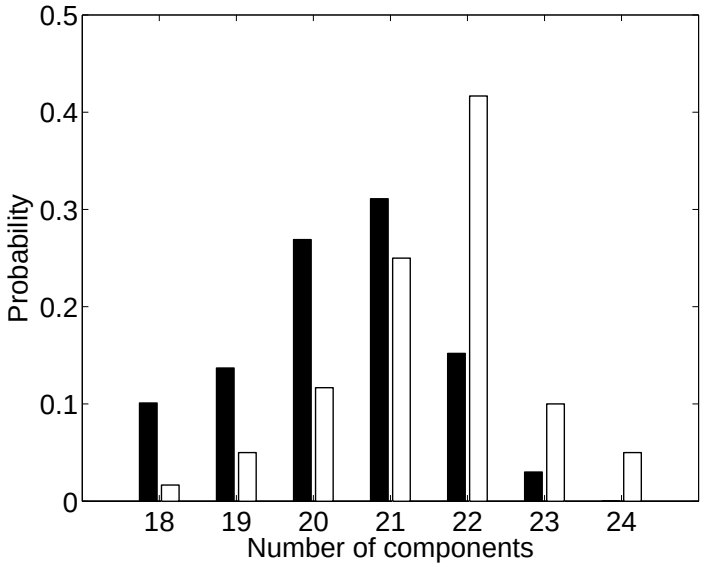

Figure 4: Posterior distributions for the number of components in the Gaussian mixture model, given only the random subsample $p(k \mid X^R)$ shown in black and given both the random and targeted subsample $p(k \mid X^R, X^T)$ shown in white.

6 Discussion

One of the key aspects of this work consists in defining the low probability region of interest and specifying the weight function used in the targeted sample. Naturally, the low probability region in sample space is strongly driven by the scientific question in hand. Based on that, and taking into account algorithmic tractability and efficiency, different weight functions may be used. In this work we presented methods relating to inferences about a specific component, defined in terms of an identifying criterion. In the flow cytometry example used in this paper, this was chosen as the component with mean closest to a specific point. Although the weight function used had a Gaussian shape, the analysis revealed a non-Gaussian structure in the region of interest; using mixtures of components as a weight function would be a straightforward extension of our methods. In fact, using a hierarchical model of mixtures of Gaussian mixtures may provide a better fit to the non-Gaussian inhomogeneous structure of the flow cytometry data; our targeted subsampling approach can be implemented using such models at little additional computational cost.

In the case of the flow cytometry data, an alternative to the identifying criterion of the component of interest would be to use the component containing a specific cell of known rare cell subtype. A natural extension to the weight functions used in this work stems from the fact that, in the original flow cytometry data, the identifying criterion for the component of interest is not defined on a fixed number of dimensions. Instead,

it is defined as the set of markers which are significant in identifying the component in the region of low probability in sample space, which itself is unknown. In other words, the Gaussian mixture may be defined only on a subset of the p markers (unknown), such that we draw inferences about the parameters of the mixture $p(\theta^q \mid X) \quad x_i \in \mathbb{R}^q$ corresponding to variable dimensions $1 : q, \quad q \leq p$. The targeted learning about θ^q can be incorporated in the analysis such that, within the sequential design, the weight function $w(x) \propto N(x \mid m, S)$ is updated at each round of re-sampling both in terms of the mean m and covariance S of the Gaussian distribution, but also in terms of the markers over which the weight function is defined. In the case of flow cytometry data, this can be viewed as soft gating of cells into cell subtypes, based on both the values of the individual markers, but also the set of significant markers.

One of the main challenges in drawing inferences about targeted subsamples is constructing efficient proposals for the parameters of interest, as the convergence of the algorithms is influenced by several factors. The size of the targeted subsample in relation to the random subsample plays a significant role. This becomes especially important when the assumption of a large number of observations within the region of interest is breached, as this would lead to a likelihood used for the targeted subsample which deviates severely from the true likelihood because of sampling without replacement. The multiplicative elements of τ also play significant roles in constructing a weight function which is wide enough to not violate the infinite weights assumption, at the same time targeting the region of interest.

Furthermore, the size of the initial subsample affects the posterior variance of the estimators. In fact, the random subsample may contain no observations in the low probability region. The weight function can be treated as having a prior mean and variance, used if no more information is available in the random subsample. As more observations are drawn sequentially from the specific region, this weight function is updated to include further information. Although the posterior estimates will be comparable, a very small initial subsample will strongly affect the efficiency of the proposal kernels, and possibly require overall longer running times to obtain similar results.

References

Balakrishnan, S. and Madigan, D. (2006). "A one-pass sequential Monte Carlo method for Bayesian analysis of massive datasets." *Bayesian Analysis*, 1(2): 345–362. 430

Bayarri, M. J. and Berger, J. (1998). "Robust Bayesian analysis of selection models." *Annals of Statistics*, 26(2): 645–659. 430

Carvalho, C., Johannes, M., Lopes, H., and Polson, N. (2010). "Particle Learning and Smoothing." *Statistical Science*. To appear. 437

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M., and Kepler, T. (2008). "Statistical mixture modeling for cell subtype identification in flow cytometry." *Cytometry A*, 73: 693–701. 429, 430, 440, 442

Chopin, N. (2002). "A sequential particle filter method for static models." *Biometrika*, 89(3): 539–552. 437

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York. 436, 437

Gordon, N. J., Salmond, D. J., and Smith, A. F. M. (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." In *IEE Proceedings*, volume 140, 107–113. 438

Heckman, J. J. (1979). "Sample selection bias as a specification error." *Econometrica: Journal of the econometric society*, 47(1): 153–161. 430

Ishwaran, H. and James, L. (2002). "Approximate Dirichlet process computing in finite normal mixtures: Smoothing and prior information." *Journal of Computational and Graphical Statistics*, 11: 508–532. 430, 431, 433

Liu, J. and West, M. (2000). "Combined parameter and state estimation in simulation-based filtering." In Doucet, A., De Freitas, J. F. G., and Gordon, N. J. (eds.), *Sequential Monte Carlo Methods in Practice*, 197–223. Springer-Verlag, New York. 430

MacEachern, S. N. (1998). "Estimating mixture of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7(2): 223–238. 430

Müller, P., Erkanli, A., and West, M. (1996). "Bayesian curve fitting using multivariate normal mixtures." *Biometrika*, 83(1): 67. 430

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jagera, P. L., and Mesirova, J. P. (2009). "Automated high-dimensional flow cytometric data analysis." *Proceedings of the National Academy of Sciences*, 106(21): 8519. 442

Ridgeway, G. and Madigan, D. (2003). "A sequential Monte Carlo method for Bayesian analysis of massive datasets." *Data Mining and Knowledge Discovery*, 7(3): 301–319. 429

Seder, R., Darrah, P., and Roederer, M. (2008). "T-cell quality in memory and protection: implications for vaccine design." *Nature Reviews Immunology*, 8(4): 247–258. 429

Suchard, M., Wang, Q., Chan, C., Frelinger, J., Cron, A., and West, M. (2010). "Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures." *Journal of Computational and Graphical Statistics*, 19: 419–438. 430

West, M. (1994). "Discovery sampling and selection models." In Gupta, S. S. and Berger, J. O. (eds.), *Statistical Decision Theory and Related Topics*, 221–235. Springer-Verlag, New York. 430

— (1996). "Inference in successive sampling discovery models." *Journal of Econometrics*, 75(1): 217–238. 430

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 2nd edition. 438

Appendix

A: Flow cytometry data

In the case of the flow cytometry dataset under study, MCMC inference on the full dataset is feasible. The results are shown below in Figure 5, comparing the 3 relevant GMM structures in the case of (a) a random subsample (b) a random and a targeted subsample and (c) the full data. Here it's not possible to draw a direct comparison between the posterior distributions of the component k^*, because the component structure in the random, targeted and full dataset case changes significantly (with a much larger number of components in the full dataset case, as is expected by the Dirichlet process prior). The comparison shows that, indeed, there are more components in the region of

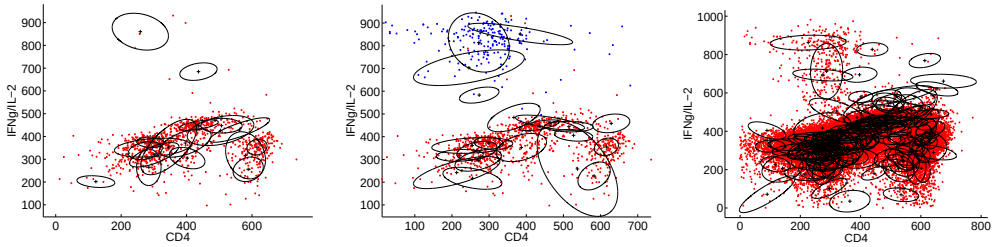

Figure 5: Samples of the Gaussian mixture model structure in the case of the (a) random subsample (b) random and targeted subsample and (c) full data. Red points represent randomly selected observations while blue represent targeted observations. Components are shown with a '+' at the mean and ellipses at one standard deviation in each direction.

interest using the full data, as is also suggested by our targeted approach.

B: Weight functions

In both the Markov chain Monte Carlo and Sequential Monte Carlo approaches described above, the targeted sample was weighted according to $w(x) = N(x \mid m, S_\tau)$, where m and S are estimates of the mean and covariance of the low-probability component k^*, and $S_\tau = TST$ where $T = \text{diag}(\tau_1^{1/2}, \ldots, \tau_p^{1/2})$ is a $p \times p$ diagonal matrix based on a set of positive *variance multipliers* τ_j, $(j = 1, \ldots, p)$. The τ_j are tuning parameters. Larger values will allow for wider dispersal of the targeted sub-sample, accounting for uncertainty of the initial estimate of ϕ_{k^*}. As elements τ_j decrease, the weights w_i in the targeted sample become more concentrated and so favour fewer potential data points. One result of this is that it the assumption of an infinite number of points with non-negligible weight becomes invalid. If our initial estimate of $(\mu_{k^*}, \Sigma_{k^*})$ is poor, small elements τ_j will restrict the targeted sample to a region away from the full low probability region of interest. Within the context of the Metropolis-Hastings updates,

as elements τ_j increase, the acceptance rate for μ, Σ increases, since the targeted sample looks more like the random sample. In that case, the posterior distribution of ϕ_{k^*} is not pulled too far from the proposed values. At the same time, as elements τ_j increase, acceptance rates for π decrease since the information about π given by the targeted sample becomes significant, and the proposed values (which are based only on the random subsample) become less acceptable.

Consider the one-dimensional case where $S_\tau = \tau S$ so that $w(x_i) \propto N(x_i \,|\, m, \tau S)$. Assume that μ, Σ, π are known and that there is an infinite number of data points. The weight function becomes $w(x_i) \approx N(x_i \,|\, \mu_{k^*}, \tau\Sigma_{k^*})$, and the coefficient τ (here scalar) may be chosen such that the probability of drawing data points from the low-probability component is maximized. An example is shown in Figure 6.

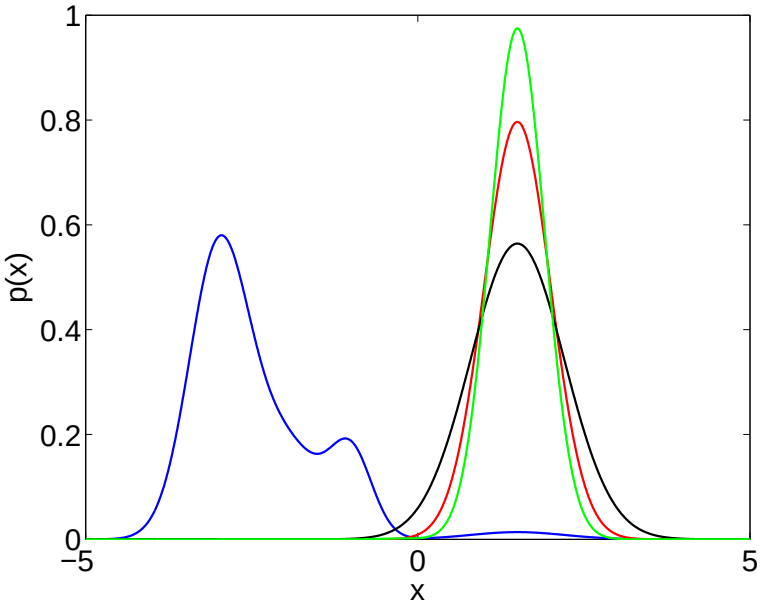

Figure 6: Example in one dimension, here the blue curve represents the mixture $f(x \,|\, \pi, \mu, \Sigma)$ and the red line the density of the low-probability component $N(x \,|\, \mu_{k^*}, \Sigma_{k^*})$. The black curve then represents the weight function $N(x \,|\, \mu_{k^*}, \tau\Sigma_{k^*})$, and the green curve the mixture distribution of the targeted sample, $\tilde{f}(x \,|\, \tilde{\pi}, \tilde{\mu}, \tilde{\Sigma})$. Ideally we want the common area of the green and red curve to be maximized.

Considering the overlap between the distribution of the targeted subsample and the low-probability component, we plot the common area, denoted by $A(\tau)$, for varying τ, and obtain the graph shown in Figure 7. As is seen from Figure 7, in terms of maximizing the overlap between the low probability component and the targeted subsample, the optimum value of τ varies. Specifically, the closer the remaining components are to the component of interest (and similarly the higher their variance) yields a lower value for the optimum τ, and the same happens when the weight of the component of interest

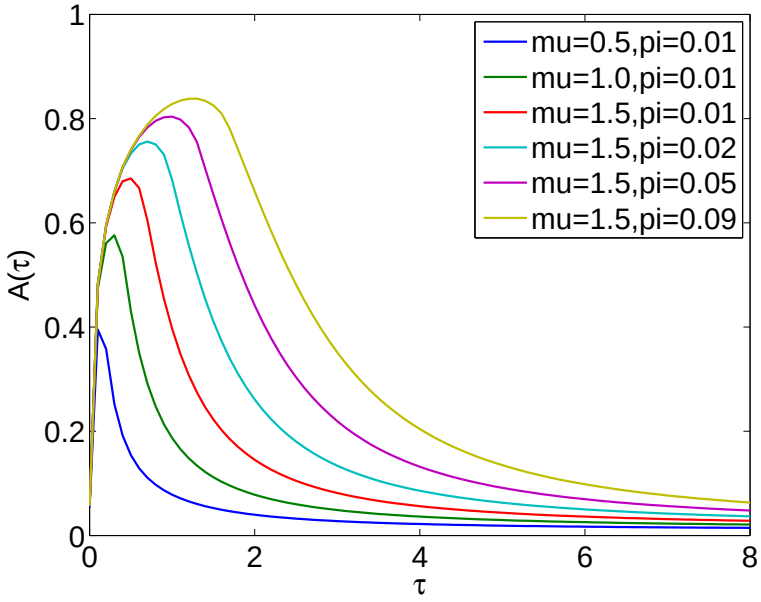

Figure 7: Example of $A(\tau)$ for several values of (μ_{k^*}, π_{k^*}), using a numerical approximation of the integral in order to calculate the common area.

decreases.

Combining the above results with the fact that a large τ will improve the acceptance rate for μ, Σ but reduce the acceptance rate for π, and taking into account uncertainty on the $S = \hat{\Sigma}_{k^*}$, it is apparent that the optimum coefficient τ is not uniquely 1, and plays a significant role which affects many levels of the analysis.

C: Software

Matlab code implementing the analysis described here, with examples, is available at http://ftp.stat.duke.edu/WorkingPapers/09-26.html.

Acknowledgments

Research was partially supported by grants to Duke University from the National Science Foundation (DMS-0342172) and the National Institutes of Health (grant P50-GM081883, P30-AI064518-0 and contract HHSN268200500019C). Aspects of the research were also partially supported by the NSF grant DMS-0635449 to the Statistical and Applied Mathematical Sciences Institute. CC and MW were also partially supported on this research by NIH grant RC1AI086032. Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NSF or NIH.