Targeted Sequential Resampling from Large Data Sets in Mixture Modeling

Ioanna Manolopoulou, Cliburn Chan and Mike West

SAMSI and Duke University

August 3, 2009

Objective

Very **big datasets** (large n, possibly large p), interested in **low probability subpopulations**. Example from flow cytometry:

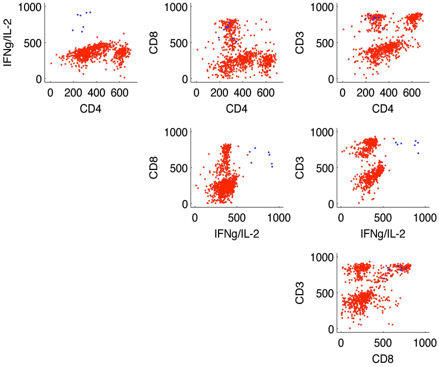


Figure: Pairwise plots of flow cytometry data: interested in large IFNg and CD3 and/or CD8.

Intuition:

Obtain estimates about parameters of those low probability regions in sample space, identify and focus on similar regions.

Flow cytometry

Measuring cells in a fluid

- Fluid passing through a tube.
- Shine laser beams at wavelengths $l_1, \ldots, l_p$.
- Record wavelength of reflected beams.

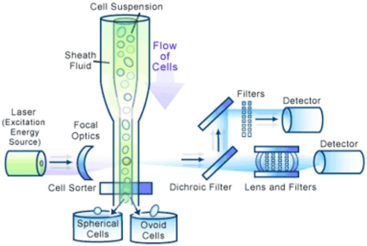


Figure: Setup of typical flow cytometry experiment, from the Science Creative Quarterly, `www.scq.ubc.ca`, by Jane Wang

Flow cytometry data

Objective: identify and characterize cell subtypes with high IFNg/IL-2, CD3, CD4, CD8.

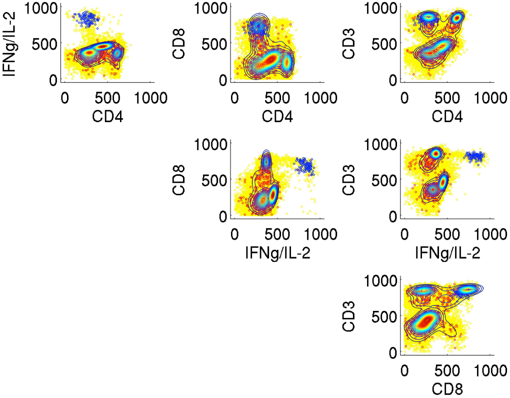


Figure: Pairwise plot of last 4 markers. Yellow: full data, red: random subsample, blue: targeted subsample. Contour plot superimposed.

Example in 1d

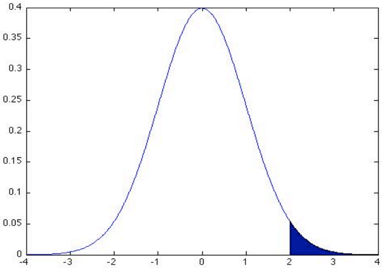


Figure: Example of outlier detection in 1d: values >2

- In higher dimension, distribution of data becomes intractable
- Threshold not necessarily known
- Rare region of interest not necessary 'separated'.

Mixture Models

Distribution $f(x|\pi, f_1, \ldots, f_K) = \sum \pi_i f_i$

- Set of distributions $f_1, \ldots, f_k$.
- Set of mixing weights $\pi_1, \ldots, \pi_k$.
- Mixture distribution $f = \sum \pi_i f_i$
- Interested in a rare component.

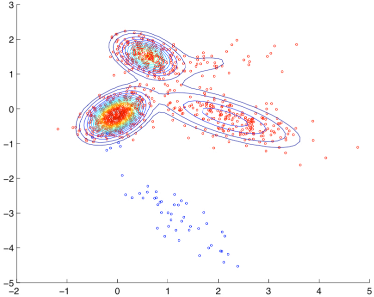


Figure: Gaussian Mixture Model example.

Sampling the targeted regions

Gaussian Mixture Model $f(x) = \sum \pi_j N(\mu_j, \Sigma_j)$, Dirichlet Process πs.

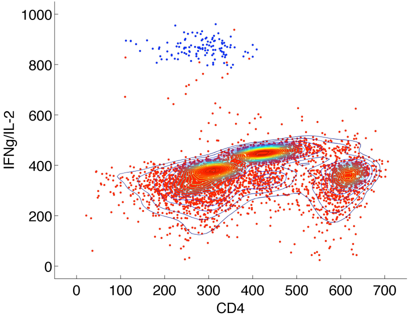


Figure: Red: random subsample, blue: targeted subsample

MCMC approach

Inferences about the distribution of a component in a pre-specified region.

- Run initial MCMC on random subsample, obtain $\hat{\phi}_K = (\hat{\mu}_K, \hat{\Sigma}_K)$
- For each data point x_i calculate weights

$$w(x_i) \propto N(x_i | \hat{\mu}_K, \hat{\Sigma}_K)$$

- Draw a *targeted* subsample without replacement, with distribution

$$f^S(x_i|\mu, \Sigma) \propto w_i f(x_i|\theta) \propto \sum_{k=1}^{K} \tilde{\pi}_k N(x_i|\tilde{\mu}_k, \tilde{\Sigma}_k).$$

- Run second MCMC on both the random and targeted subsample.

Focusing on parameters of the rare subpopulation

How can we use the initial samples?

- Use initial posterior estimates to approximate

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

  using that $p(z^R|X^R, X^T) \approx p(z^R|X^R)$.
- Only update π, μ, Σ, z^T.
- Conditioning on Z^R **de-couples** the z-dependence of the random and targeted subsamples.

SMC algorithm

For each particle, draw a sample of $(Z^R, \pi, \mu, \Sigma)|X^R$ from the posterior of the MCMC. Then iteratively draw targeted data points through the following steps.
For $i = 1 : M$

1. Draw a 'targeted' observation X_i^S without replacement according to weights $w_j \propto f(x_j|\hat{\mu}_K, \hat{\Sigma}_K)$.
2. Using Metropolis Hastings, update

$$\mu_K, \Sigma_K | X^R, X^S_{1:i}, Z^R, \pi, \mu_{1:K-1}, \Sigma_{1:K-1}$$

In other words, for each particle $j = 1 : J$, propagate

$$(\mu_K^{(j)}, \Sigma_K^{(j)}) \xrightarrow{propagate} (\mu_K^{(j)}, \Sigma_K^{(j)})'.$$

After propagating ALL particles, draw

$$\mu_K, \Sigma_K \xrightarrow{drawdatapoint} X_i^S.$$

Introduce a decision rule such that we stop augmenting the targeted subsample.

3. If no other observations fall within the 50% contour of the weight function, stop.

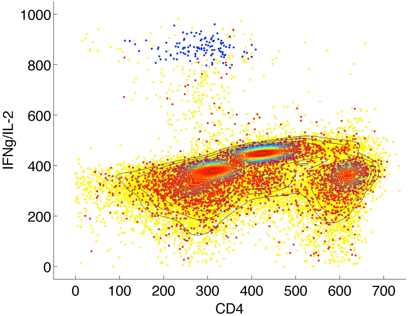


Figure: Full sample in yellow, random subsample in red, targeted subsample in blue

Comparison of posterior distributions

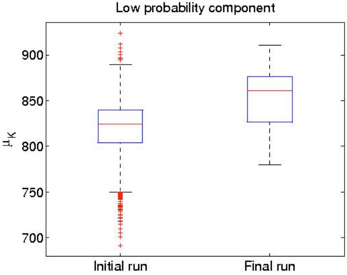
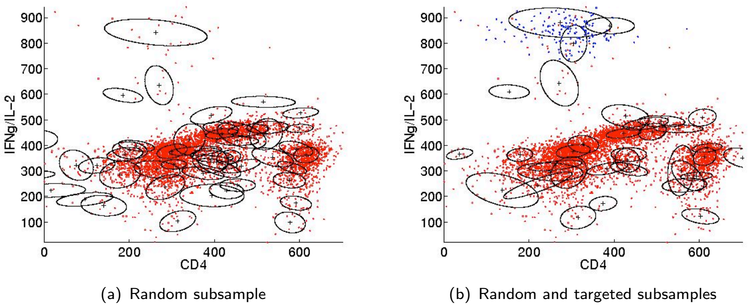


Figure: Box plots of one of the markers for the component of interest only on the random subsample (left) and on both the random and targeted subsamples (right).

Comparison of mixture distributions



(a) Random subsample

(b) Random and targeted subsamples

Figure: Sample of the mixture model using the random sample only (left) and both the random and targeted sample (right).

Weight functions

Weight functions chosen based on scientific question and computational efficiency/tractability.

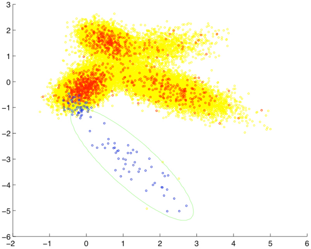


Figure: Example of a weight function which is too wide

Weight functions

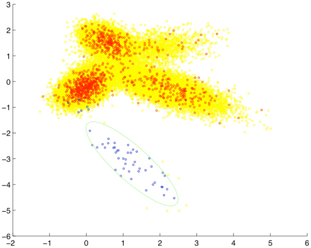


Figure: Example of a weight function which collects (almost) all data points from the low probability component

Weight functions

How can we 'collect' as many points as possible?

Maximize overlap between the targeted subsample and the rare component. Let

$$w(x) \propto N(x|\mu_K, \tau\Sigma_K)$$

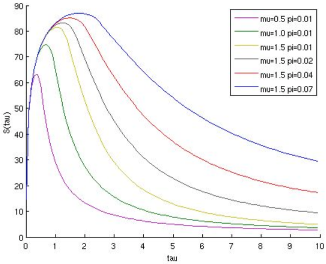


Figure: Common area between distribution of x^T and component K: optimizing τ for different values of μ, π.

Weight functions

How does the weight function affect the convergence?

- As τ increases, more information about π, harder convergence in π.
- As τ decreases, more information about μ_K, Σ_K, harder convergence in μ, Σ.

Current & future work

- Dimension selection: which markers matter?
- Draw inferences about the location as well as the patterns of covariation of the low probability subregion of the flow cytometry data.
- Compare case and control datasets.
- Construct efficient weight functions to address scientific question.
- Implement for non-Gaussian mixtures $f(x) = \sum_j \pi_j f_j(x)$.
- Try different priors for mixing weights π.

Thanks!