Sequential Dirichlet Process Mixtures of Multivariate Skew t-distributions for Model-based Clustering of Flow Cytometry Data

Boris P. Hejblum, Chariff Alkhassim, Raphael Gottardo, François Caron, Rodolphe Thiébaut

Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

SEQUENTIAL DIRICHLET PROCESS MIXTURES OF MULTIVARIATE SKEW *T*-DISTRIBUTIONS FOR MODEL-BASED CLUSTERING OF FLOW CYTOMETRY DATA

By Boris P. Hejblum[*,†,‡] , Chariff Alkhassim[†,‡] , Raphael Gottardo[§] , François Caron[¶] and Rodolphe Thiébaut[†,‡]

Univ. Bordeaux, ISPED, Bordeaux Population Health Research Center Inserm U1219, Inria SISTM, 33000 Bordeaux, France[†], Vaccine Research Institute (VRI), 94010 Créteil, France[‡], Fred Hutchinson Cancer Research Center, Seattle, Washington, U.S.A.[§], Department of Statistics, University of Oxford, Oxford, U.K.[¶]

Flow cytometry is a high-throughput technology used to quantify multiple surface and intracellular markers at the level of a single cell. This enables us to identify cell sub-types, and to determine their relative proportions. Improvements of this technology allow us to describe millions of individual cells from a blood sample using multiple markers. This results in high-dimensional datasets, whose manual analysis is highly time-consuming and poorly reproducible. While several methods have been developed to perform automatic recognition of cell populations, most of them treat and analyze each sample independently. However, in practice, individual samples are rarely independent, especially in longitudinal studies. Here we analyze new longitudinal flow-cytometry data from the DALIA-1 trial which evaluates a therapeutic vaccine against HIV, by proposing a new Bayesian nonparametric approach with Dirichlet process mixture (DPM) of multivariate skew t-distributions to perform model based clustering of flow-cytometry data. DPM models directly estimate the number of cell populations from the data, avoiding model selection issues, and skew t-distributions provides robustness to outliers and non-elliptical shape of cell populations. To accommodate repeated measurements, we propose a sequential strategy relying on a parametric approximation of the posterior. We illustrate the good performance of our method on simulated data and on an experimental benchmark dataset. This sequential strategy outperforms all other methods evaluated on the benchmark dataset, and leads to improved performance on the DALIA-1 data.

[*]Corresponding author: boris.hejblum@u-bordeaux.fr

MSC 2010 subject classifications: Primary 62H30, 62P10; secondary 62L12

Keywords and phrases: Automatic gating, Bayesian Nonparametrics, Dirichlet process, Flow cytometry, HIV, Mixture model, Skew t-distribution

1. Introduction. Flow cytometry is a high-throughput technology used to quantify multiple surface and intracellular markers at the level of single cell. More specifically, cells are stained with multiple fluorescently-conjugated monoclonal antibodies directed to cell surface receptors (such as CD4) or intracellular markers (such as cytokines) to determine the type of cell, their differentiation, and their functionality. With the improvement of this technology leading currently to the measurement of up to 18 markers at the same time (using 18 colors for flow cytometry), multi-parametric description of millions of individual cells can be generated.

Analysis of such data is generally performed manually. This results in analyses that are: i) poorly reproducible (Aghaeepour et al., 2013), ii) expensive (highly time-consuming) and consequently iii) only focused on specific cell populations (i.e. specific combination of markers), ignoring other cell populations. There has been an effort in the recent years to offer automated solutions to overcome these limitations (Lo, Brinkman and Gottardo, 2008; Aghaeepour et al., 2013; Gondois-Rey et al., 2016). Quite a lot of different methodological approaches have been proposed to perform automatic recognition of cell populations from flow cytometry data. Clustering methods related to the k-means were proposed, including L2kmeans (Aghaeepour et al., 2013), flowMeans (Aghaeepour et al., 2011). Model based clustering methods relying on finite mixture models such as flowCust/merge (Lo, Brinkman and Gottardo, 2008; Finak et al., 2009), FLAME (Pyne et al., 2009), SWIFT (Naim et al., 2014) were also proposed, as well as dimension reduction methods such as MM and MMPCA (Sugár and Sealfon, 2010), SamSPECTRAL (Zare et al., 2010), FLOCK (Qian et al., 2010). All those approaches require the number of cell populations to be fixed in advance, determining its optimal value according to various criteria. Finally, several authors (Chan et al., 2008; Lin et al., 2013; Cron et al., 2013; Dundar et al., 2014) proposed nonparametric Bayesian mixture models of Gaussian distributions, that directly estimate this number of cell populations. All these methods, except those of Lin et al. (2013), of Cron et al. (2013) and of Dundar et al. (2014), were evaluated by Aghaeepour et al. (2013).

However, there is still room for improvement, especially in the estimation of the suitable number of cell populations, as well as in the identification of rare cell populations. In addition, most of those previous approaches have been proposed for single sample analysis, except for Cron et al. (2013) who proposed to use hierarchical Dirichlet process mixture (DPM) of Gaussian distribution models to analyze multiple samples simultaneously. Yet in the case of repeated measurements of flow cytometry data, it can be useful to perform a sequential analysis as the samples are acquired (samples are often

collected across several time points in a population of patients). In such a case, one would want to use previously acquired samples as prior information in the analysis of a new sample. In this paper, the proposed approach includes a strategy of sequential approximations of the posterior distribution for multiple data samples, presented in Section 3.2. Our approach offers three advantages: i) it quantifies the uncertainty of the posterior clustering, ii) it can make use of prior knowledge to inform on the structure of the data, potentially building up on previous analyses, and iii) it allows the analysis of multiple samples without requiring to process all the data at once, alleviating both the computational burden and the necessity for all data to be readily available before any analysis can be performed.

The automatic recognition of cell populations from flow cytometry data is a difficult task which can be seen as an unsupervised clustering problem (Lo, Brinkman and Gottardo, 2008). It is characterized by two big challenges. First, the total number of cell populations to identify is unknown. Second, the empirical distributions of the populations are heavily skewed, even when optimal transformation of the data is applied (Lo, Brinkman and Gottardo, 2008; Pyne et al., 2009; Lo and Gottardo, 2012), and the data generally present many outliers. To address all these points together, our approach considers a Bayesian nonparametric model-based approach, where the flow cytometry data are assumed to be drawn from a DPM of multivariate skew-t distributions. First, this approach enables the number of cell populations to be inferred from the data and avoids the challenging problem of model selection. Second, it has been demonstrated that the Gaussian assumption for the parametric shape of a cell population fits poorly flow cytometry data (Mosmann et al., 2014). Indeed, even after state-of-the-art transformation of raw cytometry data, such as the biexponential transformation (Finak et al., 2010), cell population distributions are typically skewed. Pyne et al. (2009) have showed the advantages of the skew t-distribution (Azzalini and Capitanio, 2003) for modeling cell populations in flow cytometry data. Numerous parameterizations have been proposed for the multivariate skew t-distribution (Lee and McLachlan, 2013; Murray, Browne and McNicholas, 2014; Azzalini et al., 2016; McLachlan and Lee, 2016), most notably the restricted and the unrestricted multivariate skew t-distributions (denoted rMST and uMST respectively) which are generalizations of the skew normal distribution (Azzalini and Valle, 1996) with a heavier tail (making it more robust to outliers). Lee and McLachlan (2016) recently proposed the canonical fundamental skew t-distribution (CFUST) as a generalization that encompasses both the rMST and the uMST. To avoid identifiability issues associated with the uMST and the CFUST (Lee and McLachlan, 2016), in

this work we will adopt the rMST formulation, and in the remainder of this article we will be referring to the rMST formulation when mentioning the skew t-distribution. Frühwirth-Schnatter and Pyne (2010) proposed a finite mixture model of rMST. We extend this model to the infinite mixture case in a Bayesian nonparametric framework. Of interest, quantifying the uncertainty around the estimated partition is straightforward in this Bayesian paradigm, from the posterior distribution of the partition. While a skewed distribution could be fitted either by a skew t or by a mixture of Gaussians, using the latter requires to estimate separately the overall number of clusters and the skewness. On the contrary, our proposed approach jointly estimates the two thus taking into account the uncertainty associated with both. Furthermore, the use of a Bayesian framework enables the use of informative priors. In the case of repeated measurements for instance, we propose to sequentially estimate the posterior partition of flow cytometry using posterior information from time point t as prior information for time point $t + 1$.

The proposed method is evaluated on simulated data and on a benchmark clinical dataset from Aghaeepour et al. (2013), and is applied to analyze an original experimental longitudinal dataset from a phase I HIV clinical trial DALIA-1 (featuring depending time-course data where the sequential approach is of particular interest). The method is implemented in the R package NPflow, available on the CRAN at `https://CRAN.R-project.org/package=NPflow`.

2. Statistical Model.

2.1. *Motivation and problem set-up.*

2.1.1. *Motivating example.* Our motivating example for developing a sequential model-based clustering approach for longitudinal flow cytometry data comes from the DALIA-1 trial. DALIA-1 is a phase I trial for a therapeutic vaccine candidate against HIV (Lévy et al., 2014). This vaccine candidate was based on ex-vivo generated interferon-α dendritic cells loaded with HIV-1 lipo-peptides, and activated with lipopolysaccharide. The primary objectives of this trial were to evaluate the safety of the vaccine strategy and to evaluate the immune response. As part of this trial, 12 HIV positive patients had their cellular populations quantified repeatedly by flow-cytometry, generating an important amount of data, for which comprehensive manual gating would take several months. Hence we aimed at developing an automatic gating approach suitable for longitudinal measurements.

2.1.2. *Problem set-up and notations.* We first consider a single sample per subject, i.e. one data matrix where each row represents a cell and each column contains the fluorescence intensities for one marker measured by the flow cytometer. The case of the sequential estimation of multiple datasets will be addressed in Section 3.2. We let $\boldsymbol{y}_c \in \mathbb{R}^d$ denote the data, $c = 1, \ldots, C$ corresponding to the vector of fluorescence intensities measured for the cell c. Typically, the observations $\boldsymbol{y}_c$ have been transformed (to help visualization and gating) from the raw measurements of fluorescence through a biexponential or Box-Cox transformation (Finak et al., 2010). We assume that these observations are independent and identically distributed (i.i.d.) from some unknown distribution F:

$$\boldsymbol{y}_c | G \overset{i.i.d.}{\sim} F \text{ for } c = 1 \ldots, C \tag{2.1}$$

where F is a mixture of distributions:

$$F(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\boldsymbol{y}) G(d\boldsymbol{\theta}) \tag{2.2}$$

with $f_\theta(\boldsymbol{y})$ is a known probability density function, parameterized by $\boldsymbol{\theta} \in \Theta$ a set of parameters, and defining the shape of a cluster. G is the unknown mixing distribution, which carries the weights and locations of the mixture components. In a parametric approach, $G = \sum_{k=1}^{K} \pi_k \delta_{\boldsymbol{\theta}_k}$ where π_k is the weight of the k^{th} mixture component. Maximum likelihood or Bayesian estimates of F can be derived for such models (Biernacki, Celeux and Govaert, 2000). In a nonparametric perspective (where the number of clusters is unknown) G is written as an infinite sum of atoms: $G = \sum_{k=1}^{+\infty} \pi_k \delta_{\boldsymbol{\theta}_k}$. The Dirichlet process is a conjugate prior for the infinite atomic discrete distribution, which makes it very useful for unsupervised clustering approaches.

2.2. *Dirichlet process mixture of skew t-distributions.*

2.2.1. *Dirichlet process mixture.* We assume that the random mixing distribution G is drawn from a Dirichlet process (Ferguson, 1973):

$$G \sim \mathrm{DP}(\alpha, G_0) \tag{2.3}$$

where $\mathrm{DP}(\alpha, G_0)$ denotes the Dirichlet process of concentration parameter $\alpha > 0$ and base probability distribution G_0. A draw $G \sim \mathrm{DP}(\alpha, G_0)$ is almost surely discrete (Sethuraman, 1994) and gives a nonparametric mixing distribution $G = \sum_{k=1}^{+\infty} \pi_k \delta_{\boldsymbol{\theta}_k}$ with $\theta_k \overset{i.i.d.}{\sim} G_0$, and $\boldsymbol{\pi} = (\pi_k)_{k=1,2,\ldots}$ drawn from a so-called "stick-breaking" distribution, written as the Griffiths-Engen-McCloskey (GEM) distribution (Pitman, 2006). The model defined

by Equations (2.1), (2.2) and (2.3) yields the following hierarchical model known as a Dirichlet process mixture model (Lo, 1984; Escobar and West, 1995; Teh, 2010) with a Gamma hyperprior on α:

$$\alpha | a, b \sim \text{Gamma}(a, b) \tag{2.4a}$$
$$\boldsymbol{\pi} \,\big|\, \alpha \sim \text{GEM}(\alpha) \tag{2.4b}$$
$$\boldsymbol{\theta}_k \,\big|\, G_0 \sim G_0 \qquad \text{for } k = 1, 2, \dots \tag{2.4c}$$
$$\ell_c \,\big|\, \boldsymbol{\pi} \sim \text{Mult}(\boldsymbol{\pi}) \qquad \text{for } c = 1, 2, \dots, C \tag{2.4d}$$
$$y_c \,\big|\, \ell_c, (\boldsymbol{\theta}_k) \sim f_{\boldsymbol{\theta}_{\ell_c}} \qquad \text{for } c = 1, 2, \dots, C \tag{2.4e}$$

where ℓ_c is the latent cluster-allocation for cell c. G_0 tunes the prior information about the cluster locations while α tunes the prior distribution on the overall number of clusters K that will be uncovered within C cells. In particular we have $\mathbb{E}[K|C] = \sum_{c=0}^{C-1} \frac{\alpha}{\alpha+c}$.

2.2.2. *The multivariate skew t-distribution.* Frühwirth-Schnatter and Pyne (2010) rely on Azzalini and Valle (1996)'s parametrization of the multivariate skew normal ($\mathcal{SN}$) to propose a truncated normal random-effects model representation of this distribution: $\boldsymbol{Y} = \boldsymbol{\xi} + \boldsymbol{\psi} Z + \boldsymbol{\varepsilon}$ with $Z \sim \mathcal{N}_{[0,+\infty[}(0, 1)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. If $\boldsymbol{X} \sim \mathcal{SN}(\mathbf{0}, \boldsymbol{\Omega}, \eta)$ and $W \sim \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2})$, (Azzalini and Capitanio, 2003) show that $\boldsymbol{Y} = \boldsymbol{\xi} + \frac{1}{\sqrt{W}}\boldsymbol{X}$ then follows a multivariate skew t-distribution: $\boldsymbol{Y} \sim \mathcal{ST}(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu)$. Following Frühwirth-Schnatter and Pyne (2010), we write the density of a multivariate skew t-distribution as:

$$f_{\mathcal{ST}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\eta}, \nu) = 2 f_{\mathcal{T}}(\boldsymbol{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \nu) T_{\nu+d}\left(\boldsymbol{\eta}' \boldsymbol{\omega}^{-1} (\boldsymbol{y} - \boldsymbol{\xi}) \sqrt{\frac{\nu + d}{\nu + Q_y}}\right) \tag{2.5}$$

with $\boldsymbol{\omega} = \sqrt{Diag(\boldsymbol{\Omega})}$, $Q_y = (\boldsymbol{y} - \boldsymbol{\xi})' \boldsymbol{\Omega}^{-1} (\boldsymbol{y} - \boldsymbol{\xi})$, $f_{\mathcal{T}}$ the multivariate Student t-distribution density, and $T_{\nu+d}$ the cumulative distribution function of the standard univariate Student's t-distribution with $\nu + d$ degrees of freedom. This parametrization of the skew t is referred as the restricted multivariate skew t distribution by Lee and McLachlan (2013), and it admits the following random-effect model representation:

$$\boldsymbol{Y} = \boldsymbol{\xi} + \boldsymbol{\psi} \frac{Z}{\sqrt{W}} + \frac{\boldsymbol{\epsilon}}{\sqrt{W}} \tag{2.6}$$

2.2.3. *Dirichlet process mixture of multivariate skew t-distributions.* Combining model (2.4) with a random-effects model representation (2.6) of the skew t-distribution, we propose the following model:

$$\alpha | a, b \sim \text{Gamma}(a, b) \tag{2.7a}$$
$$\boldsymbol{\pi} \,\big|\, \alpha \sim \text{GEM}(\alpha) \tag{2.7b}$$

for $k = 1, 2, \ldots$

$$\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k, \nu_k \sim G_0 \tag{2.7c}$$

for $c = 1, 2, \ldots, C$

$$\ell_c \,|\, \boldsymbol{\pi} \sim \text{Mult}(\boldsymbol{\pi}) \tag{2.7d}$$

$$\gamma_c \,|\, \ell_c, \{\nu_k\} \sim \text{Gamma}\left(\frac{\nu_{\ell_c}}{2}, \frac{\nu_{\ell_c}}{2}\right) \tag{2.7e}$$

$$s_c \,|\, \gamma_c \sim \mathcal{N}_{[0,+\infty[}\left(0, \frac{1}{\gamma_c}\right) \tag{2.7f}$$

$$\boldsymbol{y}_c \,|\, \ell_c, \gamma_c, s_c, (\boldsymbol{\xi}_k, \boldsymbol{\psi}_k, \boldsymbol{\Sigma}_k) \sim \mathcal{N}\left(\boldsymbol{\xi}_{\ell_c} + \boldsymbol{\psi}_{\ell_c} s_c, \frac{1}{\gamma_c}\boldsymbol{\Sigma}_{\ell_c}\right) \tag{2.7g}$$

where G_0 is the product of a structured normal-inverse-Wishart ($sNiW$) and of a prior on ν: $G_0 = sNiW(\xi_0, \psi_0, B_0, \Lambda_0, \lambda_0)P_{0,\nu}$.

2.3. *Discussion on the model assumptions.* In model (2.7), the base distribution parameter G_0 conveys the prior information on the cluster parametric shape. For the parameters $\boldsymbol{\xi_k}$, $\boldsymbol{\psi_k}$ and $\boldsymbol{\Sigma_k}$, we have conditional conjugacy with the random-effects model representation using joint priors taking the form of a structured normal-inverse-Wishart distribution. See Online Supplement A for details (Hejblum et al., 2018). Frühwirth-Schnatter and Pyne (2010) pointed out that the prior on $\boldsymbol{\Sigma}_k$ can have a big impact on the posterior number of clusters. Indeed, setting the scale of the prior on $\boldsymbol{\Sigma}_k$ too small will result in an inflated number of clusters in the posterior, whereas too large values tend to cluster all the observations together. Adding a Wishart hyperprior on $\boldsymbol{\Sigma}_k$, that carries on conjugacy with the inverse-Wishart, enables us to reduce this impact of the prior (Frühwirth-Schnatter and Pyne, 2010; Huang and Wand, 2013). Assuming prior independence between each ν_k and also from the three parameters mentioned above, we can use any of the three priors proposed in Juárez and Steel (2010) for instance (such as an objective Jeffrey's prior, see Online Supplement A (Hejblum et al., 2018)).

3. Estimation.

3.1. *Posterior Estimation via Gibbs sampling.* For making inference on the model (2.7), MCMC methods can be used to sample the partition $\{\ell_{1:C}\}$ and the corresponding cluster parameters $\{\theta_k^*\} = \{\{\boldsymbol{\xi}_k^*\}, \{\boldsymbol{\psi}_k^*\}, \{\boldsymbol{\Sigma}_k^*\}, \{\nu_k^*\}\}$ from the marginal posterior distribution. Extending results from Frühwirth-Schnatter and Pyne (2010) and Caron, Teh and Murphy (2014), it is possible

to implement an efficient and valid partially collapsed Gibbs sampler with a Metropolis-Hastings step (van Dyk and Park, 2008; van Dyk and Jiao, 2015). The use of slice sampling (Neal, 2003; Kalli, Griffin and Walker, 2011) enables the straightforward parallelization of the latent allocation sampling (thanks to conditional conjugacy) in such an MCMC algorithm (even in the skew normal and skew t cases), which can lead to substantial computation speed up when the number of observations C (cells) per sample increases. Each iteration of our Gibbs sampler proceeds in the following order (details are provided in Online Supplement A (Hejblum et al., 2018)):

1. Update the concentration parameter α given the previous partition $\{\ell_{1:C}\}$ using the data augmentation technique from Escobar and West (1995).
2. Update the mixing distribution G given α, $\{\boldsymbol{\xi}_k\}$, $\{\boldsymbol{\psi}_k\}$, $\{\boldsymbol{\Sigma}_k\}$ and the base distribution G_0 via slice sampling.
3. For $c = 1, \ldots, C$ update the individual skew parameter s_c given $\{\boldsymbol{\xi}_k\}$, $\{\boldsymbol{\psi}_k\}$, $\{\boldsymbol{\Sigma}_k\}$ and the new ℓ_c.
4. Update $\{\boldsymbol{\xi}_k\}$, $\{\boldsymbol{\psi}_k\}$, $\{\boldsymbol{\Sigma}_k\}$ given the base distribution G_0, the updated partition $\{\ell_{1:C}\}$ and the updated individual skew parameters $\{s_{1:C}\}$.
5. Finally jointly update the degrees of freedom and the individual scale factors $(\{\nu_k\}, \{\gamma_{1:C}\})$ in an Metropolis-Hastings (M-H) within Gibbs step. First an M-H step is performed to update the $\{\nu_k\}$ where the $\{\gamma_{1:C}\}$ are integrated out, immediately followed by a Gibbs step to sample the $\{\gamma_{1:C}\}$ from their full conditional distribution. This ensures that the reduced conditioning performed in the M-H step does not change the stationary distribution of the Markov chain (van Dyk and Jiao, 2015) – see Online Supplement A (Hejblum et al., 2018).

3.2. *Sequential Posterior Approximation.* In flow cytometry experiments, it is common to actually have multiple datasets $\boldsymbol{y}^{(i)}$ (with $i = 1, \ldots, I$) corresponding to multiple individuals or repeated measurements of the same individual. In such cases, it is of interest to use previous time points or previous samples results as prior information, in order to leverage all the information available to estimate the mixture. However, incorporating prior information into Dirichlet process mixture models is not straightforward (Kessler, Hoff and Dunson, 2015). Here we propose to use the posterior MCMC draws obtained from previous dataset $\boldsymbol{y}^{(i)}$ as prior information to analyze the next dataset $\boldsymbol{y}^{(i+1)}$. To do so, first we consider the hierarchical

model using all observations from both $\boldsymbol{y}^{(i)}$ and $\boldsymbol{y}^{(i+1)}$ at once :

$$\alpha \sim Gamma(a, b) \tag{3.1a}$$

$$G|\alpha \sim DP(\alpha, G_0) \tag{3.1b}$$

$$\boldsymbol{y}^{(i)}, \boldsymbol{y}^{(i+1)}|G \overset{i.i.d.}{\sim} \int_{\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\cdot) dG(\boldsymbol{\theta}) \tag{3.1c}$$

We are interested in estimating $p(G|\boldsymbol{y}^{(i)}, \boldsymbol{y}^{(i+1)}) \propto p(G|\boldsymbol{y}^{(i)}) p(\boldsymbol{y}^{(i+1)}|G)$. The idea is to first approximate $p(G|\boldsymbol{y}^{(i)})$ by a Dirichlet process through MCMC draws from the model described in 2.1:

$$p(G|\boldsymbol{y}^{(i)}) \simeq \int DP(G; \alpha, G_1) Gamma(\alpha; a_1, b_1) d\alpha \tag{3.2}$$

where G_1, a_1, b_1 are parameters to be estimated from the MCMC approximation of the true posterior: i) $\widehat{a_1}$ and $\widehat{b_1}$ can be taken as MLE estimates from the MCMC samples $\alpha^{(j)}$; ii) $\widehat{G_1}$ is a parametric approximation of the posterior mixing distribution G_1 (the true posterior is not suitable for being directly plugged in as a base distribution parameter of another DP as it is nonparametric). In the case of a skew t-distribution mixture model, we approximate G_1 with the following joint distribution: $G_1 \simeq (sNiW, P_{0,\nu})$ where $P_{0,\nu}$ is the chosen prior for the skew t-distribution degrees of freedom. To estimate G_1, we estimate the maximum *a posteriori* (MAP) from the posterior MCMC samples (see Online Supplement B (Hejblum et al., 2018)).

Now using this posterior parametric approximation, we have the same hierarchical model as before but conditional on $\boldsymbol{y}^{(i)}$:

$$\alpha|\boldsymbol{y}^{(i)} \sim Gamma(\widehat{a_1}, \widehat{b_1}) \tag{3.3a}$$

$$G|\alpha, \boldsymbol{y}^{(i)} \sim DP(\alpha, \widehat{G_1}) \tag{3.3b}$$

$$\boldsymbol{y}^{(i+1)}|G, \boldsymbol{y}^{(i)} \overset{i.i.d.}{\sim} \int_{\boldsymbol{\Theta}} f_{\boldsymbol{\theta}}(\cdot) dG(\boldsymbol{\theta}) \tag{3.3c}$$

Note that under this approximate posterior model, the cluster parameters $\boldsymbol{\theta}_k^*$ are i.i.d. according from G_1. Such an approach can be iterated a number of times, if for instance several time points are observed, iteratively approximating the successive posteriors. This approach allows us to finally account for all the previous information in the mixture model estimation. This model hypothesizes that all the data are originating from the same mixture model and as more data are acquired, the successive posteriors will concentrate:

since the overall posterior is our target, this concentration effect is desired. When the data are anticipated to be non stationary, which might be the case in real life applications, we propose a forgetful factor by adding a vague component in the base distribution of the prior.

3.3. *Point estimate of the clustering.* Getting a representation of the partition posterior distribution is difficult (Medvedovic and Sivaganesan, 2002). One can use the maximum *a posteriori*, i.e. using the point estimation from the MCMC sample that maximizes the posterior density. However this ignores all the information about the uncertainty around the partition provided by the Bayesian approach.

Another way is to rather consider a co-clustering posterior probability (or similarity) matrix ζ on each pair (c, d) of observations. Such a matrix can be estimated by averaging the co-clustering matrices from all the explored partitions in the posterior MCMC draws:

$$\widehat{\zeta}_{cd} = \frac{1}{N} \sum_{i=1}^{N} \delta_{\ell_c^{(i)} \ell_d^{(i)}} \tag{3.4}$$

where N is the number of MCMC draws from the posterior, i the MCMC iteration, and $\delta_{kl} = 1$ if $k = l$, 0 otherwise. An optimal partition point estimate $\{\widehat{\ell}_{1:C}\}$ can then be derived in regard of this similarity matrix through stochastic search with the explored partitions in the posterior MCMC draws (Dahl, 2006), by using a pairwise coincidence loss function (Lau and Green, 2007) such as the one proposed by Binder (1978, 1981) which optimizes the Rand index (Fritsch and Ickstadt, 2009):

$$\{\widehat{\ell}_{1:C}\} = \underset{\{\ell_{1:C}^{(i)}\} \in \left\{\{\ell_{1:C}^{(1)}\}, \ldots, \{\ell_{1:C}^{(N)}\}\right\}}{\arg\min} \sum_{c=1}^{C-1} \sum_{d=c+1}^{C} 2 \left(\delta_{\ell_c^{(i)} \ell_d^{(i)}} - \widehat{\zeta}_{cd}\right)^2 \tag{3.5}$$

The computational complexity of this approach however is of the order $\mathcal{O}(NC^2)$ due to the necessity of computing all the similarity matrices.

A different optimal partition point estimate $\{\tilde{\ell}_{1:C}\}$ can also be derived using the $\mathcal{F}$-measure as our loss function. The $\mathcal{F}$-measure is widely used as a way to summarize the accordance between 2 methods, one being considered as a reference (gold-standard). It is the harmonic mean of the precision and recall:

$$\mathcal{F} = \frac{2PrRe}{Pr + Re} \tag{3.6}$$

In order to use the $\mathcal{F}$-measure to evaluate our clustering method, we rely on the definition proposed in the online methods from Aghaeepour et al. (2013).

In this unsupervised clustering setting, the precision Pr is the number of cells correctly assigned to a given cluster divided by the total number of cells assigned to that cluster (also called Positive Predictive Value). The recall Re is the number of cells correctly assigned to a given cluster divided by the number of cells that should be assigned to this cluster according to the gold-standard. Since in our problem the labels of the different clusters are exchangeable, the $\mathcal{F}$-measure is computed for each combination of the reference clusters and the predicted clusters. Let $\mathcal{G} = \{g_1, \ldots, g_m\}$ be a set of m reference clusters and $H = \{h_1, \ldots, h_n\}$ be set of n predicted clusters. For each combination pair of a reference cluster g_q and a predicted cluster h_r, the $\mathcal{F}$-measure is computed as follows:

$$Pr(h_r, g_q) = \frac{|g_q \cap h_r|}{|h_r|} \quad \text{and} \quad Re(h_r, g_q) = \frac{|g_q \cap h_r|}{|g_q|} \tag{3.7}$$

$$\mathcal{F}(h_r, g_q) = \frac{2Pr(g_q, h_r)Re(g_q, h_r)}{Pr(g_q, h_r) + Re(g_q, h_r)} \tag{3.8}$$

This $\mathcal{F}$-measure is comprised in $[0, 1]$ and the closer it is to 1 the better the agreement is between the predicted cluster and the reference cluster. The total $\mathcal{F}$-measure for a predicted partition H given a gold-standard G is then defined as the weighted sum of the best matched $\mathcal{F}$-measure:

$$\mathcal{F}_{tot}(H, \mathcal{G}) = \frac{1}{\sum_{q=1}^{m} |g_q|} \sum_{q=1}^{m} |g_q| \max_{r \in \{1,\ldots,n\}} \mathcal{F}(h_r, g_q) \tag{3.9}$$

This total $\mathcal{F}$-measure is again between 0 and 1, and the closer it is to 1 the better the predicted partition agrees with the gold-standard. The optimal partition point estimate in respects of this $\mathcal{F}$-measure is then obtained with the partition that maximizes its average $\mathcal{F}$-measure over all the other explored partitions in the posterior MCMC draws:

$$\{\tilde{\ell}_{1:C}\} = \underset{\{\ell_{1:C}^{(i)}\} \in \left\{\{\ell_{1:C}^{(1)}\}, \ldots, \{\ell_{1:C}^{(N)}\}\right\}}{\arg\max} \frac{1}{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathcal{F}_{tot}\left(\{\ell_{1:C}^{(i)}\}, \{\ell_{1:C}^{(j)}\}\right) \tag{3.10}$$

Note the $\mathcal{F}$-measure is computed here only between sampled partitions, and a gold-standard partition is unnecessary.

4. Simulation Study.

4.1. *Weakly informative prior.* First, to assess the performance of the Dirichlet process mixture of skew t distributions model in a simple clustering case, 100 simulations in 2-dimensions were performed. In each simulation, 10,000 observations were drawn from 5 distinct clusters representing respectively 50%, 29.9%, 15%, 5%, and 0.1% of the data. After 20,000 MCMC iterations (18,000 iterations burnt and a thining of 20 gave 100 partitions sampled from the posterior; the chain was initialized with 30 clusters), we compared the partition point estimate obtained from our approach with the true clustering of the simulated data through the resulting mean $\mathcal{F}$-measure. When the clustering problem is well characterized, i.e. when the true clusters are well separated (Tibshirani, Walther and Hastie, 2001), the median $\mathcal{F}$-measure was 0.998. When considering overlapping clusters, i.e. when the true clustering is not entirely recoverable, our approach was nevertheless able to maintain good performance with a median $\mathcal{F}$-measure of 0.895. Figure 1 shows an example of the partition point estimate obtained for one of those simulation runs in both cases, where one can see that NPflow is able to correctly recover the 5 clusters, including the extremely small one of 0.1% (green diamond). Of course, if this extremely small cluster was not well separated from larger clusters in Figure 1B, the data would hardly contain any evidence of its presence and the model would likely not recover it. As a comparison, k-means only reached a $\mathcal{F}$-measure of 0.920 and 0.823 in the well-separated and overlapping scenarios respectively, in spite of having the correct number (5) of true clusters specified.

Comparison between skew-t and Gaussian kernels. Figure 2 illustrates the improvement due to the use of a skew t kernel over a Gaussian kernel in a nonparametric mixture model. Figures 2A and 2C both show that the $\mathcal{F}$-measure is significantly better with a skew t kernel than with a Gaussian kernel for well-separated and overlapping settings respectively, while Figures 2B and 2D both show that skew t kernels allow to accurately recover the true number of clusters in a majority of cases for well-separated and overlapping settings respectively. Both criteria also highlight the increased difficulty in the overlapping setting as the clustering problem becomes less well characterized.

4.2. *Sequential posterior approximation plugged-in as informative prior.* To illustrate how the sequential posterior approximation strategy compares to the standard weakly informative prior setting, we ran simulations where we considered two samples derived from the same infinite mixture model. The first sample is simulated for a time t, and the second sample at $t+1$. As all observations originate from the exact same distribution, regardless of the

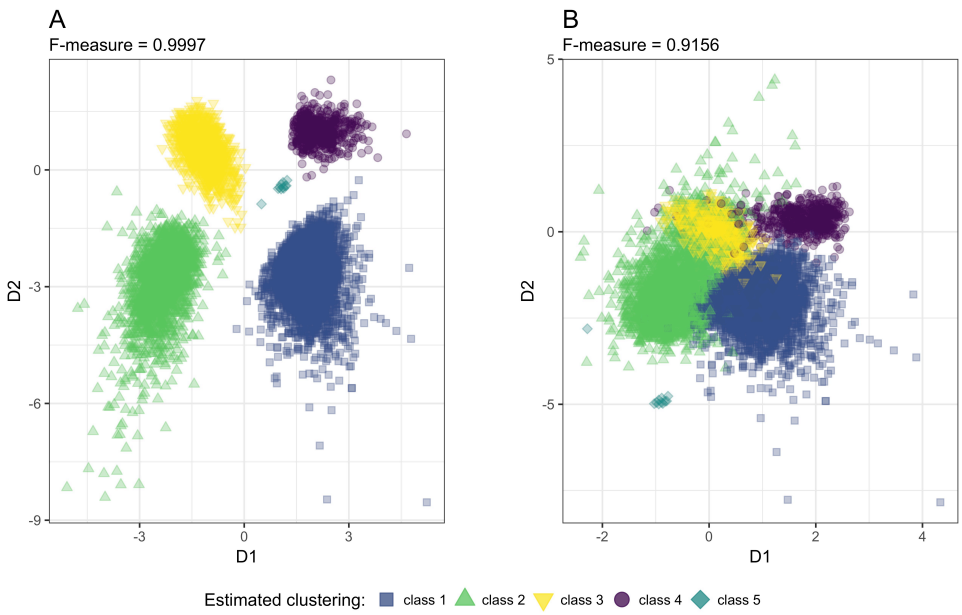

FIG 1. *Partition point estimate from one of the 100 2-dimensional simulations with weakly informative prior (10,000 observations with 5 clusters representing respectively 50%, 29.9%, 15%, 5% and 0.1% of the data). A: well separated clusters ; B: overlapping clusters*

sample, the hypothesis of the sequential posterior approximation strategy is satisfied. One of the major gain observed is the time to convergence for the partition. Using an informative prior derived from the sample at time t to estimate the partition of the sample from $t+1$ makes it more than three time faster to converge according to the Gelman-Rubin statistics.

In further simulations, we also investigated the performance of this sequential posterior approximation strategy. As opposed to using the standard weakly informative prior strategy, it shows substantial gains when the amount of information brought by the prior is substantial compared to the amount available from the data at time $t+1$ alone. As the amount of information available at time $t+1$ increases, the gain from using this strategy can become less noticeable, as shown using the $\mathcal{F}$-measure in Figure 3. But even when the number of observations available at time $t+1$ is the same as at time t, the accuracy for rare cell populations is still improved by using an informative prior. This is not necessary visible at the scale of the total $\mathcal{F}$-measure, because it is masked by the larger clusters. However, when computing a limited $\mathcal{F}$-measure, that only takes into account smaller clusters, the use of an informative prior in this sequential strategy seems to always im-

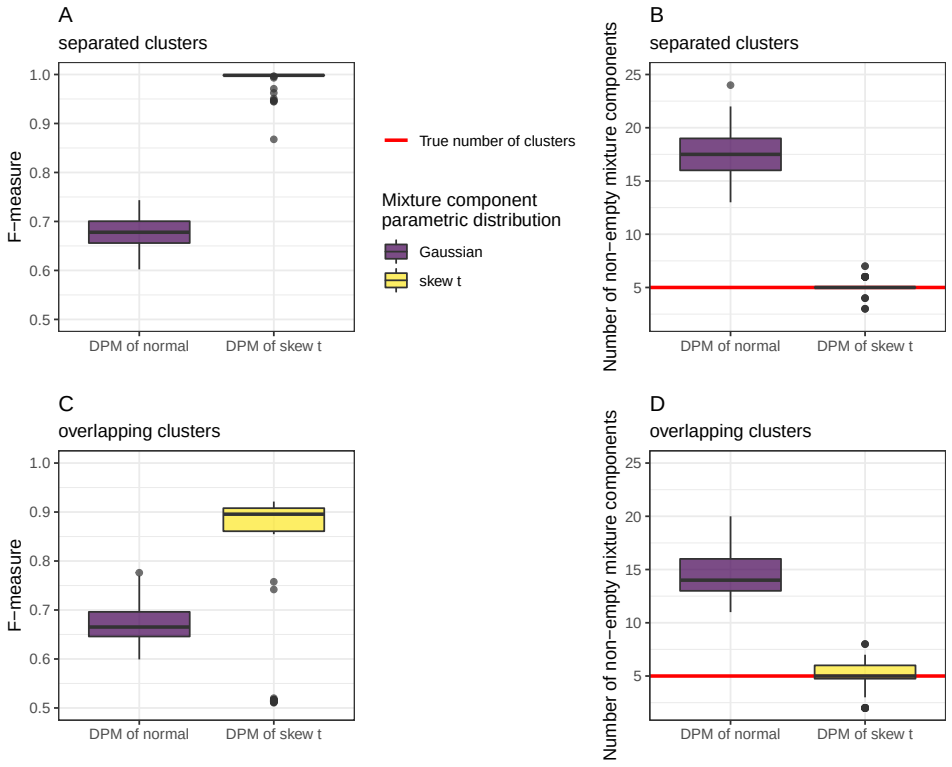

FIG 2. *Comparison of a Gaussian kernel versus a skew t kernel. A, B: well separated clusters ; C,D: overlapping clusters*

prove the clustering accuracy for smaller clusters (see Supplementary Figure S1 in Online Supplement C (Hejblum et al., 2018)).

5. Application to real datasets.

5.1. *Benchmark dataset.* The Graft versus Host Disease (GvHD) dataset is a public dataset that was first analyzed (manually gated) in Brinkman et al. (2007), with the objective of identifying a cellular signature that correlates or predicts the Graft versus Host disease. These GvHD data were used as benchmark data in the FlowCAP challenge (Aghaeepour et al., 2013). Flow cytometry data was collected for 12 samples, and original manual gates are being regarded as the true cell clustering (actually a consensus over eight different manual operators). In an attempt to mitigate further the well known reproducibility issues with manual gating (Ge and Sealfon, 2012; Aghaeepour et al., 2013), only the most concordant clusters

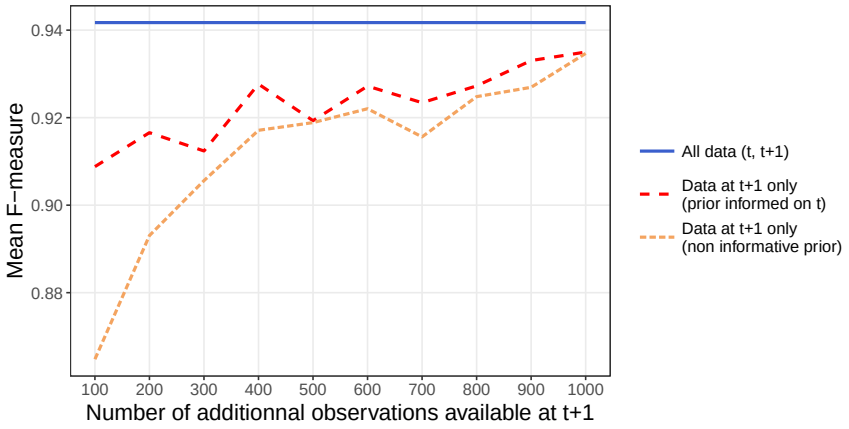

FIG 3. *Mean $\mathcal{F}$-measure according to the number of observations available at time $t+1$, while 1,000 observations are available at time t, over 300 simulations*

between the eight gatings (i.e. with an $\mathcal{F}$-measure above 0.8) were used for comparison with the automated results, as was done in Aghaeepour et al. (2013). The data were downloaded from the FlowCAP project website [http://flowcap.flowsite.org/] as part of the FlowCAP-I challenge [http://flowcap.flowsite.org/codeanddata/FlowCAP-I.zip]. Table 1 shows the performance of our proposed approach NPflow on this dataset, compared to the other approaches reviewed by Aghaeepour et al. (2013). The $\mathcal{F}$-measure is computed for all samples available for a given dataset and the mean over all samples is reported, as well as a bootstrap 95% confidence interval. No algorithm is performing significantly better than NPflow thus placing NPflow among the top methods for automatic gating. Additionally, we compared the use of a skew t kernel by NPflow with the use of a Gaussian kernel (denoted NPflowG in Table 1), the latter reaching a mere 0.61 $\mathcal{F}$-measure on average thus demonstrating the benefit of the skew t distributions for modeling real flow-cytometry data.

Thanks to its use of sequential information, the sequential posterior model would ideally improve results by analyzing each individual sample sequentially. The GvHD benchmark data are not longitudinal but, as long as the different samples are similar enough, one can expect an improvement. On this benchmark, our sequential approach NPflow-seq reaches a mean $\mathcal{F}$-measure of 0.89 (0.85, 0.94) compared to a value of 0.85 (0.80, 0.90) with the standard NPflow model (Table 1). Of note, the only other approach not analyzing the samples independently, and that relies on a hierachical Dirichlet process Gaussian mixture model (HDPGMM), only reaches a value of

TABLE 1
Mean $\mathcal{F}$-measures across all the 12 samples from the GvHD benchmark dataset

Method	$\mathcal{F}$-measure*
NPflow	0.85 (0.80, 0.90)
NPflow-seq$^{\measuredangle}$	0.89 (0.85, 0.94)
NPflowG	0.61 (0.57, 0.66)
ADICyt†	0.81 (0.72, 0.88)
CDP†	0.52 (0.46, 0.58)
FLAME†	0.85 (0.77, 0.91)
FLOCK†	0.84 (0.76, 0.90)
flowClust/Merge†	0.69 (0.55, 0.79)
flowMeans†	0.88 (0.82, 0.93)
FlowVB†	0.85 (0.79, 0.91)
L2kmeans†	0.64 (0.57, 0.72)
MM†	0.83 (0.74, 0.91)
MMPCA†	0.84 (0.74, 0.93)
SamSPECTRAL†	0.87 (0.81, 0.93)
SWIFT†	0.63 (0.56, 0.70)
HDPGMM$^{\ddagger\measuredangle}$	0.35 (0.30, 0.39)

*95% Confidence Intervals are calculated on 10,000 bootstrap samples of the $\mathcal{F}$-measures.
$^{\measuredangle}$methods that do not analyze the 12 samples independently.
†estimates from Aghaeepour et al. (2013).
‡estimates are from Johnsson, Wallin and Fontes (2016).

0.35 (0.30, 0.39) Cron et al. (2013); Johnsson, Wallin and Fontes (2016). This illustrates that integrating all samples in a simultaneous model does not necessarily yield better results (e.g. if the global model across samples is misspecified or not flexible enough). For the GvHD benchmark dataset, our sequential strategy thus exhibits the highest $\mathcal{F}$-measure compared to standard NPflow and to HDPGMM, and also to competing unsupervised automatic gating methods evaluated in Aghaeepour et al. (2013) that however analyze each sample independently. It is worth noting that since our sequential strategy performs sequential approximations of the posterior and provides intermediate results for each sample, the order in which observations are included can have an impact (especially for the first sample). Here we analyzed the GvHD samples in their original order as provided by their identifiers in FlowCAP-I (from 001 to 012).

5.2. *Original DALIA-1 data: a longitudinal real data study.* We applied our method to analyze an original dataset from DALIA-1, a phase I trial evaluating a therapeutic vaccine against HIV (Lévy et al., 2014). For our purpose here, we are interested in the 12 HIV positive patients who had their cellular populations quantified at 18 time points during the trial. More specifically, we focused on two time points (at week 24 and week 26 of the

trial) immediately following antiretroviral treatment (HAART) interruption which took place at week 24. Following this interruption, the increase of viral replication is associated with changes in cell populations (Thiébaut et al., 2005; Lévy et al., 2012). Here we especially looked at the CD4+ effector T-cells, defined as CD45RA+CD27- among the CD3+CD4+ cells (Larbi and Fulop, 2014), that are one of the first cell populations to be affected during the viral rebound (Lévy et al., 2012). Since flow-cytometry measurements were repeated at each time point for each patient, we used the sequential strategy at week 26, in the hope to use the information from week 24 to better identify the CD4+ effector T-cell population at the next time point. Figure 4 illustrates the overall efficiency gain at week 26 from using the sequential strategy. The average limited $\mathcal{F}$-measure (considering available manual gating as gold-standard) on those 12 samples is 0.58 for NPflow with a non-informative prior, and increases to 0.63 with the sequential strategy. By comparison, flowMeans (the second best method on the benchmark GvHD dataset) gives an average limited $\mathcal{F}$-measure of 0.49 (see Online Supplement D for details). We also compared our approach to the HDPGMM proposed by Cron et al. (2013) that is specifically focusing on small cell populations (even if it had the lowest $\mathcal{F}$-measure on the benchmark dataset - see Table 1). In spite of this example representing its ideal use-case, it performed slightly worse than our approach giving an average limited $\mathcal{F}$-measure of 0.62 (see Online Supplement D for more details (Hejblum et al., 2018)). Figure 5 gives an example of a patient for who the sequential strategy was especially improving the identification of the CD4+ effector T-cells. In this case, the percentage of CD4+ effector T-cells was estimated at 31.7 by the manual gating, at 7.6 by NPflow, and at 38.1 by the sequential strategy. Figure 6 shows the slight increase (of about 2%) of CD4+ effector T-cell proportions after treatment interruption (see Online Supplement D for more details (Hejblum et al., 2018)).

In addition to providing a point estimate of the partition, our method also quantifies the uncertainty around the posterior clustering through posterior co-clustering probabilities. Figure 7 displays such a co-clustering posterior probability matrix where we can clearly identify four core clusters, with some uncertainty between two overlapping clusters that both belong to the CD4 effector T-cell population (with moderate and high expression of the CD45RA marker, respectively).

6. Discussion. We analyzed longitudinal flow cytometry data from the DALIA-1 trial, focusing on the CD4+ effector T-cell population among 12 HIV positive patients. Compared to state-of-the-art automatic gating ap-

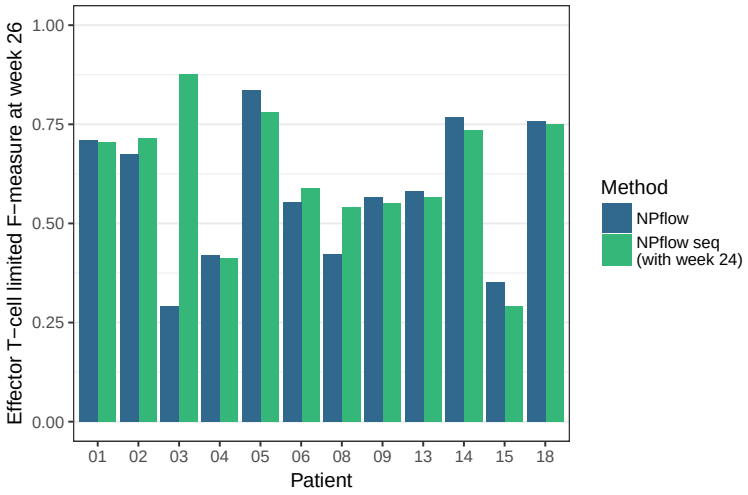

FIG 4. *Limited $\mathcal{F}$-measures for the CD4+ effector T-cell population from the DALIA-1 trial two weeks after HAART interruption for NPflow with or without the sequential strategy, compared to manual gating.*

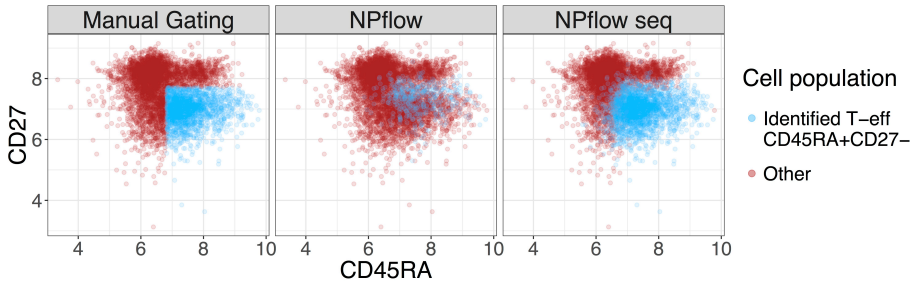

FIG 5. *CD3+CD4+ cells of patient 3 from the DALIA-1 trial two weeks after HAART interruption (at week 26).*

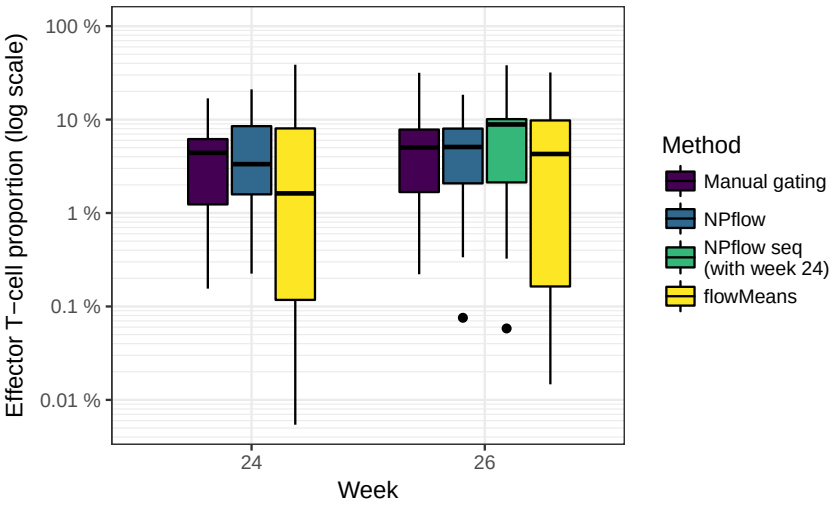

FIG 6. *Proportion of CD4+ effector T-cells in the DALIA-1 trial following HAART interruption (from manual gating).*

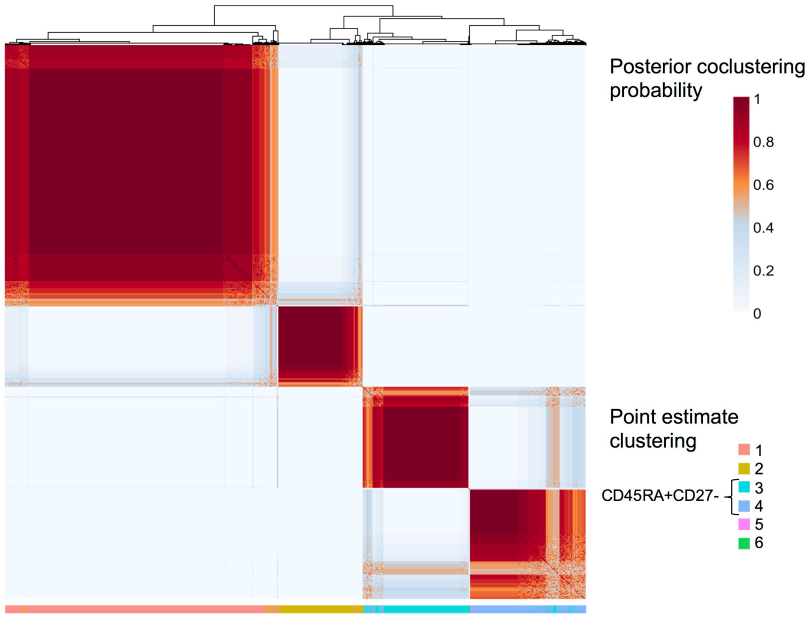

FIG 7. *Heatmap of the posterior co-clustering probabilities for the CD3+CD4+ cells of patient 3 at week 26 from DALIA-1 with NPflow seq.*

proaches, our sequential strategy using Dirichlet process mixtures of multivariate skew t-distributions allowed a better recovery of the effector T-cell population after a meaningful perturbation of this targeted population following HAART interruption, highlighting their expected increase.

Our proposed method extends the classical multivariate Dirichlet process Gaussian mixture model to multivariate skew t-distribution mixtures, based on Frühwirth-Schnatter and Pyne (2010) parametrization of the restricted multivariate skew t-distribution (Lee and McLachlan, 2013). Automatic gating of cell populations from flow cytometry data is an open research problem and the proposed approach features two important characteristics for this task: i) it avoids the difficult issue of model selection by estimating directly the number of components in the mixture ; ii) it uses skew and heavy tailed distributions in the form of skew t-distributions, of which the skew normal and the normal are particular cases. Further domain-knowledge can be incorporated in the proposed model by specifying more informative priors on the Dirichlet process parameters for instance. Thanks to the use of the rMST formulation of the skew t, we avoid the identifiability issues mentioned by Lee and McLachlan (2016). Estimation of the pairwise posterior co-clustering probabilities allows to quantify the uncertainty about the posterior partition, and an optimal point estimate of the clustering is provided by minimizing a cost function in regards to the average posterior co-clustering matrix. We have developed and implemented an efficient collapsed Metropolis within Gibbs sampler for estimating such models. One of the advantages of our proposed sampler is the absence of label switching issue, as it uses directly the partition of the data without having to deal with labels (Jasra, Holmes and Stephens, 2005). The computational cost of fitting our model is linear in the number of observations as well as in the number of clusters, whilst the computational cost of the partition point estimate depends of the optimal criterion chosen. The use of state-of-the-art MCMC techniques along with inner parallelization allow us to mitigate the computational cost that comes with such approaches on large data. As an indication of runtime, around 3,000 MCMC iterations can be run on average for a real dataset of around 30,000 observations over 6 dimensions, using one Intel® Xeon® x5675 processor for one hour. Besides, instead of using a partially collapse Gibbs sampler algorithm, it could be of interest to also investigate the use of sequential Monte-Carlo algorithms, especially for the sequential modeling strategy or other possible dynamic extensions of the model proposed here (Caron et al., 2008, 2017).

In case of repeated measurements of flow cytometry data, we propose to use sequential parametric approximations of the posterior as refined infor-

mative priors. The proposed sequential analysis strategy enables to analyze each sample sequentially, as the data are acquired. It does not require to wait for the last sample to perform the automatic gating nor to analyze all data at once, but it still uses available prior knowledge. This contrasts with hierarchical extensions of the Dirichlet Process Mixture Model such as those proposed by Cron et al. (2013) or Dundar et al. (2014), where the complete dataset must be analyzed at once. This sequential strategy allows one to analyze the samples as they are acquired, which can be useful in clinical trials where there are often intermediate analyses for instance. Moreover in large studies the size of the data can make it challenging to analyze all samples at once, and such a sequential approach then makes practical sense (Huang and Gelman, 2005). Futhermore, this use of sequentially informed priors does not face the usual complications of cluster matching arising when an algorithm is run on each sample separately (Cron et al., 2013). In our simulation study, this sequential posterior approximation strategy improves the fit of the model. In addition, such a strategy exhibits accelerated convergence and greater accuracy for small clusters, as long as the different samples are similar enough. Besides, the parametric prior can also be specified to inform the model with expert knowledge, e.g. to favor a range for the expected number of clusters. On real flow-cytometry, data we showed that the sequential strategy also improves the clustering performances. On the benchmark dataset, it outperformed all other methods investigated in by Aghaeepour et al. (2013), and in the DALIA-1 trial, the sequential strategy also improved the automatic gating results. It is worth noting however that in other cases, for instance if the data distributions were too different between samples, the sequential posterior model would not necessarily improve the clustering results, and could even gave a diminished F-measure compared to the non sequential strategy.

Manual gating is still considered the gold-standard when evaluating an automatic gating strategy on real flow cytometry data. Yet one should keep in mind that manual gating has reproducibility issues, often resulting in a partial and subjective clustering (Ge and Sealfon, 2012; Welters et al., 2012; Aghaeepour et al., 2013; Gondois-Rey et al., 2016). Therefore using manual gating as the gold-standard might not be actually the best way to assess the performance of automatic gating algorithms on real data, because of its inherent flaws.

Mass cytometry is a technology very similar to flow cytometry. Using ions in place of colors, CyTOF is able to measure up to 40 cell markers at once, generating even more data than flow cytometry. Efficient automated gating method are therefore all the more needed in the context of

CyTOF(Melchiotti et al., 2017). The approach proposed here could be directly applied to such data. More generally, we propose here a framework for Dirichlet process mixtures of multivariate skew t-distributions modeling that is suitable for any kind of data modeled as such a mixture, especially when the number of mixture components is unknown. We provide an efficient implementation of our method within the R package `NPflow` that is available on CRAN at https://CRAN.R-project.org/package=NPflow.

Acknowledgements. The authors are extremely grateful to Jean-Louis Palgen for his time and efforts to manually gate the effector T-cells in the DALIA-1 trial at weeks 24 and 26, as well as to the DALIA-1 study group. The authors also thank Cedric Lachat for his precious help in getting the `dpmix` software to run despite its lack of maintenance. The authors also thank Nima Aghaeepour for his help in using supplementary data provided with his publication (Aghaeepour et al., 2013). Boris P. Hejblum was a recipient of a Ph.D. fellowship from the École des Hautes Études en Santé Publique (EHESP) Doctoral Network. Part of this work has been supported by the BNPSI ANR project n° ANR-13-BS-03-0006-01. Part of this work has been supported by the IMI2 grant EBOVAC2. Computer time for this study was partly provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour.

SUPPLEMENTARY MATERIAL

Online Supplement to "Sequential Dirichlet Process Mixtures of Multivariate Skew t-distributions for Model-based Clustering of Flow Cytometry Data"
(doi: COMPLETED BY THE TYPESETTER; .pdf). We provide additional mathematical details for the proposed Gibbs samplers and the parameter estimations, as well as additional plots showing the good performance of the sequential strategy.

References.

Aghaeepour, N., Nikolic, R., Hoos, H. H. and Brinkman, R. R. (2011). Rapid cell population identification in flow cytometry data. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **79** 6–13.

Aghaeepour, N., Finak, G., Hoos, H., Mosmann, T. R., Brinkman, R. R., Gottardo, R. and Scheuermann, R. H. (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nature methods* **10** 228–238.

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 367–389.

Azzalini, A. and Valle, A. D. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726.

Azzalini, A., Browne, R. P., Genton, M. G. and McNicholas, P. D. (2016). On nomenclature for, and the relative merits of, two formulations of skew distributions. *Statistics and Probability Letters* **110** 201–206.

Biernacki, C., Celeux, G. and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence* **22** 719–725.

Binder, D. A. (1978). Bayesian Cluster Analysis. *Biometrika* **65** 31–38.

Binder, D. A. (1981). Approximations to Bayesian Clustering Rules. *Biometrika* **68** 275–285.

Brinkman, R. R., Gasparetto, M., Lee, S.-J. J., Ribickas, A. J., Perkins, J., Janssen, W., Smiley, R. and Smith, C. (2007). High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation* **13** 691–700.

Caron, F., Teh, Y. W. and Murphy, T. B. (2014). Bayesian nonparametric Plackett-tLuce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics* **8** 1145–1181.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2008). Bayesian Inference for Linear Dynamic Models With Dirichlet Process Mixtures. *IEEE Transactions on Signal Processing* **56** 71–84.

Caron, F., Neiswanger, W., Wood, F., Doucet, A. and Davy, M. (2017). Generalized Pólya Urn for Time-Varying Pitman-Yor Processes. *Journal of Machine Learning Research* **18** 1–32.

Chan, C., Feng, F., Ottinger, J., Foster, D., West, M. and Kepler, T. B. (2008). Statistical mixture modeling for cell subtype identification in flow cytometry. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **73** 693–701.

Cron, A., Gouttefangeas, C., Frelinger, J., Lin, L., Singh, S. K., Britten, C. M., Welters, M. J. P., van der Burg, S. H., West, M. and Chan, C. (2013). Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples. *PLoS computational biology* **9** e1003130.

Dahl, D. B. (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In *Bayesian Inference for Gene Expression and Proteomics* (K.-A. Do, P. Müller and M. Vannucci, eds.) 10, 201–218. Cambridge University Press, Cambridge.

Dundar, M., Akova, F., Yerebakan, H. Z. and Rajwa, B. (2014). A non-parametric Bayesian model for joint cell clustering and cluster matching: identification of anomalous sample phenotypes with random effects. *BMC Bioinformatics* **15** 314.

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90** 577–588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1** 209–230.

Finak, G., Bashashati, A., Brinkman, R. and Gottardo, R. (2009). Merging mixture components for cell population identification in flow cytometry. *Advances in bioinformatics* **2009** 247646.

Finak, G., Perez, J.-M., Weng, A. and Gottardo, R. (2010). Optimizing transformations for automated, high throughput analysis of flow cytometry data. *BMC Bioinformatics* **11** 546.

Fritsch, A. and Ickstadt, K. (2009). Improved criteria for clustering based on the

posterior similarity matrix. *Bayesian Analysis* **4** 367–392.

Frühwirth-Schnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics* **11** 317–36.

Ge, Y. and Sealfon, S. C. (2012). flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics* **28** 2052–2058.

Gondois-Rey, F., Granjeaud, S., Rouillier, P., Rioualen, C., Bidaut, G. and Olive, D. (2016). Multi-parametric cytometry from a complex cellular sample: Improvements and limits of manual versus computational-based interactive analyses. *Cytometry Part A* **89** 480–490.

Hejblum, B. P., Alkhassim, C., Gottardo, R., Caron, F. and Thiébaut, R. (2018). Supplement to "Sequential Dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data".

Huang, Z. and Gelman, A. (2005). Sampling for Bayesian Computation with Large Datasets. *SSRN Electronic Journal* 1–21.

Huang, A. and Wand, M. P. (2013). Simple Marginally Noninformative Prior Distributions for Covariance Matrices. *Bayesian Analysis* **8** 439–452.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science* **20** 50–67.

Johnsson, K., Wallin, J. and Fontes, M. (2016). BayesFlow: latent modeling of flow cytometry cell populations. *BMC Bioinformatics* **17** 25.

Juárez, M. A. and Steel, M. F. J. (2010). Model-Based Clustering of Non-Gaussian Panel Data Based on Skew- t Distributions. *Journal of Business & Economic Statistics* **28** 52–66.

Kalli, M., Griffin, J. E. and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and Computing* **21** 93–105.

Kessler, D. C., Hoff, P. D. and Dunson, D. B. (2015). Marginally specified priors for non-parametric bayesian estimation. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **77** 35–58.

Larbi, A. and Fulop, T. (2014). From "truly naïve" to "exhausted senescent" T cells: When markers predict functionality. *Cytometry Part A* **85** 25–35.

Lau, J. W. and Green, P. J. (2007). Bayesian Model-Based Clustering Procedures. *Journal of Computational and Graphical Statistics* **16** 526–558.

Lee, S. X. and McLachlan, G. J. (2013). On mixtures of skew normal and skew t-distributions. *Advances in Data Analysis and Classification* **7** 241–266.

Lee, S. X. and McLachlan, G. J. (2016). Finite mixtures of canonical fundamental skew t-distributions: The unification of the restricted and unrestricted skew t-mixture models. *Statistics and Computing* **26** 573–589.

Lévy, Y., Thiébaut, R., Gougeon, M.-L., Molina, J.-M., Weiss, L., Girard, P.-M., Venet, A., Morlat, P., Poirier, B., Lascaux, A.-S., Boucherie, C., Sereni, D., Rouzioux, C., Viard, J.-P., Lane, C., Delfraissy, J.-F., Sereti, I., Chêne, G. and ILIADE Study Group (2012). Effect of intermittent interleukin-2 therapy on CD4+ T-cell counts following antiretroviral cessation in patients with HIV. *AIDS* **26** 711–720.

Lévy, Y., Thiébaut, R., Montes, M., Lacabaratz, C., Sloan, L., King, B., Pérusat, S., Harrod, C., Cobb, A., Roberts, L. K., Surenaud, M., Boucherie, C., Zurawski, S., Delaugerre, C., Richert, L., Chêne, G., Banchereau, J. and Palucka, K. (2014). Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load.

European journal of immunology **44** 2802–2810.

Lin, L., Chan, C., Hadrup, S. R., Froesig, T. M., Wang, Q. and West, M. (2013). Hierarchical Bayesian mixture modelling for antigen-specific T-cell subtyping in combinatorially encoded flow cytometry studies. *Statistical Applications in Genetics and Molecular Biology* **12** 309–331.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* **12** 351–357.

Lo, K., Brinkman, R. R. and Gottardo, R. (2008). Automated gating of flow cytometry data via robust model-based clustering. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **73** 321–332.

Lo, K. and Gottardo, R. (2012). Flexible mixture modeling via the multivariate t distribution with the Box-Cox transformation: An alternative to the skew-t distribution. *Statistics and Computing* **22** 33–52.

McLachlan, G. J. and Lee, S. X. (2016). Comment on On nomenclature, and the relative merits of two formulations of skew distributions by A. Azzalini, R. Browne, M. Genton, and P. McNicholas. *Statistics & Probability Letters* **116** 1–5.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* **18** 1194–1206.

Melchiotti, R., Gracio, F., Kordasti, S., Todd, A. K. and de Rinaldis, E. (2017). Cluster stability in the analysis of mass cytometry data. *Cytometry Part A* **91** 73–84.

Mosmann, T. R., Naim, I., Rebhahn, J., Datta, S., Cavenaugh, J. S., Weaver, J. M. and Sharma, G. (2014). SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 2: Biological evaluation. *Cytometry Part A* **85** 422–433.

Murray, P. M., Browne, R. P. and McNicholas, P. D. (2014). Mixtures of skew-t factor analyzer. *Computational Statistics & Data Analysis* **77** 326–335.

Naim, I., Datta, S., Rebhahn, J., Cavenaugh, J. S., Mosmann, T. R. and Sharma, G. (2014). SWIFT-scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, Part 1: Algorithm design. *Cytometry Part A* **85** 408–421.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics* **31** 705–767.

Pitman, J. (2006). *Combinatorial Stochastic Processes. Lecture Notes in Mathematics* **1875**. Springer-Verlag, Berlin Heidelberg.

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 Jager, P. L. and Mesirov, J. P. (2009). Automated high-dimensional flow cytometric data analysis. *Proceedings of the National Academy of Sciences of the United States of America* **106** 8519–8524.

Qian, Y., Wei, C., Eun-Hyung Lee, F., Campbell, J., Halliley, J., Lee, J. a., Cai, J., Kong, Y. M., Sadat, E., Thomson, E., Dunn, P., Seegmiller, A. C., Karandikar, N. J., Tipton, C. M., Mosmann, T., Sanz, I. and Scheuermann, R. H. (2010). Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. *Cytometry. Part B, Clinical cytometry* **78 Suppl 1** S69–82.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4** 639–650.

Sugár, I. P. and Sealfon, S. C. (2010). Misty Mountain clustering: application to fast unsupervised flow cytometry gating. *BMC Bioinformatics* **11** 502.

Teh, Y. W. (2010). Dirichlet Process. In *Encyclopedia of Machine Learning* 280–287.

Springer US, Boston, MA.

Thiébaut, R., Pellegrin, I., Chêne, G., Viallard, J. F., Fleury, H., Moreau, J. F., Pellegrin, J. L. and Blanco, P. (2005). Immunological markers after long-term treatment interruption in chronically HIV-1 infected patients with CD4 cell count above 400 x 10(6) cells/l. *AIDS* **19** 53–61.

Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** 411–423.

van Dyk, D. A. and Jiao, X. X. (2015). Metropolis-Hastings within Partially Collapsed Gibbs Samplers. *Journal of Computational and Graphical Statistics* **24** 301–327.

van Dyk, D. A. and Park, T. (2008). Partially Collapsed Gibbs Samplers. *Journal of the American Statistical Association* **103** 790–796.

Welters, M. J. P., Gouttefangeas, C., Ramwadhdoebe, T. H., Letsch, A., Ottensmeier, C. H., Britten, C. M. and Van Der Burg, S. H. (2012). Harmonization of the intracellular cytokine staining assay. *Cancer Immunology, Immunotherapy* **61** 967–978.

Zare, H., Shooshtari, P., Gupta, A. and Brinkman, R. R. (2010). Data reduction for spectral clustering to analyze high throughput flow cytometry data. *BMC Bioinformatics* **11** 403.

B.P. Hejblum
C. Alkhassim
R. Thiébaut
ISPED - Université de Bordeaux
146 rue Léo Saignat
33076 Bordeaux France
E-mail: boris.hejblum@u-bordeaux.fr
chariff.alkhassim@u-bordeaux.fr
rodolphe.thiebaut@u-bordeaux.fr

R. Gottardo
Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N., Mail Stop M1-B514
Seattle, WA 98109
USA
E-mail: rgottard@fredhutch.org

F. Caron
Department of Statistics
University of Oxford
24-29 St Giles
OX1 3LB, Oxford United Kingdom
E-mail: caron@stats.ox.ac.uk