Article

Single-cell Bayesian deconvolution

Highlights

A Bayesian method robustly removes autofluorescence from flow cytometry data

The method provides unbiased error estimates and natural clustering of the data

The code is publicly available as an efficient Julia package
Individual cells exhibit substantial heterogeneity in protein abundance and activity, which is frequently reflected in broad distributions of fluorescently labeled reporters. Since all cellular components are intrinsically fluorescent to some extent, the observed distributions contain background noise that masks the natural heterogeneity of cellular populations. This limits our ability to characterize cell-fate decision processes that are key for development, immune response, tissue homeostasis, and many other biological functions. It is therefore important to separate the contributions from signal and noise in single-cell measurements. Addressing this issue rigorously requires deconvolving the noise distribution from the signal, but approaches in that direction are still limited. Here, we present a non-parametric Bayesian formalism that performs such a deconvolution efficiently on multidimensional measurements, providing unbiased estimates of the resulting confidence intervals. We use this approach to study the expression of the mesodermal transcription factor Brachyury in mouse embryonic stem cells undergoing differentiation.

INTRODUCTION

The inherent stochasticity of biological processes leads to substantial heterogeneity even among genetically identical cells in the same environment. The degree to which this heterogeneity affects, or even dictates, cellular decision-making in most situations is still an open question. This issue is of paramount importance in processes such as mammalian development, where hundreds (if not thousands) of distinct cell types (cell states) emerge from a small number of identical undifferentiated cells. Identifying the molecular mechanisms underlying these cell-fate decision programs and their interplay with cellular heterogeneity requires a rigorous quantification of cellular states across large numbers of cells.

Flow cytometry enables monitoring the distributions of abundances and activities of selected proteins for thousands of cells at a time, using fluorescently labeled markers. Cells, however, have a non-negligible amount of autofluorescence in the emission spectrum of most fluorescent probes (500–700 nm). In what follows, we indistinctly refer to this autofluorescence as noise, since for our purposes it corresponds to an undesired random signal on top of our signal of interest. This background noise must be subtracted from the total signal emitted by fluorescently labeled cells, in order to adequately relate the signal distribution provided by the cytometer to the mechanisms regulating the expression and/or activity of the protein of interest. Several standard methods exist for addressing this issue on a cell-by-cell basis. One can use, for instance, additional fluorescence channels outside the emission spectrum of the fluorophores serving as regressor variables. Other experimental setups have been developed to deal with autofluorescence, such as introducing a second laser system that provides independent measurements of both the total and autofluorescence signals in the same sample, and even combining in silico simulations with experimental measurements of quantum-dot markers to track cellular proliferation. Beyond issues of cost or accessibility, the effectiveness of such solutions is limited, because there is no guarantee that the autofluorescence from another channel, or from another excitation source, is a good proxy for the autofluorescence in our channel of interest.

Commonly, rather than dedicating measurement resources to assess autofluorescence, control measurements of unlabeled cells are used to set a baseline of the signal coming from the naturally present autofluorescent components in the cell. This procedure, however, does not lead to a quantitative determination of the distribution of the signal coming exclusively from the fluorescent probe. Such quantitative assessment would require deconvolving the fluorescence distribution obtained in labeled cells from the one produced by unlabeled cells. Here, we propose a non-parametric Bayesian approach to this deconvolution problem, applicable to multichannel measurements. The method is robust and efficient, requiring cell numbers not larger than those typically considered in standard flow cytometry runs, and gives natural confidence intervals of the target distributions, which makes it attractive for a variety of applications.

There is an extensive statistics literature addressing the additive deconvolution problem (We do not discuss in what follows other deconvolution problems that are not relevant to our situation, such as those related with composition of distributions, which usually require complete knowledge of the noise distribution). A common set of deconvolution methods are kernel-based approaches, such as those relying...
on Fourier transforms,\textsuperscript{17–23} which use the fact that in Fourier space, a deconvolution is simply the product of two functions. Two problems arise from such methods that limit their applicability in practical cases. First, Fourier transforms (and other methods that use orthogonal local bases, such as wavelets\textsuperscript{24}) can have negative values in part of their domain, which is incompatible with the positive-valued requirement that probability densities must obey. Consequently, kernel-based methods lead to deconvolved pseudo-distributions with artificial features, which are hardly interpretable for practical applications. Second, these methods usually lead to point estimates, and therefore do not provide native confidence intervals (i.e., without applying additional statistical approximations) that allow us to assess the quality of the inferred target distribution. Having a confidence interval around the deconvolved distribution is crucial for practical applications. In general, depending on the levels of noise and the size of the dataset, the final convolved distribution could be equally well described by many different target distributions with quantitative and qualitative dissimilarities. A confidence interval introduces an error estimate representing how different are the various compatible solutions given the data and thus how much one should trust the inferred target distribution in each case.

A second class of deconvolution approaches are likelihood-based methods,\textsuperscript{25} which estimate the unknown target distribution using maximum likelihood approaches. As in the case of kernel-based methods, these approaches provide us with point estimates, and usually assume exact knowledge of the noise distribution.\textsuperscript{16} Additionally, these methods commonly use kernel density estimation to represent the data,\textsuperscript{10} which is rather limited when dealing with multidimensional datasets. Finally, a third class of methods involves Bayesian inference,\textsuperscript{26–28} which does not require complete knowledge of the noise distribution and naturally provides confidence intervals of the estimates obtained. Bayesian approaches have also been applied, in combination with mixture models (described in the following section), to cluster data in flow cytometry analyses.\textsuperscript{27} However, those methods have only been applied so far to repeated measurements of the same individual entities that are being monitored (in our case, cells), which is not a realistic possibility in standard flow cytometry. Our non-parametric Bayesian approach does not require repeated measurements and retains all the aforementioned advantages of Bayesian methods. We have implemented the procedure in a Julia package available in GitHub (https://github.com/dsb-lab/scBayesDeconv.jl).

The work is structured as follows. First, we introduce the mathematical description of the additive convolution problem in the context of flow cytometry data, and explain our proposed deconvolution method based on non-parametric Bayesian models. Second, we validate our method using synthetic datasets with known target distributions, and compare its results to other existing methods. Third, we further test our method in real flow cytometry data of mouse embryonic stem cells undergoing differentiation. We test the dataset in two conditions. Initially, we treat our cells with a low concentration of a fluorescent dye (which masks the real flow cytometry signal and acts as an external noise), and validate our method by deconvolving the noise coming from the dye and comparing it to the control case where the dye was not added. This offers evidence of the robustness of the method in real applications while having a ground truth. Additionally, we perform the deconvolution using several channels from the dataset, to show the effectiveness of the method in multichannel measurements, and comment on the complementarity of our approach to other processing steps common in analysis pipelines of flow cytometry data. We conclude by discussing the limitations of the method and possible ways to improve it in future work.

RESULTS

Theoretical definition of the problem

Consider a population of cells containing fluorescent markers that label the abundances or activities (e.g., phosphorylation states) of certain proteins of interest. Flow cytometry measurements provide us with the distribution $p_C(C)$ of total fluorescence signal $C$ emitted by each individual cell in the population, as measured by the flow cytometer detectors. These signals have two components: the fluorescence $T$ emitted exclusively by the target fluorophores that report on the proteins of interest, and the autofluorescence $\xi$ emitted by cellular components other than our fluorescent labels:

$$C = T + \xi$$

(Equation 1)

If these two components are independent of one another, the distribution $p_C(C)$ of the measured signal takes the form of a convolution of the distributions of $T$ and $\xi$:

$$\left(p_T * p_\xi\right)(C) = \int_0^\infty p_T(C - \xi) p_\xi(\xi) \, d\xi$$

(Equation 2)

Similarly to $p_C$, the distribution $p_T$ of the autofluorescence $\xi$ can be measured in the flow cytometer by using unlabeled cells that are otherwise identical to the labeled ones. On the other hand, the probability distribution of $T$, $p_T$, cannot be measured directly. Our goal is to extract (deconvolve) the distribution $p_T$ from the measured distributions $p_C$ and $p_T$, considering that we only have a finite set of samples (cells) of $C$ and $\xi$.

In what follows, we first introduce the way in which we describe the distributions involved in the problem. Next, we define the posterior distribution given by the model and the data, and finally we discuss the methods used to explore the parameter space for the deconvolution problem.

Mixture model

The vast majority of real datasets (including those generated in flow cytometry) result from complex combinations of variables that cannot be explained in general with simple distributions. In order to adapt flexibly to such conditions, we use mixtures of probability distributions
as our basis set. Any function can be arbitrarily well approximated using an adequate choice of basis functions, provided enough components are included in the mixture. We can describe both the target and the noise distributions by independent mixtures of $K$ components:

$$p(x|\varphi) = \sum_{\eta=1}^{K} \omega_{\eta} B(x|\varphi_{\eta}) \quad \text{(Equation 3)}$$

where $x$ is multivariate in the case of multichannel measurements. $\omega_{\eta}$ denotes the weight of each base $B(x|\varphi_{\eta})$ in the mixture, and $\varphi_{\eta}$ represents its parameters (the means and covariance matrices in the case of multivariate normal distributions) (Throughout the paper we use greek subindices, in particular $\eta$ and $\lambda$, when summing over basis functions, and roman subindices when summing over samples cells). The sets of $\omega_{\eta}$ and $\varphi_{\eta}$ are in turn represented by the vector $\varphi$ in what follows. Using these basis functions, the distribution of the observed signal $C$ can be described as a superposition of all the convolutions between them:

$$p_{c}(c|\varphi_{T}, \varphi_{T}) = \sum_{\eta=1}^{K_{T}} \sum_{\xi=1}^{K_{T}} \omega_{\xi} \omega_{\eta} \{B \ast B\}(c|\varphi_{\xi}^{T}, \varphi_{\eta}^{T}) \quad \text{(Equation 4)}$$

where $(B \ast B)(c|\varphi_{\xi}^{T}, \varphi_{\eta}^{T})$ represents the convolution of two basis distributions with parameters $\varphi_{\xi}^{T}$ and $\varphi_{\eta}^{T}$, respectively, and $K_{T}$ and $K_{T}$ denote the number of bases used in each of the two mixtures.

The choice of basis functions to describe our data are crucial. We propose to use multivariate normal distributions to describe the data. Normal distributions with unknown mean and covariance have been shown to be flexible enough to represent datasets with high quality, without requiring many more components than non-normal, skewed basis functions such as gamma distributions. Normal basis functions have the additional advantage that their convolution turns out to be another normal distribution with modified parameters.\(^{30}\) This last property is especially interesting for deriving exact analytical results for the Bayesian sampling process.

**Posterior distribution and likelihood**

Our goal is to extract, starting from samples of the distributions of total signal $C$ and noise $\xi$, the parameters of the distribution of the target signal $T$. Defining the problem in terms of probability distributions leads very naturally to work with Bayesian methods. According to Bayes’ rule, the posterior distribution that represents the probability of the parameters given the data are

$$p(\varphi_{T}, \varphi_{T} | c, \xi) \propto \left[ p(\xi|\varphi_{T}, \varphi_{T}) p(\xi|\varphi_{T}) \right] \left[ p(\varphi_{T}) p(\varphi_{T}) \right], \quad \text{(Equation 5)}$$

where $c = \{c_{i} : i = 1, \ldots, N_{c}\}$ and $\xi = \{\xi_{i} : i = 1, \ldots, N_{\xi}\}$ are the sets of observed samples of the total signal and the noise, respectively, with $N_{c}$ and $N_{\xi}$ representing the number of samples in each case.

The first bracket on the right-hand side of Equation 5 is the likelihood function, which corresponds to the probability of the data given the parameters. Under the assumption of independent and identically distributed (iid) observations, this function is given by

$$L(\varphi_{T}, \varphi_{T}) = \prod_{i=1}^{N_{c}} p_{c}(c_{i}|\varphi_{T}, \varphi_{T}) \prod_{j=1}^{N_{\xi}} p_{\xi}(\xi_{j}|\varphi_{T}) \quad \text{(Equation 6)}$$

As can be seen, the information about the noise parameters is contained in both datasets, while the information about the target parameters only appears in the convolved data.

As shown in Equation 5 previously, the posterior distribution can be estimated by multiplying the likelihood by the prior distribution of parameter values. Since the datasets required for a good deconvolution are large, the impact of the prior distribution should be negligible. Therefore, the posterior landscape is effectively described by Equation 6, and the prior distribution is only present in our approach for formal and computational reasons. In any case, since prior information exists on the parameters that describe the noise and target mixture components, we impose vague priors over the plausible set of parameters (see STAR Methods “Hyperparameter selection”).

**Sampling**

The posterior distribution is a complex multimodal (multi-peaked) object containing all the configurations of the target and noise distributions that are consistent with the observed data. A correct and efficient exploration of this distribution is essential to set bounds on the candidate models that explain the data. The Bayesian model, as stated in Equation 5, has no direct closed form. Sampling from this model as stated would require the introduction of complex sampling algorithms such as Markov Chain Monte Carlo (MCMC)\(^{30}\) and nested sampling.\(^{31}\) These sampling processes are computationally very expensive and scale linearly with sample size, which in the case of flow cytometry datasets is of the orders of tens to hundreds of thousands samples per dataset.

In order to overcome these shortcomings, we propose two approximations to the posterior distribution. The first one decouples the inference of the noise and target distributions parameters as two independent problems (see STAR Methods “Approximated decomposition of the posterior into two separate problems”). The parameters of the noise distribution can be fitted using well-known mixtures of normal distributions, with the target being described by a convolved mixture model. The second approximation simplifies the sampling of the convolved mixture.
model to enable sampling using standard Gibbs sampling algorithms (see STAR Methods “Gibbs sampling algorithms”). The combination of both approximations allows the posterior to be sampled using Gibbs sampling, which is an efficient MCMC method of sampling from the posterior.

In addition to these approximations, we offer two alternative Bayesian mixtures, namely finite or infinite normal mixture models. The latter approach has the added advantage over the former of not requiring fixing the number of basis functions to describe the data. However, for most practical purposes, finite normal mixture models are able to fit the data with high accuracy and in a more efficient manner regarding how the corresponding algorithms work.

**Analysis pipeline**

Given the concepts and tools described previously, the analysis of the data are performed as shown in Figure 1. First, the distributions of the target and autofluorescence signals are assumed to be described by mixtures (top left panel in Figure 1), as defined by Equations 3 and 4. The data measured experimentally correspond to the autofluorescence signal (noise) and the total signal of the labeled cells (which includes the autofluorescence), as shown in the top right panel of Figure 1. The mixture assumption and the observed data samples allow us to construct the posterior distribution (Equation 5 and middle panel in Figure 1) from the likelihood function defined by Equation 6 and the prior distributions discussed in the STAR Methods “Hyperparameter selection”. The posterior distribution is a function of the model parameters (weights of the mixtures and parameters of the basis functions). We next explore (sample) the posterior distribution in parameter space using the Gibbs sampling approximation (bottom left panels in Figure 1). This algorithm provides us with a representative sampling of the parameters of the target distribution, which allows us to compute an average of this distribution and its confidence interval (bottom right panel in the figure).
Figure 2. Deconvolution instance for a target bimodal distribution (light blue in B and C) corrupted by a normally distributed noise with a signal-to-noise ratio $SNR = 2$ (magenta distribution in the inset of A), applying the Bayesian Deconvolution method (B) and the FFT method (C).

(A and B) The total distribution from which the noise is deconvolved is shown in green in panel (A). The red lines in panel (B) depict samples of the Bayesian fitting process. In this case, the noise distribution is a single normal function with mean $\mu_1 = 0$ and standard deviation $\sigma_1 = 0.5$, and the target is a mixture of two normal functions with means $\mu_2^1 = -0.43$ and $\mu_2^2 = 1.67$, standard deviations $\sigma_2^1 = 0.6$, and weights $\omega_2^1 = 0.8$ and $\omega_2^2 = 0.2$.

Application to synthetic data

First, we benchmark the ability of our method to recover the target distribution by using collections of synthetic datasets. In those datasets, the target and noise distributions are known, so we can compare the result of the deconvolution against a ground truth. To do this, we applied our method to the synthetic data described in the STAR Methods “Synthetic data”. To prove the robustness of the implementation, we ran the test using four components for both the noise and target distributions. We also avoided checking the full convergence of the algorithm manually. We ran the algorithm in these suboptimal conditions to avoid having to fine-tune the specific parameters, which could lead to positive bias favoring our method in comparison with FFT-based approaches. We contrasted the results of our method with those of a specific FFT approach that does not require knowledge of the autofluorescence distribution. Figure 2 shows a typical instance of the deconvolution performance of the two methods.

Panel (A) shows in green the total (convolved) signal mimicking the output of a flow cytometry experiment. In this synthetic case, the signal is obtained by forward convolving a target distribution with the characteristics given earlier, shown in light blue in panels (B) and (C), with a noise (auto-fluorescence) distribution, shown in magenta in the inset of panel (A). The goal in this case is to recover (deconvolve) the ground-truth target distribution from the total and noise distributions. Even in the case of the relatively simple target distribution considered here (a bimodal normal mixture), our Bayesian deconvolution method (Figure 2B) performs qualitatively better than the FFT-based method (Figure 2C). Additionally, the FFT method leads to oscillatory components in the deconvolution, and correspondingly to artifactual negative values in the probability distribution. Additionally, the Bayesian deconvolution method naturally provides a confidence interval, shown by the different samples (red lines) in panel (B) of Figure 2.

For benchmarking purposes and given the relative simplicity of the bimodal normal mixture considered in Figure 2 previously, we next turned a wide variety of synthetic distributions (shown in Figure S1), for which we compared systematically our Bayesian deconvolution method with the FFT-based approach discussed above. To quantify the similarity between the deconvolved and the ground-truth target distribution we used the mean integrated overlap (MIO), as defined in STAR Methods “Efficiency assessment”. According to its definition (Equation 108 in STAR Methods), a MIO value of 1 corresponds to a perfect overlap, while the measure is 0 when the two distributions do not overlap at all. We preferred this measure to the more commonly used mean integrated squared error (MISE) for theoretical reasons, since the MIO measure is easier to interpret, as it corresponds directly to the absolute overlap of two probability distributions. We also avoid other common measures of distribution dissimilarity, such as the Kolmogorov-Smirnov test, since such methods would underestimate the ability of FFT-based methods to converge to the ground-truth deconvolution, given that they lead to artifacts in the resulting distributions, as shown previously.

Figure 3 compares the MIO of the two methods for each of the 9 test cases shown in Figure S1, for which we compared systematically our Bayesian deconvolution method and the FFT-based approach discussed above. To quantify the similarity between the deconvolved and the ground-truth target distribution we used the mean integrated overlap (MIO), as defined in STAR Methods “Efficiency assessment”. According to its definition (Equation 108 in STAR Methods), a MIO value of 1 corresponds to a perfect overlap, while the measure is 0 when the two distributions do not overlap at all. We preferred this measure to the more commonly used mean integrated squared error (MISE) for theoretical reasons, since the MIO measure is easier to interpret, as it corresponds directly to the absolute overlap of two probability distributions. We also avoid other common measures of distribution dissimilarity, such as the Kolmogorov-Smirnov test, since such methods would underestimate the ability of FFT-based methods to converge to the ground-truth deconvolution, given that they lead to artifacts in the resulting distributions, as shown previously.

Additionally, it is worth noting that while the single-cell Bayesian deconvolution method is able to reproduce almost perfectly the target distribution (MIO close to 1) for large enough dataset sizes, the Fourier based method always saturates to a suboptimal level even as the number of samples increases, even for low noise levels (large SNR, crosses in Figure 3). Qualitatively, this is due to the oscillatory features in the distribution produced by the noise during the deconvolution with the FFT method, as a consequence of the oscillatory nature of the Fourier basis functions. In general, it has been proved that in FFT methods the convergence to the actual distribution grows sublinearly with the sample size, with scaling dependencies that make the method unfeasible for practical purposes. Since Bayesian deconvolution makes use of local basis functions, more parsimonious solutions can be obtained, preventing the degradation of the target distribution due to the noise.

In addition to the direct comparison performed between FFT and Bayesian deconvolutions, we compared the results of the deconvolution against a model that simply fits the convolved data ignoring the convolution problem. A condition to impose on any practical deconvolution
algorithm is that the result of the deconvolution is not worse than not doing the decomposition at all. The results of this comparison can be observed in Figure S2. The results show that, while the Fourier based method generally tends to make worse predictions than the null model (fitting without deconvolution), the Bayesian approach consistently generates results virtually as good as when the noise is negligible (SNR = 10), and improves systematically the distribution in the high-noise case (SNR = 1) over the null model. In conclusion, the Bayesian method that we propose has combined benefits, both qualitatively and quantitatively, over other basis methods such as FFT and it consistently improves the result (or at least it does not degrade the quality of the distribution).

An experimental dataset with external noise

Next, we applied our method to an experimental dataset in which the noise distribution is unknown. To generate a ground truth, we used cells with the Brachyury reporter T/BraGFP in two media conditions (N2B27 supplemented with 3 μM CHIR99 and DMSO as control). At day 3, for each condition, one of the replicas was treated with 20 nM Green CMFDA dye and incubated for 3 min prior to flow cytometry, and the second one was incubated for 3 min with N2B27 (see STAR Methods). Given that the dye incorporates in the cell cytoplasm and its emission spectrum is similar to the one of GFP, the dye acts as a noise source to the GFP signal coming from the Brachyury reporter. Consequently, we have the following four conditions, with their potential outcomes in terms of the signal measured in the FITC-A channel.

- **(c1) Control**: Brachyury expression is minimal, and the signal comes mainly from the intrinsic autofluorescence of the cells.
- **(c2) Control + CMFDA**: Brachyury expression is minimal, and the signal comes mainly from the CMFDA dye, plus intrinsic autofluorescence of the cells.
- **(c3) CHIR99**: Brachyury expression is upregulated, and thus the signal contains both the T/BraGFP reporter and the intrinsic autofluorescence of the cells.
- **(c4) CHIR99 + CMFDA**: Brachyury expression is upregulated, and thus the signal contains both the T/BraGFP reporter, the signal from the CMFDA dye and the intrinsic autofluorescence of the cells.

The four signal distributions corresponding to these four conditions are shown in Figure S3. We note that the distribution of the signal coming only from the dye cannot be measured independently, given the intrinsic autofluorescence of the cells. Moreover, the dye might be absorbed differently depending on the state of the cells (CHIR99 treatment), and the low concentration of the dye (≈ nM) might lead to significant cell-to-cell variability on the absorbance of the dye. These features make it a suitable dataset to test the robustness of our method.

We first used our method to deconvolve the signal of the dye from the total signal observed in the experimental conditions (c2), using condition (c1) to define the “noise” distribution. The results of the deconvolution are shown in Figure S4, which presents the dye distribution deconvolved from experimental conditions (c1) and (c2). Once the dye distribution was obtained from the deconvolution of conditions (c1) and (c2), we tested the consistency of our approach by considering the dye signal as the noise in condition (c4). Deconvolving
the dye distribution from the one measured in (c4) should lead to the distribution obtained in condition (c3), which we can measure experimentally and thus serves as the ground truth in this case. The result can be seen in Figure 4. The measured signal (top panel) is mildly skewed but distinguishable from a normal distribution (its skewness coefficient is 11.6, while that of a log-normal distribution with the same mean and variance is 5.3). Additionally, we note that here we use a deconvolved dataset for a second deconvolution. Even in these conditions, the inferred distribution resulting from our method is in good agreement with the experimentally measured distribution in condition (c3) (MIO = 0.81 ± 0.04) and improves over the null model that ignores the noise (MIO = 0.46 ± 0.01). Overall, this experiment shows that our method is robust even when our conditions (target and noise independence) are not fully met, and over other fluctuations present in real datasets.

Multidimensional distributions

Finally, we applied our method to a multichannel flow cytometry dataset. Specifically, we measured the expression of the mesodermal gene Brachyury through a GFP reporter (T/BraGFP) in mouse embryonic stem cells (see STAR Methods). In this dataset, two populations are expected to appear: a Brachyury-negative population of pluripotent cells, and a positive population capturing the cells exiting from pluripotency toward mesodermal fates.

The Brachyury signal of the GFP reporter was our target signal, and it was acquired through the FITC-A channel. Additionally, we use in what follows five more channels produced by the cytometer, including two light-scattering channels and three extra fluorescence channels (see Figure S5 for scatterplots representing the data in 2-dimensional projections of the 6-dimensional measurement space). These additional channels are barely uncorrelated with the Brachyury signal of interest, and although none of them provide any useful biological signal in our case, we use them to represent potential additional channels to which the deconvolution method described previously can be applied in multidimensional space.

Figure 5 shows the flow cytometry signal from the T/BraGFP cells (represented as light-green circles) in the plane formed by the Brachyury channel FITC-A and one of the other five channels that were used in the deconvolution. The signal distribution was fitted to a mixture of four normal basis functions, only two of which (2 and 3, see inset in the figure) are enough to reproduce the data. The extent of the two subsets (which we associate with the two populations described previously) is represented by blue and red ellipses that mark one standard deviation away from the means (blue and red solid circles), as determined by the respective covariant matrices in two dimensions (having traced out the other four dimensions).

Brachyury is expressed in the two populations. In particular, there seems to be a non-zero basal expression of the gene in the pluripotent population (the blue ellipse is fully above the horizontal zero axis). These cells, however, are not expected to express Brachyury, as confirmed by the absence of mRNA in the pluripotent states.33 Reassuringly, when the data are deconvolved through our Bayesian method, the result shows that the pluripotent population peaks actually around zero (small blue ellipse encompassing the light-blue circles in Figure 5). Similar cases of apparent basal signals in negative cells due to autofluorescence have been reported, such as the false positive detection of MHCII in microglia,34 and the false discovery of Foxp3-expressing cell types that are not part of the T cell lineage.35
We can compare in this multidimensional case the results of our method with the classical approach of removing the mean value of the autofluorescence from the signal distribution. The autofluorescence data points are shown in orange in Figure 5, while gray circles are the result of removing the mean autofluorescence from the total signal distribution. Whereas this method also corrects the lower peak to zero level expression, the width of the resulting mean distribution is much larger. Other approaches that use additional channels such as regressors are expected to give results similar to this simpler method, since there are no appropriate regressor channels that can be used to remove the autofluorescence (see STAR Methods “Generalization to multichannel measurements”). Thus, the autofluorescence substraction method and other alternative approaches do not quantify correctly the variance of data. Our method outperforms those other approaches in this regard by using the fact that the variability of FSC-A is a main contribution to the autofluorescence signal, and therefore removing it produces narrower distributions that capture the real variability of the biological phenomenon.

In summary, the Bayesian deconvolution method proposed here is applicable in a straightforward manner to multichannel datasets, and its results are consistent with the expected biological observations. Our method is thus able to correct for artifacts arising from the presence of background signals in multiple detection channels.

**DISCUSSION**

In this paper, we propose a Bayesian approach to obtain flow cytometry distributions of protein abundance or activity convolved with a known source of noise. The method, which relies on non-parametric Bayesian techniques, is freely available as a Julia package (https://github.com/dsb-lab/scBayesDeconv.jl) and can be used in a straightforward manner in a purely computational way, without the need of dedicated measurement channels or additional laser sources. It only requires measuring the fluorescently labeled and unlabeled cells and, unlike previously proposed deconvolution methods, it provides well-defined and robust probability distributions, described by mixtures of basis functions.

We measure the quality of the results obtained with our method by comparing the deconvolved distributions with known (ground-truth) target distributions using synthetic data and ad hoc experiments with mouse embryonic stem cells. We argue that the use of local basis distributions to describe all the distributions involved in the problem (both measured and unknown), and the corresponding use of a relatively small number of degrees of freedom, leads to an efficient inference. Finally, the Bayesian nature of the method gives rise to a set of candidate target probability distributions. This reflects the natural indeterminacy present in any process corrupted by noise. The ability to express indeterminacy in the solutions is crucial for any real application of a deconvolution algorithm. We further show that the method is robust and consistent even under strong noise in real datasets.

The approach that we propose here is applicable to flow cytometry datasets with multiple channels, which is the current outcome of modern flow cytometers. Furthermore, the method is consistent with already well-established methods for flow cytometry processing, such as the...
calculation of spillover matrices and the reduction of autofluorescence using additional channels for autofluorescence regression, as discussed in STAR Methods “Generalization to multichannel measurements”. In this sense, our deconvolution approach could be applied to calculate corrected one-fluorophore samples for the calculation of the spillover coefficients, or it could be used after calculating the spillover uncoupling with another method, in order to remove the remaining sources of noise in the dataset.

**Limitations of the study**

A potential limitation of the algorithm is the computing time requirements. For $K_n$, mixture components of the noise and $K_t$, components of the target, a single evaluation of the likelihood function scales as $O(K_n \times K_t \times N)$, where $N$ is the number of measures in the Bayesian model. The current finite mixture implementation is capable of dealing with datasets of 100000 cells with $K_n = 4$ and $K_t = 6$ in about 10 min in a desktop with an i7 processor. This efficiency will be sufficient for most part of the analysis, but can be a drawback for the analysis of very large datasets when compared with non-Bayesian methods. The computational time required by the infinite mixture model depends strongly on the parameter $\alpha$, which governs the probability of generating new basis functions, but in general this model is more than an order of magnitude slower than its finite counterpart for the same degree of complexity. Handling large datasets is a well-established problem in Bayesian statistics, and some lines of work have explored solutions to reduce this computational cost, which might be worth exploring in the context of the single-cell Bayesian deconvolution method proposed here. 37–39

A second potential improvement of the method presented here would be to address the degeneracy under label exchange characteristic of mixture models. Such degeneracy may confound the convergence of the samples to a stationary distribution and the use of the samples for the discovery of clusters, as the same cluster may be assigned to different basis in different samples. A potential solution to this problem would be the implementation of label reassignment methods. 40

Finally, it would be interesting to generalize this method to other basis functions. While normal distributions have been used in this work because of their well-behaved properties and the existence of analytical results (without losing quality when representing the data), they may not be the most efficient basis functions to describe the skewed distributions typical of flow cytometry datasets. A mixture of gamma or log-normal distributions could adapt to the target datasets using less components than normal distributions. In order to generalize to such basis distributions and keep the efficiency of the method, approximate analytical expressions of the convolved distributions should be derived.

**STAR+ METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS**
- **METHOD DETAILS**
  - Biological statement of the problem
  - Mathematical statement of the problem
  - Relevant probability distributions
  - Hyperparameter selection
  - Gibbs sampling algorithms
  - Efficiency assessment
  - Synthetic data
  - Software

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107941.

**ACKNOWLEDGMENTS**

We thank the reviewers of the paper for their insightful comments, and the CRG Tissue Engineering Unit and the UPF/CRG Flow Cytometry Unit for continuous support. This work was supported by the Spanish Ministry of Science and Innovation and FEDER, under projects PIS2017-92551-EXP and PID2021-127311NB-I00, by the “Maria de Maeztu” Program for Units of Excellence in R&D (grant CEX2018-000792-M), and by the Generalitat de Catalunya (ICREA Academia program). GTC is supported by an FPU doctoral fellowship from the Spanish Ministry of Education and Universities (reference FPU18/05091). D.O. acknowledges funding from Juan de la Cierva Incorporación with Project no. IJC2018-035298-I, from the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI).
REFERENCES

1. Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S. (2002). Stochastic gene expression in a single cell. Science 297, 1183–1186. https://doi.org/10.1126/science.1070919.

2. Suel, G.M., Kulkarni, R.P., Dworkin, J., Garcia-Ojalvo, J., and Elowitz, M.B. (2007). Tunability and noise dependence in differentiation dynamics. Science 315, 1716–1719. https://doi.org/10.1126/science.1137455.

3. Chang, H.H., Hemberg, M., Barahona, M., Ingber, D.E., and Huang, S. (2008). Transcription-wide noise controls lineage choice in mammalian progenitor cells. Nature 453, 544–547. https://doi.org/10.1038/nature06965.

4. Osumi-Sutherland, D., Xu, C., Keays, M., Levine, A.P., Kharchenko, P.V., Regen, A., Lein, E., and Teichmann, S.A. (2021). Cell-type ontologies of the human cell atlas. Nat. Cell Biol. 23, 1129–1135. https://doi.org/10.1038/s41556-021-00767-7.

5. Nowotschin, S., Setty, M., Kuo, Y.-Y., Liu, V., Garg, V., Sharma, R., Simon, C.S., Saiz, N., Gardiner, R., Boutet, S.C., et al. (2019). The emergent landscape of the mouse gut endoderm at single-cell resolution. Nature 569, 361–367. https://doi.org/10.1038/s41586-019-1127-1.

6. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science 367, eaaw3381. https://doi.org/10.1126/science.aaw3381.

7. Mittnenzweig, M., Mayshar, Y., Cheng, S., Ben-Yair, R., Hadad, R., Raic, Y., Chomsky, E., Reines, N., Uzoniy, A., Lumerman, L., et al. (2021). A single-embryo, single-cell time-resolved model for mouse gastrulation. Cell 184, 2825–2842.e22. https://doi.org/10.1016/j.cell.2021.04.004.

8. Garcia-Ojalvo, J., and Martinez Arias, A. (2012). Towards a statistical mechanics of cell fate decisions. Curr. Opin. Genet. Dev. 22, 619–626. https://doi.org/10.1016/j.gde.2012.10.004.

9. Ribe, P. and Garcia-Ojalvo, J. (2013). Modeling gene expression in time and space. Annu. Rev. Biophys. 42, 605–627. https://doi.org/10.1146/annurev-biophys-080112-130335.

10. Corsetti, J.P., Sotirchos, S.V., Cox, C., Cowles, J.W., Leary, J.F., and Blumberg, N. (1988). Correction of cellular autofluorescence in flow cytometry by mathematical modeling of cellular fluorescence. Cytometry 9, 539–547. https://doi.org/10.1002/cyto.990090606.

11. Roederer, M., and Murphy, R.F. (1986). Cell-by-cell autofluorescence correction for low signal-to-noise systems: Application to epidermal growth factor endocytosis by 3T3 fibroblasts. Cytometry 7, 538–565. https://doi.org/10.1002/cyto.990070610.

12. Alberti, S., Parks, D.R., and Herzenberg, L.A. (1987). A single laser method for subtraction of cell autofluorescence in flow cytometry. Cytometry 8, 114–119. https://doi.org/10.1002/cyto.990080203.

13. Steinkamp, J.A., and Stewart, C.C. (1986). A single laser method for subtraction of cell autofluorescence in flow cytometry. Cytometry 7, 925–932. https://doi.org/10.1002/cyto.a.20936.

14. Brown, M.R., Summers, H.D., Rees, P., Chappell, S.C., Silvestre, O.F., Khan, I.A., Smith, P.J., and Errington, R.J. (2010). Time-lapse series analysis of quantum dot encoded cells by deconvolution of the autofluorescence signal. Cytometry A 77, 925–932. https://doi.org/10.1002/cyto.a.20936.

15. Efron, B. (2016). Empirical Bayes deconvolution estimates. Biometrika 103, 1–20. https://doi.org/10.1093/biomet/asv068.

16. Madrid-Padilla, O.-H., Polson, N.G., and Scott, J. (2018). A deconvolution path for mixture estimation. Electron. J. Statist. 12, 1717–1751. https://doi.org/10.1214/18-EJS1430.

17. Stefanski, L.A., and Carroll, R.J. (1990). Deconvolving kernel density estimators. Statistics 21, 169–184. https://doi.org/10.1080/02331889008802238.

18. Liu, M.C., and Taylor, R.Y. (1990). Simulations and computations of nonparametric density estimates for the deconvolution problem. J. Stat. Comput. Simulat. 35, 145–167. https://doi.org/10.1080/00949699008811241.

19. Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. Ann. Stat. 19, 1257–1272. https://doi.org/10.1214/aos/1176342848.

20. Diggle, P.J., and Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. J. Roy. Stat. Soc. B 55, 523–531. https://doi.org/10.1111/j.2517-6161.1993.tb01920.x.

21. Neumann, M.H., and Hössjer, O. (1997). On the effect of estimating the error density in nonparametric deconvolution. J. Nonparametric Statistics 7, 307–330. https://doi.org/10.1080/10485259708832708.

22. Comte, F., and Lacour, C. (2011). Data-driven density estimation in the presence of additive noise with unknown distribution. J. Roy. Stat. Soc. B 73, 601–627. https://doi.org/10.1111/j.1467-9868.2011.00775.x.

23. Delagie, A., and Hall, P. (2014). Parametrically assisted nonparametric estimation of a density in the deconvolution problem. J. Am. Stat. Assoc. 109, 717–729. https://doi.org/10.1080/01621459.2013.857611.

24. Fan, J., and Koo, J.-Y. (2000). Wavelet deconvolution. IEEE Trans. Inf. Theor. 48, 734–747. https://doi.org/10.1109/1.86021.

25. Lee, M., Hall, P., Shen, H., Marron, J.S., Tolle, J., and Burch, C. (2013). Deconvolution estimation of mixture distributions with boundaries. Electron. J. Stat. 7, 323–341. https://doi.org/10.1214/13-EJS774.

26. Staudenmayer, J., Ruppert, D., and Buonaccorsi, J.P. (2008). Density estimation in the presence of heteroscedastic measurement error. J. Am. Stat. Assoc. 103, 726–736. https://doi.org/10.1198/016214508000000328.

27. Sarkar, A, Mallick, B.K., Staudenmayer, J., Pati, D., and Carroll, R.J. (2014). Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors. J. Comput. Graph Stat. 23, 1101–1125. https://doi.org/10.1080/10618600.2014.999231.

28. Sarkar, A, Pati, D., Chakraborty, A, Mallick, B.K., and Carroll, R.J. (2018). Bayesian semiparametric multivariate density deconvolution. J. Am. Stat. Assoc. 113, 401–416. https://doi.org/10.1080/01621459.2016.1260467.

29. Amalfitano, S., Fazi, S., Ejarque, E., Freixa, A., Romani, A.M., and Buttunni, A. (2018). Deconvolution model to resolve cytometric microbial community patterns in flowing waters. Cytometry A 93, 194–200. https://doi.org/10.1002/cyto.a.23304.
30. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). Bayesian Data Analysis, Third Edition (Taylor & Francis).
31. Skilling, J. (2006). Nested sampling for general Bayesian computation. Bayesian Anal. 1, 833–859. https://doi.org/10.1214/06-BA127.
32. Li, Y., Schofield, E., and Gönen, M. (2019). A tutorial on dirichlet process mixture modeling. J. Math. Psychol. 91, 128–144. https://doi.org/10.1016/j.jmp.2019.04.004.
33. Anlas, K., Gritti, N., Oriola, D., Arató, K., Nakaki, F., Lim, J.L., Sharpe, J., and Trivedi, V. (2021). Dynamics of Anteroposterior axis Establishment in a Mammalian Embryo-like System. https://doi.org/10.1101/2021.02.24.432766.
34. Li, Q., and Barres, B.A. (2018). Microglia and macrophages in brain homeostasis and disease. Nat. Rev. Immunol. 18, 225–242. https://doi.org/10.1038/nrni.2017.125.
35. Li, F., Yang, M., Wang, L., Williamson, I., Tian, F., Qin, M., Shah, P.K., and Sharifi, B.G. (2012). Autofluorescence contributes to false-positive intracellular Foxp3 staining in macrophages: A lesson learned from flow cytometry. J. Immunol. Methods 386, 101–107. https://doi.org/10.1016/j.jim.2012.08.014.
36. Roca, C.P., Burton, O.T., Gergelits, V., Prezzemolo, T., Whyte, C.E., Halpert, R., Kreft, K., Collier, J., Botzki, A., Spidlen, J., et al. (2021). AutoSpill is a principled framework that simplifies the analysis of multichromatic flow cytometry data. Nat. Commun. 12, 2890. https://doi.org/10.1038/s41467-021-23126-8.
37. Huang, Z., and Gelman, A. (2005). Sampling for Bayesian Computation with Large Datasets (SSRN-Elsevier). available at Social Science Research Network. https://doi.org/10.2139/ssrn.1010107.
38. Nemeth, C., and Sherlock, C. (2018). Merging MCMC subposteriors through Gaussian-process approximations. Bayesian Anal. 13, 507–530. https://doi.org/10.1214/17-BA1063.
39. Vehtari, A., Gelman, A., Sivula, T., Jylinki, P., Tran, D., Sahai, S., Blomstedt, P., Cunningham, J.P., Schimovich, D., and Robert, C.P. (2020). Expectation propagation as a way of life. A framework for Bayesian inference on partitioned data. J. Mach. Learn. Res. 21, 1–53. https://doi.org/10.48550/arXiv.1412.4869.
40. Cron, A.J., and West, M. (2011). Efficient Classification-Based Relabeling in Mixture Models. Am. Statistician 65, 16–20. https://doi.org/10.1198/tast.2011.10170.
41. Fehling, H.J., Lacaud, G., Kubo, A., Kennedy, M., Robertson, S., Keller, G., and Kouskoff, V. (2003). Tracking mesoderm induction and its specification to the hemangioblast during embryonic stem cell differentiation. Development 130, 4217–4227. https://doi.org/10.1242/dev.00589.
42. Turner, D.A., Girgin, M., Alonso-Crisostomo, L., Trivedi, V., Baillie-Johnson, P., Glodowski, C.R., Hayward, P.C., Collignon, J., Gustavsen, C., Serup, P., et al. (2017). Anteroposterior polarity and elongation in the absence of extra-embryonic tissues and of spatially localised signalling in gastruloids: mammalian embryonic organoids. Development 144, 3894–3906. https://doi.org/10.1242/dev.150391.
STEREO METHODS

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Chemicals, peptides, and recombinant proteins | | |
| CHIR99021 | Sigma | Cat # SML1046 |
| Activin A | Bio-Technne | Cat # 338-AC |
| CellTracker Green CMFDA Dye | Thermofisher | Cat # C7025 |

| Deposited data | | |
| Flow cytometry data | This paper | https://github.com/dsb-lab/scBayesDeconv.jl |
| Synthetic data | This paper | https://github.com/dsb-lab/scBayesDeconv.jl |

| Experimental models: Cell lines | | |
| E14 T/Bra::GFP | Gift from Alfonso Martinez-Arias | N/A |

| Software and algorithms | | |
| BD FACSDiva™ | BD Biosciences | | |
| Python | Python Software Foundation | www.python.org |
| Julia | Julia Computing | www.julialang.org |
| scBayesDeconv: A Julia package for noise deconvolution | This paper | https://github.com/dsb-lab/scBayesDeconv.jl |

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jordi Garcia-Ojalvo (jordi.g.ojalvo@upf.edu).

Materials availability
This study did not generate new unique reagents.

Data and code availability
- Experimental data have been deposited at https://github.com/dsb-lab/scBayesDeconv.jl and are publicly available as of the date of submission.
- All original code has been deposited at https://github.com/dsb-lab/scBayesDeconv.jl and is publicly available as of the date of submission.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

E14 mouse embryonic stem cells containing a knock-in fluorescence reporter for the mesodermal transcription factor Brachyury, T/Bra::GFP were used.31 Cells containing the T/Bra::GFP reporter were cultured in ES-Lif (ESL) medium (KnockOut Dulbecco’s Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 1x Non-essential aminoacids (NEEA), 50 U/mL Pen/Strep, 1x GlutaMax, 1x Sodium Pyruvate, 50 μM 2-Mercaptoethanol and leukemia inhibitory factor (LIF). Cells adhered to tissue culture-treated 25 cm² (T25) plates (Corning, catalog number 353108) coated with 0.1% gelatin (Millipore, catalog number ES-006-B), and were passaged every second day, as previously described.42 Cells were kept at 37°C with 5% CO₂, and were routinely tested and confirmed to be free from mycoplasma.

Flow cytometry experiments were performed as follows: On day 1, cells were trypsinized and seeded into gelatin-coated 6-well plates (Corning, catalog number 352244) to a final density of ~ 10⁵ cells/well in 3 mL ESL media. On day 2, the media was replaced by first washing twice with 3 mL DPBS++ (Phosphate buffered saline containing Mg++ and Ca++), Sigma, catalog number D8662) and then adding NDiff227 media (N2B27) (Takara Bio, catalog number Y40002)32 with the appropriate combination of Brachyury activators: 3 μM CHIR99021 (Sigma, catalog number SML1046) and/or 25 ng/mL Activin A (Bio-Technne, catalog number 338-AC). The same procedure was followed on day 3. For the control conditions, the corresponding volume of dimethyl sulfoxide (DMSO) was added to the medium.
Flow cytometry data was acquired on day 2 or day 3, depending on the type of experiment. For the extrinsic noise experiment, cells were incubated with 1 mL of 20 nM CellTracker Green CMFDA dye (Thermofisher, catalog number C7025) for 3 min prior to trypsinization and flow cytometry. The data was acquired using an LSRIIb flow cytometer, with 2×10^5 cells being analyzed per condition. DAPI labelling was used to discard death cells and debris. Furthermore, cell doublets were discarded from the analysis. The readout of protein expression (peak of the signal) was obtained through the FITC-A channel. Measurements were extracted using the BD FACSDiva™ (BD Biosciences) software and exported in a Python format for subsequent analysis.

**METHOD DETAILS**

**Biological statement of the problem**

Flow cytometry provides us with measurements of a target signal \( T \) in large numbers of single cells. The target signal is emitted by a fluorophore that reports on the abundance (or activity) of a protein of interest within each cell. This signal is affected by autofluorescence (which can be considered a source of noise) produced by elements of the cell other than the fluorophore. Due to the noise, the total signal \( C \) measured by the device is not directly \( T \), but

\[
C = T + \xi
\]

(Equation 7)

**Deconvolving signal from noise**

We are interested in the case in which we cannot measure both the signal \( C \) and the noise \( \xi \) independently in the same cell, and thus \( T \) in that cell cannot be calculated trivially via Equation 7. This limitation is typical of flow cytometry experiments, in which cells can only be measured once. In this case, the only information that can be extracted from the device consists of the distributions of the measured signal, \( p_T \), and of the background noise by itself, \( p_\xi \) (by measuring cells without fluorophore), over large populations of cells (tens of thousands in a typical flow cytometry run). We can assume the samples to be independent and identically distributed (iid).

If \( T \) and \( \xi \) independent of each other, the distributions defined above are related to one another by means of a convolution:

\[
p_T(C) = \int_0^\infty \int_0^\infty p_T(T) p_\xi(\xi) \delta(C - (T + \xi)) dT d\xi = \int_0^\infty p_T(C - \xi) p_\xi(\xi) d\xi \equiv (p_T * p_\xi)(C)
\]

(Equation 8)

In what follows, we describe a method to extract the distribution \( p_T \) of the target variable \( T \) from the observed distributions of \( C \) and \( \xi \) via a deconvolution of Equation 8. The method is applicable to any measurement technique that provides distributions of a signal affected by noise.

**Generalization to multichannel measurements**

Flow cytometry systems have multiple detectors to measure light at different emission frequencies. This allows to target different proteins in the same cell with different fluorophores, measuring their emission at either the peak frequency (classical flow cytometry) or in a set of frequencies that define the emission spectrum (spectral flow cytometry). Independently of the method, the emission of a fluorophore extends over the spectrum, and hence it spills over the different channels. Considering the additive effect of the emission of different fluorophore markers on the measured channels, Equation 7 can be more generally stated as

\[
c_k = \sum_{j=1}^{N_c} t_j S_{jk} + \xi_k, \quad k = 1...N_{ch}
\]

(Equation 9)

where \( c_k \), \( t_j \) and \( \xi_k \) are realizations of the random variables \( C \), \( T \) and \( \xi \) defined in Equation 7 above. The subindex \( j \) runs over the \( N_c \) fluorophores (one per target), and the subindex \( k \) runs over the \( N_{ch} \) channels (so that \( \xi_k \) represents the contribution of the autofluorescence to channel \( k \)). The terms \( S_{jk} \) define the spillover matrix \( S \), that quantifies how the fluorophore signals are spread over all the measuring channels.

The spillover matrix can be estimated from single fluorophore controls by using regression methods. To that end, one can perform control experiments in which only one fluorophore is present, and measure the resulting signal in all the channels:

\[
c_k^{(j)} = t_j^{(j)} S_{jk} + \xi_k
\]

(Equation 10)

where the subindex \( j \) now corresponds to the only fluorophore present in the system, the superindex \( (j) \) indicates that the experiments were performed in the single-fluorophore condition, and \( c_k^{(j)} \) denotes a total signal in channel \( k \) coming from cell \( i \) when only the fluorophore \( j \) is present in the system. This set of equations is underdetermined, as we do not know neither the real target signal \( t_j^{(j)} \) nor the components \( S_{jk} \) of the spillover matrix. However, if we focus for the moment on the channel that corresponds to our single fluorophore \( (k = j) \) and impose \( S_{jj} = 1 \), we can write down that \( c_k^{(j)} = t_j^{(j)} \) (ignoring for now the autofluorescence of that channel). This allows us to establish a set of linear regression problems whose solution enables the estimation of the spillover matrix components for which we have single-fluorophore controls:

\[
c_k^{(j)} = c_k^{(j)} S_{jk} + \xi_k
\]

(Equation 11)
The spillover coefficients $S_{ik}$ can then be estimated using robust regression techniques that remove the noise coming from outliers.36 A similar method can be used to reduce the noise coming from the autofluorescence. To that end, one can define an effective “fluorophore signal” $t_n$ coming from the autofluorescence, with an associated additional channel $n$ not linked to a real fluorophore used in the sample. If the signal in this channel is correlated with the emission of the autofluorescence $\xi$ over the other channels, we can decompose the autofluorescence signal at every channel $k$ into a regression term and a noise term:

$$\xi_k \equiv t_n S_{nk} + \epsilon_k \quad \Rightarrow \quad c_k = \sum_j b_j S_{jk} + t_n S_{nk} + \epsilon_k$$  \hspace{1cm} (Equation 12)

The coefficients $S_{nk}$ indicate how the autofluorescence signal is spread over the other channels. The fluorophore signals and the autofluorescence “signal” can be grouped in a single term:

$$c_k = \sum_{j=1}^{N_p+1} b_j S_{jk} + \epsilon_k$$  \hspace{1cm} (Equation 13)

where now the sum over $j$ also includes the autofluorescence signal. We can then calculate the spillover coefficients with the method above, by using the additional channel $n$ as an additional “fluorophore control” for the autofluorescence signal. We also note that writing the autofluorescence as a regression problem in Equation 12 will lead in general to a reduction of the noise in all channels:

$$\text{Var}(\xi_k) < \text{Var}(\epsilon_k)$$  \hspace{1cm} (Equation 14)

Finally, to obtain the original signal, we can multiply Equation 13 by the inverse of the spillover matrix:

$$C S^{-1} = T + \xi S^{-1} \approx T + \xi''$$  \hspace{1cm} (Equation 15)

We emphasize that this method requires that the autofluorescence signal at the “noise channel” is sufficiently correlated with its contribution at the other channels, which is not necessarily true a priori. Also, Equation 15 shows that the need to deconvolve the noise stands even after the application of the spillover and autofluorescence corrections proposed in the literature. In fact, the deconvolution method that we propose in this article is compatible with existing methods of autofluorescence correction such as the regression method reviewed above, which can be performed before applying our method to the corrected system (Equation 15).

Mathematical statement of the problem

Since we do not know the exact underlying distributions, we model them as potentially infinite mixtures of normal basis functions. The Gaussian mixture models for the target and the noise distributions can be represented as

$$p(t|\omega^T, \{\mu^j\}, \{\Sigma^j\}) = \sum_{i=1}^{K_t} \omega^i N(t|\mu^i, \Sigma^i)$$  \hspace{1cm} (Equation 16a)

$$p(\xi|\omega^T, \{\mu^j\}, \{\Sigma^j\}) = \sum_{i=1}^{K_t} \omega^i N(\xi|\mu^i, \Sigma^i)$$  \hspace{1cm} (Equation 16b)

where $N(x|\mu, \Sigma)$ denotes a normal distribution on the variable vector $x$ (whose components are the different measurement channels), with mean $\mu$, and variance $\Sigma$. $K_t$ and $K_r$ represent the number of bases used to describe each distribution, $\omega = \{\omega_i : i = 1, \ldots, K_r\}$ are their weights, and $\{\mu\} = \{\mu_i : i = 1, \ldots, K_t\}$ and $\{\Sigma\} = \{\Sigma_i : i = 1, \ldots, K_r\}$ are their characteristic parameters. Note that we have removed the subindices from the distributions of the target and the noise, since according to the mixture representation, $p^T$ and $p^f$ depend exclusively on the parameters defined above.

The mixture decomposition defined in Equation 16 allows for a flexible and robust representation of unknown and generic distributions. Moreover, exploiting the fact that the convolution of two Gaussian distributions is Gaussian, we have an analytical expression for the distribution of the total variable $C$:

$$p(c|\varphi^T, \varphi^f) = \sum_{i=1}^{K_t} \sum_{j=1}^{K_r} \omega^i \omega^j N\left(c|\mu^i + \mu^j, \Sigma^i + \Sigma^j\right)$$  \hspace{1cm} (Equation 17)

where $\varphi^T = \{\omega^T, \mu^T, \Sigma^T\}$ and $\varphi^f = \{\omega^f, \mu^f, \Sigma^f\}$ represent all the parameters of the target and noise distributions, from which we can parametrize the distribution of the total measured signal. We note that the combination of normal basis function $N$ in Equation 17 above is invariant under changes in the individual basis functions of $T$ and $F$ provided the sums of means, $\mu^T + \mu^f$, and variances, $\Sigma^T + \Sigma^f$, are constant.

As discussed in the main text, according to Bayes’ rule, the posterior distribution that represents the probability of the parameters given the data is

$$p(\varphi^T, \varphi^f | c, \xi) \propto p((c|\varphi^T, \varphi^f)p((\xi|\varphi^T)) [p(\varphi^f)p(\varphi^T)].$$  \hspace{1cm} (Equation 18)
where \( \{ \mathbf{c} \} = \{ \mathbf{c}_i : i = 1, \ldots, N_c \} \) and \( \{ \xi \} = \{ \xi_i : i = 1, \ldots, N_c \} \) represent the samples of the total signal and noise, respectively, with the different components of \( \mathbf{c} \) and \( \xi \) corresponding to the different channels of the cytometer. In Equation 18, the first bracket on the right-hand side corresponds to the likelihood, namely the joint probability of observing the data given the parameters, which we can redefine as

\[
\mathcal{L} = p(\{ \mathbf{c} \} | \mathbf{c}^T, \psi^f) p(\{ \xi \} | \psi^f) = \prod_{i=1}^{N_c} p(\mathbf{c}_i | \mathbf{c}^T, \psi^f) \prod_{i=1}^{N_c} p(\xi_i | \psi^f), \tag{Equation 19}
\]

The form of the likelihood (Equation 19) breaks the symmetry between the target and noise signals, and thus lifts the above-mentioned degeneracy between their parameters exhibited by Equation 17. The second bracket in the right-hand side of Equation 18, in turn, corresponds to the prior distributions of all the parameters of the problem, which we define in what follows.

**Approximated decomposition of the posterior into two separate problems**

It is worth noting that the posterior distribution (Equation 18) can be decomposed as follows:

\[
p(\psi^T, \psi^f | \{ \mathbf{c} \}, \{ \xi \}) = p(\psi^T | \mathbf{c}, \{ \xi \}) p(\psi^f | \{ \mathbf{c} \}, \{ \xi \}) \tag{Equation 20}
\]

where we have removed the dependency of the noise signal in the first term of the right-hand side, since only the total signal \( \mathbf{c} \) defines the parameters of the target distribution, as can be seen from the likelihood (Equation 19). The second term, on the other hand, is conditioned by both the noise and the total signal. Since we usually have as much data for the noise signal \( \{ \xi \} \) as for the signal of interest \( \{ \mathbf{c} \} \), we can approximately consider that the posterior distribution of the noise mixture parameter \( \psi^f \) is well represented by the noise data alone:

\[
p(\psi^T, \psi^f | \{ \mathbf{c} \}, \{ \xi \}) \approx p(\psi^T | \mathbf{c}, \{ \xi \}) p(\psi^f | \{ \xi \}) \tag{Equation 21}
\]

With this approximation, the problem can be decomposed in two separate subproblems: first, finding the probability distribution of the parameters \( \psi^f \) of the noise mixture; and second, finding the distribution of the parameters \( \psi^T \) of the convolution mixture, conditioned on the noise mixture parameters.

In the following section, we go over the mathematical details of the probability distributions that will allow us to sample from the posterior distribution.

**Relevant probability distributions**

In this section, we derive the main expressions that we need to sample our model. Our aim is to have a self-contained derivation of the sampling process of the posterior of a Gaussian normal mixture.

**Multivariate normal distribution with unknown mean and error**

Here we derive the main results of interest on multivariate normal distributions, which we will use when sampling the posterior distribution using normal bases.

**Likelihood.** The multivariate normal distribution for a set of independent identical samples (iid) \( \{ \mathbf{x} \} \) has the form, up to a scaling parameter,

\[
\mathcal{N}(\{ \mathbf{x} \}; \mu, \Sigma) \propto \det[\tau]^n/2 \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)^T \tau (\mathbf{x}_i - \mu) \right\} \tag{Equation 22}
\]

where \( \mu \) is the mean and \( \tau \) is the precision parameter and \( n \) is the number of cells being measured. The precision parameter relates to the covariance matrix as

\[
\tau = \Sigma^{-1} \tag{Equation 23}
\]

It is convenient to rearrange the distribution to make explicit the dependency of the summary statistics:

\[
\mathcal{N}(\{ \mathbf{x} \}; \mu, \Sigma) \propto \det[\tau]^n/2 \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)^T \tau (\mathbf{x}_i - \mu) \right\} 
\propto \det[\tau]^{n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)^T \tau (\mathbf{x}_i - \mu) + n(\bar{\mathbf{x}} - \mu)^T \tau (\bar{\mathbf{x}} - \mu) \right\} \tag{Equation 24}
\]

where the summary statistics are the mean,
\[
\mathbf{x} = \frac{1}{n} \sum_{i} x_i \tag{Equation 25}
\]

and the covariance,
\[
S^2 = \frac{1}{n} \sum_{i} (x_i - \bar{x})(x_i - \bar{x})^T \tag{Equation 26}
\]

**Conjugate prior.** A convenient prior for the multivariate distribution is a conjugate prior that allows us to obtain the analytic form of the other distributions:
\[p(\mu | \Sigma) = N(\mu; \mu_0, \Sigma)\] (Equation 27)
\[p(\Sigma) = \mathcal{W}^{-1}(\Sigma; \Sigma_0, \nu_0)\] (Equation 28)
where \(\mathcal{W}^{-1}\) is the inverse Wishart distribution. The hyperparameter \(\nu_0\) represents our confidence in the estimation of the mean: the higher \(\nu_0\) is, the closer to the mean we will be. The inverse Wishart distribution has the following shape, up to scaling terms:
\[
\mathcal{W}^{-1}(\mathbf{X}; \mathbf{\Sigma}, \nu) \propto \frac{\det[\mathbf{\Sigma}]^{(n/2)}}{\det[\mathbf{X}]^{(\nu+n_p)/2}} \exp\left[-\frac{1}{2} \text{Tr}[\mathbf{X}^{-1}\mathbf{\Sigma}]\right] \tag{Equation 29}
\]
where \(p\) is the number of dimensions. The parameter \(\mathbf{\Sigma}\) is the correlation matrix and \(\nu\) represents our confidence in the estimation of the \(\mathbf{\Sigma}\) correlation matrix.

**Posterior distribution.** The posterior distribution will have the following shape:
\[
p(\mu, \Sigma | \{x\}) \propto p(\{x\} | \mu, \Sigma)p(\mu | \Sigma, \nu_0, \mu_0)p(\Sigma | \Sigma_0, \nu_0)
\]
\[
\propto \det[\mathbf{\tau}]^{n/2} \exp\left\{ \frac{-1}{2} \left[ n \text{Tr}[\mathbf{S}^{\mathbf{\Sigma}}] + n(\mathbf{x} - \mathbf{\mu})^T \mathbf{\tau}(\mathbf{x} - \mathbf{\mu}) \right] \right\} \times \det[\mathbf{\tau}]^{(n+p)/2} \exp\left\{ \frac{-1}{2} \text{Tr}[\mathbf{\Sigma}_0 \mathbf{\tau}] \right\} \tag{Equation 30}
\]

We can use the posterior distribution to obtain different distributions of relevance.

**Conditional distribution: mean.** Retaining the terms involving the mean \(\mu\), the conditional distribution of the mean takes the form of a multivariate distribution:
\[
p(\mu | \mathbf{x}) \propto \exp\left\{ \frac{-1}{2} \left[ n(\mathbf{x} - \mathbf{\mu})^T \mathbf{\tau}(\mathbf{x} - \mathbf{\mu}) + \nu_0(\mathbf{\mu} - \mathbf{\mu}_0)^T \mathbf{\tau}(\mathbf{\mu} - \mathbf{\mu}_0) \right] \right\} \times \exp\left\{ \frac{-1}{2} \left[ \nu + 2\nu_0 \right] \mu - \mathbf{\mu}_0 \right\} \times N(\mu; \mathbf{\hat{\mu}}, \mathbf{\hat{\tau}}) \tag{Equation 31}
\]

where in the last step we completed the squares. This multivariate normal distribution has effective parameters
\[
\mathbf{\hat{\mu}} = \frac{n\mathbf{x} + \nu_0 \mathbf{\mu}_0}{n + \nu_0}, \quad \mathbf{\hat{\tau}} = (n + \nu_0) \mathbf{\tau}
\]

where \(\mathbf{\hat{\mu}}\) is a weighted version between the mean statistic and the prior mean, and \(\mathbf{\hat{\tau}}\) is an effective precision parameter. The \(\nu_0\) factor indicates how close we are from the prior mean.

**Conditional distribution: covariance.** Retaining the terms involving the covariance matrix (and its inverse, the precision matrix), the conditional distribution can be shown to take the form of an inverse Wishart distribution:
\[
p(\Sigma | \mu, \{x\}) \propto \det[\mathbf{\tau}]^{(n+p)/2} \exp\left\{ \frac{-1}{2} \left[ \text{Tr}[\mathbf{S}^2] + n(\mathbf{x} - \mathbf{\mu})^T \mathbf{\tau}(\mathbf{x} - \mathbf{\mu}) + \nu_0(\mathbf{\mu} - \mathbf{\mu}_0)^T \mathbf{\tau}(\mathbf{\mu} - \mathbf{\mu}_0) + \text{Tr}[\mathbf{\Sigma}_0 \mathbf{\tau}] \right] \right\}
\]
\[ \propto \det [r]^{(n+p+1)/2} \exp \left\{ - \frac{1}{2} \left[ \text{Tr} \left[ \left( nS_n^2 + n(\bar{x} - \mu)(\bar{x} - \mu)^T + k_0(\mu - \mu_0)(\mu - \mu_0)^T + \Sigma_0 \right) \tau \right] \right] \right\} \]

where the effective parameters are

\[ \bar{n} = n + \tau_0 + 1 \]

\[ \Sigma = nS_n^2 + n(\bar{x} - \mu)(\bar{x} - \mu)^T + k_0(\mu - \mu_0)(\mu - \mu_0)^T + \Sigma_0 \]

In this last expression, it is worth noting that each term represents a different kind of uncertainty: the first term is the uncertainty coming from the mean statistic, the second corresponds to the uncertainty with which the mean statistic represents the actual mean, the third term is the uncertainty coming from the prior mean, and the last term is the uncertainty coming from the prior itself.

Marginal distribution: covariance. One additional distribution that we will need is the marginal distribution of the variance:

\[ p(\Sigma|\{x\}) = \int p(\Sigma, \mu|\{x\})d\mu \propto \int \det [r]^{n/2} \exp \left\{ - \frac{1}{2} \left[ n\text{Tr}[S_n^2] + n(\bar{x} - \mu)^T \tau(\bar{x} - \mu) \right] \right\} \times \det [r]^{(n+p+2)/2} \exp \left\{ - \frac{1}{2} \left[ n\text{Tr}[S_n^2] + \text{Tr}[\Sigma_0] \right] \right\} \times \left[ \int \exp \left\{ - \frac{1}{2} \left[ k_0(\mu - \mu_0)^T \tau(\mu - \mu_0) + n(\bar{x} - \mu)^T \tau(\bar{x} - \mu) \right] \right\} d\mu \right]\]

Reorganizing the elements in (*)&

\[ (\ast) = (k_0 + n)\mu^T \tau \mu - 2\mu^T \tau (k_0 \mu_0 + n\bar{x}) + (n\bar{x}^T \tau \bar{x} + k_0 \mu_0 \tau \mu_0) \]

\[ = (\mu - \bar{\mu})^T \tau (\mu - \bar{\mu}) - \bar{\mu}^T \tau \mu + (n\bar{x}^T \tau \bar{x} + k_0 \mu_0 \tau \mu_0), \]

where we have defined,

\[ \bar{\mu} = \frac{k_0 \mu_0 + n\bar{x}}{k_0 + n} \]

\[ \bar{\tau} = (k_0 + n)\tau, \]

we can further regroup the part in (**).

\[ (** = -\frac{k_0 \mu_0^T \tau \mu_0 - n^2 \bar{x}^T \tau \bar{x} - 2k_0 n \mu_0 \bar{x}^T \tau \bar{x} + n^2 \tau \bar{x} \bar{x}^T \tau + k_0 \mu_0 \tau \mu_0 + nK_0 \bar{x}^T \tau \bar{x} + k_0 \mu_0 \tau \mu_0}{k_0 + n} \]

Inserting these terms

\[ p(\Sigma|\{x\}) \propto \det [r]^{(n+p+2)/2} \exp \left\{ - \frac{1}{2} \left[ \text{Tr} \left[ \left( nS_n^2 + \Sigma_0 + k_0 \mu_0 \right) (\bar{x} - \mu_0)(\bar{x} - \mu_0)^T \right] \right] \right\} \times \left[ \int \exp \left\{ - \frac{1}{2} \left[ (\mu - \bar{\mu})^T \tau (\mu - \bar{\mu}) \right] \right\} d\mu \right]\]

now we can calculate the integral as a multivariate normal:

\[ \propto \det [r]^{(n+p+2)/2} \exp \left\{ - \frac{1}{2} \left[ \text{Tr} \left[ \left( nS_n^2 + \Sigma_0 + k_0 \mu_0 \right) (\bar{x} - \mu_0)(\bar{x} - \mu_0)^T \right] \right] \right\} \det [\tau]^{-1/2} \]

\[ \propto \det [r]^{(n+p+1)/2} \exp \left\{ - \frac{1}{2} \left[ \text{Tr}[\Sigma] \right] \right\} \propto \mathcal{W}^{-1}(\Sigma, \bar{n}) \]
where in the last step we use the fact that \( \det \tilde{\mathbf{r}} = \det \mathbf{r} \) and the effective parameters of the inverse Wishart distribution are

\[
\hat{n} = n + k_0 + 1
\]  
(Equation 44)

\[
\Sigma = nS^2 + \frac{k_0n}{k_0 + n}(\mathbf{x} - \mu_0)(\mathbf{x} - \mu_0)^\top + \Sigma_0
\]  
(Equation 45)

**Posterior predictive distribution.** One last distribution of interest is the probability of a new data point given the already observed set of observations:

\[
p(y(x)) = \int \int p(y|\mu, \Sigma)p(\mu, \Sigma|\{x_i\})d\mu d\Sigma \sim \int \det \left[ \mathbf{r} \right]^{(n+k_0+1)/2} \times \exp \left\{ -\frac{1}{2} \left[ (y - \mu)^\top (y - \mu) + n(\mathbf{x} - \mu)^\top \mathbf{r}(\mathbf{x} - \mu) + k_0(\mathbf{\mu} - \mu_0)^\top \mathbf{r}(\mathbf{\mu} - \mu_0) \right] \right\} \times \exp \left\{ -\frac{1}{2} \left[ \text{Tr}(\{nS^2 + \Sigma_0\}\mathbf{r}) \right] \right\} d\mu d\Sigma
\]  
(Equation 46)

We can rearrange the elements in \((*)\), completing squares as we did when calculating the mean conditional distribution (Multivariate normal distribution with unknown mean and error):

\[
(*) = \frac{n\mathbf{x}^\top \mathbf{r} + y^\top \mathbf{r} + k_0 \mathbf{\mu}_0^\top \mathbf{r} \mathbf{\mu}_0 - \mu^\top \mathbf{r} \mu + \mu^\top \mathbf{r} \mu - \mathbf{\mu}^\top \mathbf{r} \mathbf{\mu}}{n + k_0 + 1},
\]  
(Equation 47)

where we have retained all the terms involving \(\mathbf{r}\) and \(\mathbf{\mu}\) this time, as they are necessary for integrating out. The effective parameters are now

\[
\hat{\mu} = \frac{n\mathbf{x} + k_0 \mathbf{\mu}_0 + y}{n + k_0 + 1}
\]  
(Equation 48)

\[
\hat{\mathbf{r}} = (n + k_0 + 1) \mathbf{r}
\]  
(Equation 49)

We can further rearrange the terms in \((**)*\) to obtain an expression with the \(y\) in quadrature:

\[
(***) = \frac{1}{n + k_0 + 1} \left[ (n\mathbf{x}^\top \mathbf{r} + y^\top \mathbf{r} + k_0 \mathbf{\mu}_0^\top \mathbf{r} \mathbf{\mu}_0)(n + k_0 + 1) - (n\mathbf{x} + k_0 \mathbf{\mu}_0 + y)^\top \mathbf{r}(n\mathbf{x} + k_0 \mathbf{\mu}_0 + y) \right]
\]

\[
= \frac{1}{n + k_0 + 1} \left[ n(\mathbf{y} - \mathbf{x})^\top \mathbf{r}(\mathbf{y} - \mathbf{x}) + k_0(\mathbf{y} - \mathbf{\mu}_0)^\top \mathbf{r}(\mathbf{y} - \mathbf{\mu}_0) + n k_0(\mathbf{\mu}_0 - \mathbf{x})^\top \mathbf{r}(\mathbf{\mu}_0 - \mathbf{x}) \right]
\]

\[
= \frac{(n + k_0)}{n + k_0 + 1}(\mathbf{y} - \mathbf{\mu})^\top \mathbf{r}(\mathbf{y} - \mathbf{\mu}) + \frac{n k_0}{n + k_0 + 1}(\mathbf{\mu}_0 - \mathbf{x})^\top \mathbf{r}(\mathbf{\mu}_0 - \mathbf{x})
\]

\[
= m^0(\mathbf{y} - \mathbf{\mu})^\top \mathbf{r}(\mathbf{y} - \mathbf{\mu}) + m^1(\mathbf{\mu}_0 - \mathbf{x})^\top \mathbf{r}(\mathbf{\mu}_0 - \mathbf{x})
\]  
(Equation 50)

where in the second to the third step we group by quadrature and complete squares to group the terms with \(y\), and then group by quadrature the terms on the left. The effective parameters are

\[
\hat{\mu} = \frac{(n\mathbf{x} + k_0 \mathbf{\mu}_0)}{n + k_0}, \quad m^0 = \frac{n + k_0}{n + k_0 + 1}, \quad m^1 = \frac{n k_0}{n + k_0 + 1}
\]  
(Equation 51)
We can now insert the obtained term in the original expression:

\[
p(y|x) \propto \int \det |\tau|^{(n+p+3)/2} \times \exp \left\{ -\frac{1}{2} \left[ m^1 (y - \tilde{\mu}^r)^T \tau (y - \tilde{\mu}^r) + m^1 (\mu_0 - \bar{x})^T \tau (\mu_0 - \bar{x}) + \text{Tr}[\tau S_x^2 + \tau \Sigma_0] \right] \right\} \times \]
\[
\exp \left\{ -\frac{1}{2} \left[ (\mu - \tilde{\mu})^T \tau (\mu - \tilde{\mu}) \right] \right\} d\mu d\Sigma
\]

(Equation 52)

= \int \det |\tau|^{(n+p+3)/2} \exp \left\{ -\frac{1}{2} \left[ \text{Tr}[\tau \Sigma_0] \right] \right\} \times \int \exp \left\{ -\frac{1}{2} \left[ (\mu - \tilde{\mu})^T \tau (\mu - \tilde{\mu}) \right] \right\} d\mu d\Sigma

where the effective covariance has the form,

\[
\tilde{\Sigma} = m^0 (y - \tilde{\mu}^r) (y - \tilde{\mu}^r)^T + m^1 (\mu_0 - \bar{x})(\mu_0 - \bar{x})^T + n S_x^2 + \Sigma_0
\]

(Equation 53)

The mean parameter in Equation 52 only appears in the term in brackets, so we can integrate it in a straightforward manner as a multivariate normal integral:

\[
p(y|x) \propto \det |\tau|^{(n+p+3)/2} \exp \left\{ -\frac{1}{2} \left[ \text{Tr}[\tau \Sigma_0] \right] \right\} \times \int \det |\tau|^{(n+p+2)/2} \exp \left\{ -\frac{1}{2} \left[ \text{Tr}[\tau \Sigma_0] \right] \right\} \times d\Sigma
\]

(Equation 54)

where \( \det[\tau^\nu] \propto \det[\tau] \). The last integral is an inverse Wishart that we can integrate directly, leading to

\[
p(y|x) \propto \det |\tilde{\Sigma}|^{-(n+p+1)/2}
\]

(Equation 55)

We can now reorganize the effective covariance. If we define the matrix

\[
\tilde{\Sigma}' = \frac{1}{m^0} \left( m^1 (\mu_0 - \bar{x})(\mu_0 - \bar{x})^T + n S_x^2 + \Sigma_0 \right)
\]

(Equation 56)

we can rewrite the effective covariance \( \tilde{\Sigma} \) as

\[
\tilde{\Sigma} = \left( y - \tilde{\mu}^r \right)^T \tilde{\Sigma}' \left( y - \tilde{\mu}^r \right) + 1
\]

(Equation 57)

Insert this expression in Equation 55 we obtain

\[
p(y|x) \propto \det \left[ \left( y - \tilde{\mu}^r \right)^T \tilde{\Sigma}' \left( y - \tilde{\mu}^r \right) + 1 \right] \times \frac{1}{n^0} \tilde{\Sigma}'
\]

\[
\propto \det \left[ \left( y - \tilde{\mu}^r \right)^T \tilde{\Sigma}' \left( y - \tilde{\mu}^r \right) + 1 \right] \left( \frac{1}{n^0} \tilde{\Sigma}' \right)^{-(n+p+1)/2}
\]

\[
\left( \frac{1}{n^0} \tilde{\Sigma}' \right)^{-(n+p+1)/2} = \left( \frac{1}{n^0} \tilde{\Sigma}' \right)^{-(n+p+1)/2}
\]

(Equation 58)

which is a multivariate T-distribution with

\[
m^0 = \frac{nK_0}{n + K_0 + 1}, \quad m^1 = \frac{nK_0 + nK_0}{n + K_0 + 1}, \quad \tilde{\nu} = n + r_0 + 1 - p
\]

(Equation 59)

\[
\tilde{\mu}^r = \frac{n(\bar{x} + K_0 \mu_0)}{n + K_0}
\]

(Equation 60)

\[
\tilde{\Sigma}' = \frac{1}{m^0} \left( m^1 (\mu_0 - \bar{x})(\mu_0 - \bar{x})^T + n S_x^2 + \Sigma_0 \right)
\]

(Equation 61)

**Finite mixture distributions**

We derive in this section the statistics relevant to mixture models.
**Likelihood.** The likelihood of a finite mixture model with $K$ components has the form

$$p(\{x\} | \{\varphi\}, \omega) = \prod_i \left( \sum_{j=1}^K \omega_j p(x_i | \varphi_j) \right).$$  \hfill (Equation 62)

where the vector sets $\{x\} = \{x_i : i = 1, \ldots, N\}$ and $\{\varphi\} = \{\varphi_j : j = 1, \ldots, K\}$ run over the number of samples (cells) $N$ and the number of mixture components $K$, respectively (in what follows we use the subindices $i$ and $j$ with those two distinct meanings) (We remind the reader that the dimension of the vectors $x_i$ and $\varphi_j$ is equal to the number of measurement channels). The set of parameters $\omega$ are called the weights of the mixture model, and $p(x_i | \varphi_j)$ are the base distributions. We can extend this model to introduce a set of hidden indicator variables $\{z\}$, defined as

$$z_{ij} = \delta_{ki}$$  \hfill (Equation 63)

for some $k \in \{1, \ldots, K\}$, and where $\delta_{ki}$ is the Kronecker delta. This variable basically tells from which distribution $p(x_i | \varphi_j)$ the variable came from. Using this set of hidden variables, our model (Equation 62) can be rewritten as

$$p(\{x\}, \{z\} | \{\varphi\}, \omega) = \prod_i \prod_j p(x_i | \varphi_j)^{z_{ij}} \omega_j$$  \hfill (Equation 64)

It is straightforward to see that, if we take the marginal distribution over the hidden variables, we recover the original distribution:

$$p(\{x\} | \{\varphi\}, \omega) = \prod_i \left( \sum_j \sum_1^K p(x_i | \varphi_j)^{z_{ij}} \omega_j \right) = \prod_i \left( \sum_j p(x_i | \varphi_j) \omega_j \right)$$  \hfill (Equation 65)

where only when the indicator variable is one of the corresponding term survives. The use of indicator variables makes it possible to compute analytically the posterior distribution of the mixture model, as the base distributions are now in product form. The hidden indicator variables are not known, thus we will have to sample from them as well.

**Conjugate prior.** A conjugate prior for the mixture model is

$$p(\omega | \alpha) = \text{Dirichlet}(\omega | \alpha)$$  \hfill (Equation 66)

where the Dirichlet distribution has the form

$$\text{Dirichlet}(\omega | \alpha) \propto \prod_i \omega_j^{\alpha_i - 1}$$  \hfill (Equation 67)

The hyperprior parameters $\alpha$ are usually set to be symmetrical and to scale with the number of mixture components:

$$\alpha_i = \alpha / K \quad \forall i$$  \hfill (Equation 68)

In this way, the prior distribution only depends on one hyperparameter $\alpha$ that indicates the strength from the uniform weights.

**Posterior distribution.** Putting together the likelihood and the prior distribution, the posterior of the mixture model is

$$p(\{\varphi\}, \omega | \{x\}, \{z\}) \propto \prod_i \left( \prod_j p(x_i | \varphi_j)^{z_{ij}} \omega_j \right)^{\alpha_i / K - 1} p(\varphi^0_j).$$  \hfill (Equation 69)

where the last term is the set of priors for each base distribution.

**Conditional distribution: weights.** Taking the terms from the posterior that involve the weights:

$$p(\omega | \{z\}) \propto \prod_i \left( \prod_j \omega_j^{\alpha_i} \right)^{\alpha_i / K - 1} = \text{Dirichlet}(\omega | \tilde{n})$$  \hfill (Equation 70)

where $\tilde{n} = \sum z_{ij} + \alpha$.

**Conditional distribution: indicator variables.** As we already mentioned, the indicator variables are not known, so we have to sample from them too. As we are considering identically independent samples, we can obtain the conditional distribution from each indicator variable independently as

$$p(z_i | \omega, \{\varphi\}, \{x\}) \propto \prod_j p(x_i | \varphi_j)^{z_{ij}} \omega_j^{\alpha_i - 1} = \text{Multinomial}(z_i ; 1, \{p(x_i | \varphi_j) \omega_j\})$$  \hfill (Equation 71)
It is worth noting that the indicator will be sampled from a particular base distribution for the weights of that base, but also by how well that sample lies inside the base distribution.

**Conditional distribution: base distribution parameters.** Finally, because the base distributions are in product form due to the introduction of the indicator variables, the parameters of each base can be computed independently as

\[ p(\phi_j | \{x\}, \{z\}) \propto p(\{x\}) p(\phi_j^z) \]  

(Equation 72)

where \( \{x\} \) represents the subsample of cells whose indicator variable belongs to the corresponding mixture component.

**Infinite mixture distributions**

In this section we define basic results from infinite Dirichlet processes (for an insightful tutorial see Li et al. 32) that allow us to consider infinite mixtures.

**Taking the infinite limit.** In order to take the limit to infinite clusters, we need to remove the dependence from the number of clusters in a mixture model. For that, we need to remove the dependence on the weights of our probability distribution. Consider the moment a basis distribution that is uniform in space. The joint probability distribution conditioned on the priors would be

\[ p(\{z\}, \omega | \alpha) = p(\{z\} | \omega) p(\omega | \alpha) \]  

(Equation 73)

where the first term in the right-hand side is the likelihood of the indicator variables as in Equation 64, also given in Equation 71, which is a multinomial distribution. The second term is the prior distribution of the mixture distribution (Equation 67), which is a Dirichlet distribution. From this expression, we can calculate the marginal distribution of the indicator variables conditioned on the prior parameter:

\[ p(\{z\} | \alpha) = \int p(\{z\} | \omega) p(\omega | \alpha) \, d\omega = \int \prod_i \left( \prod_j \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)} \Gamma(\alpha/K) \prod_j \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)}  \right) \, d\omega \]

(Equation 74)

where the third equality makes use of the statistic \( n_j = \sum z_{ij} \). The integral can be identified as a multinomial distribution without the scaling factor.

The expression above still contains explicitly the dependence on the number of components in the mixture \( K \), and thus its limit \( K \to \infty \) cannot be computed in a straightforward manner. To have a more amenable expression, let us consider that we take out the sample \( l \) and reassign it to a new cluster. The probability of this sample to be assigned to any of the clusters, conditioned on all the other indicator variables, is

\[ p(z_k = 1 | \{z\}_{l \neq k}, \alpha) = \frac{\frac{\Gamma(n_l + \delta_k + \alpha / K)}{\Gamma(n_l + \alpha / K)}}{\frac{\Gamma(n_l + \alpha / K)}{\Gamma(n_l + \alpha / K)}} \]

(Equation 75)

where have made explicit the range of cells over which the samples \( \{z\} \) are taken in each case. To that end, we define \( \{z\}_{l \neq k} = \{z\}_{l \neq 1, \ldots, l \neq k, \ldots, n} \) and \( n_{l \neq k} = \sum_{l \neq k} z_{ij} \) to refer to the subsets that contain all elements except \( l \). With this expression, it is straightforward to take the limit to infinite components:

\[ p_\alpha(z_k = 1 | \{z\}_{l \neq k}, \alpha) = \lim_{K \to \infty} p(z_k = 1 | \{z\}_{l \neq k}, \alpha) = \frac{n_{l \neq k}}{n + \alpha - 1} \]  

(Equation 76)

We now have the probability that a sample is assigned to a base that has other indicator variables. Let us now consider for the moment that we have \( K \) bases with assigned samples. The probability that a sample is assigned to a new base will be:

\[ p_\alpha(z_k = 1 | \{z\}_{l \neq k}, \alpha) = 1 - \sum_{j=1}^{K} p_\alpha(z_k = 1 | \{z\}_{l \neq k}, \alpha) \]

(Equation 77)

\[ = 1 - \sum_{j=1}^{K} \frac{n_{kj}}{n + \alpha - 1} = 1 - \frac{n - 1}{n + \alpha - 1} = \frac{\alpha}{n + \alpha - 1} \]

\[ ll \]

\[ 107941, October 20, 2023 \]

\[ iScience 26, 107941, October 20, 2023 \]
There is a non-zero probability that the cell will be assigned to a new base that was not populated before, and the probability of populating this new cluster will depend on the hyperparameter $\alpha$.

**Adding a non-uniform basis.** The results above have been derived considering a uniform basis distribution. In the most general case, the basis will be non-uniform. From Equation 64 it is very easy to see that if we had a non-uniform distribution, the corresponding term will drop out of the integral (Equation 74). Proceeding in the same way as before, we can directly predict the new outcomes in the case of an infinite mixture as a function of already observed data. The probabilities of a new assignment will be

\[ p_\alpha(z_k = 1 | \{ z \}, \{ \varphi \}, \alpha) \propto \frac{n_{-i}}{n + \alpha} p(x_i | \varphi_k) \]  

(Equation 78)

for a basis with a populated sample and

\[ p_\alpha(z_{i,k+1} = 1 | \{ z \}, \{ \varphi \}, \alpha) \propto \alpha \frac{1}{n + \alpha} p(x_i | \varphi^0) \]  

(Equation 79)

for the creation of a new basis.

**Using the predictive posterior distribution.** Similar way to the approach used in Multivariate normal distribution with unknown mean and error, we can directly predict the new outcomes in the case of an infinite mixture as a function of already observed data. The probabilities of a new assignment will be

\[ p_\alpha(z_k = 1 | \{ z \}, \{ \varphi \}, \alpha) \propto \frac{n_{-i}}{n + \alpha} p(x_i | \{ x \}_k) \]  

(Equation 80)

for a basis with a populated sample and

\[ p_\alpha(z_{i,k+1} = 1 | \{ z \}, \{ \varphi \}, \alpha) \propto \alpha \frac{1}{n + \alpha} p(x_i | \varphi^0) \]  

(Equation 81)

for a new basis.

**Modifications of the convolution distribution.** Introducing indicator variables as indicated in Finite mixture distributions, the convolution of two multivariate mixture distributions like the one described by Equation 17 takes the form

\[ p(\{ x \}, \{ z \} | \{ \varphi^i \}, \{ \omega^i \}, \{ \varphi^j \}, \{ \varphi^T \}) = \prod_i \prod_j \prod_{i,j}^{K^i} N(x_i | \mu^i_j + (\omega^i_j)^T (\varphi^i)_{1:k}^{T} (\varphi^j), \Sigma^i_j + \Sigma^T) \]  

(Equation 82)

where the indicator variable is now 1 if the sample $i$ belongs to the noise base $j$ and target base $k$. Considering the approximation described in Approximated decomposition of the posterior into two separate problems, the only sampling parameters that we have to go over will be $\{ \varphi^i \}$, the weights $\omega^T$ and the indicator variables $z$. The main challenge is that there is no close form to group all the terms involving $\Sigma^T$ in single effective distributions as the ones derived in Multivariate normal distribution with unknown mean and error. In order to get a tractable expression, we would like to transform the convolved covariances in such a way that the following expression follows:

\[ M(\Sigma^i_j + \Sigma^T) = (M + A) \Sigma^{-1}_{k} \]  

(Equation 83)

for given matrices $A$, $M$. Isolating, $A$ we obtain that

\[ A = -M(\Sigma^i_j + \Sigma^T)^{-1} \Sigma^T \]  

(Equation 84)

Now, in most practical cases we can consider that the convolved covariance will be close to the expected covariance matrix, and we can approximate the expression above by

\[ A \sim -M \Sigma^T \Sigma^T \]  

(Equation 85)

where the effective expected covariance is

\[ \Sigma_k = n_k \Sigma^2_k + \kappa_0 (\overline{x}_k - \mu_0) (\overline{x}_k - \mu_0)^T + \Sigma_0 \]  

(Equation 86)

All the results derived in the preceding sections take into account this approximation:

\[ \mu \rightarrow \mu^T + \mu^i \]  

(Equation 87)

\[ \Sigma \rightarrow \Sigma^T + \Sigma^T \]  

(Equation 88)
The weights of the target distribution (Finite mixture distributions) will be computed using the sum over the samples for all the noise samples

\[ n_k^T = \sum_{ij} n_{jk} \]  

(Equation 90)

Sampling from the mean and covariance parameters in the distribution needs, however, a closer look.

**Prior distribution.** In order for the prior to be conjugated, we have to scale the prior distribution in terms of the prior of the noise distribution.

\[ p(\mu|\Sigma) = N(\mu; \mu_0, \kappa_0 \Sigma) \]  

(Equation 92)

\[ p(\Sigma) = W^{-1}(\Sigma; \Sigma_0, \Sigma^i, \rho_0) \]  

(Equation 93)

where the effective covariance is

\[ \hat{\Sigma} = \Sigma^T + \Sigma^i \]  

(Equation 94)

**Conditional distribution: covariance.** If we focus the analysis in a single set of target parameters \( \phi \), the posterior probability of the variance and the mean has the form

\[ p(\phi^i|\{\phi^j\}, \{x\}, \{z\}) \propto \prod_i N(x|\mu^i, \mu^i + \mu^i, \Sigma^i + \Sigma^i)^{n_k} p(\mu^i + \mu^i)p(\Sigma^i + \Sigma^i) \]  

(Equation 95)

In contrast with the case without convolution, the normal distributions cannot be grouped together in a simple distribution that can be sampled by standard procedures. We can group the data coming from a specific dataset, 

\[ p(\phi^i|\{\phi^j\}, \{x\}, \{z\}) \propto \det \left( \Sigma^i + \Sigma^i \right)^{-\gamma_{n_k+\kappa_0}} \exp \left\{ - \frac{1}{2} \left[ \Sigma_{jk} \left( \Sigma^i + \Sigma^i \right)^{-1} \right] \right\} \]  

(Equation 96)

where the variance is

\[ \Sigma_{jk} = n_k \Sigma^2 + n_k \left( \Sigma^2 - \mu^2 - \mu^2 \right) \left( x - \mu^2 + \mu^2 \right)^T + \kappa_0 (\mu^2 - \mu^2) (\mu^2 - \mu^2)^T + \Sigma_0 \]  

(Equation 97)

Within the proposed approximation, we can now group all the different distributions:

\[ \Sigma_{jk} = \sum_i \Sigma_{jk} \left( 1 - x_j \right) \]  

(Equation 98)

and we can sample new covariance matrices for the target with the inverse Wishart distribution.

\[ \Sigma^i \sim W^{-1}(\Sigma, n_k + \rho_0 + 1) \]  

(Equation 99)

**Conditional distribution: mean.** The mean distribution can be easily computed, regrouping all the terms in the following summary statistics:

\[ \tilde{\tau}_k = \sum_i (n_k + \kappa_0) \left( \Sigma^i + \Sigma^i \right)^{-1} \]  

\[ \tilde{\mu}_k = \Sigma \left( \sum_i \left( \Sigma^i + \Sigma^i \right)^{-1} \left( n_k \Sigma^2 + \kappa_0 \mu^2 \right) \right). \]

and sampling from a multivariate normal,

\[ \mu^i \sim N(\mu, \tilde{\tau}) \]  

(Equation 100)

**Hyperparameter selection**

Our approach the has five hyperparameters:

- \( \alpha \): Determines how close we are from a uniform distribution of weights (finite mixtures), or the potential of generating a new basis (infinite mixture).
- \( \mu_0 \): The center of the prior normal distribution \( \mu_0 \).
- \( k_0 \): The confidence we have of being close to \( \mu_0 \).
- \( \Sigma_0 \): The covariance of the prior normal distribution.
- \( \nu_0 \): The confidence we have of being close to \( \nu_0 \).

Flow cytometry applications have in general large datasets, making the approach quite insensitive to the choice of \( \alpha, k_0 \) and \( \nu_0 \), as the statistics will dominate the model. The choice of mean and covariance matrix, however, depend on the data distribution. Appropriate estimators of these parameters are the mean and the covariance matrix of the whole dataset. This imposes a soft-informative prior over the region where the density should lie. In cases where the distribution has fat tails, the density prior can be very flat because of the outliers, making it difficult for the algorithm to find the correct distribution as the prior spreads over the basis components. In these cases, narrower prior covariance matrices can help to fit the model correctly.

**Gibbs sampling algorithms**

In this section, we describe the algorithms to sample the noise distribution according to the approximation described in Approximated decomposition of the posterior into two separate problems.

**Efficient updating of summary statistics**

Once we have computed the mean or the variance over a set of samples \( \{x\} \) and we add or remove a single sample, it is possible to update the statistic without having to recompute the metric fully over the new set, which in general will be more time-consuming as we will have to go over a sum over all the size of the set. If we remove a sample:

\[
\begin{align*}
    n_j^{rem} &= n_j - 1 \quad \text{(Equation 101)} \\
    \bar{x}_j^{rem} &= \frac{n_j \bar{x}_j - x_i}{n_j^{rem}} \quad \text{(Equation 102)} \\
    S_j^{2,rem} &= \frac{1}{n_j^{rem}} \left( n_j^2 S_j^2 + n_j \bar{x}_j \bar{x}_j^T - x_i x_i^T \right) - \bar{x}_j^{rem} \bar{x}_j^{rem T} \quad \text{(Equation 103)}
\end{align*}
\]

On the other hand, if we add a sample:

\[
\begin{align*}
    n_j^{add} &= n_j + 1 \quad \text{(Equation 104)} \\
    \bar{x}_j^{add} &= \frac{n_j \bar{x}_j + x_i}{n_j^{add}} \quad \text{(Equation 105)} \\
    S_j^{2,add} &= \frac{1}{n_j^{add}} \left( n_j^2 S_j^2 + n_j \bar{x}_j \bar{x}_j^T + x_i x_i^T \right) - \bar{x}_j^{add} \bar{x}_j^{add T} \quad \text{(Equation 106)}
\end{align*}
\]

**Finite normal mixture distribution**

The procedure in this case is as follows:

0 Initialize the parameters.

0.1 Initialize the indicator variables \( \{z\} \): Assign to each cell one of the \( K \) normal distributions using any initialization procedure (random assignment, k-means...).

0.2 Compute the statistics of each base:

\[
\begin{align*}
    n_j &= \sum_i z_{ij} \quad x_j = \frac{1}{n_j} \sum_i z_{ij} x_i \quad S_j^2 = \frac{1}{n_j} \sum_i z_{ij} x_i x_i^T - x_j x_j^T
\end{align*}
\]

0.3 Initialize parameters of each base:

\[
\begin{align*}
    \mu_j &= x_j \quad \Sigma_j = S_j^2 \quad \omega_j = \frac{1}{n} \sum_i z_{ij}
\end{align*}
\]

where \( n = \sum_j n_j \) is the size of the dataset.
Sampling. For $N$ iterations do

1. Sample indicator variables as in Finite mixture distributions. For each sample $i$:
   1.1. Compute weights: $w_i = \omega_j p(x_i|\omega_j)$.
   1.2. Sample new indicator: $z \sim \text{Multinomial}(1, w)$.

2. Sample weights as from Finite mixture distributions. For each base $j = 1, \ldots, K$ do
   2.1. Compute $\hat{n}_j = \sum z_{ij} + \alpha_1 / \alpha_2$.
   2.2. Sample new weights: $\omega_j \sim \text{Dirichlet}(\hat{n})$.

3. Sample variances and mean. For each base $j = 1, \ldots, K$ do
   3.1. Compute summary statistics:
   
   $$n_j = \sum z_{ij}, \quad \bar{x}_j = \frac{1}{n_j} \sum z_{ij} x_i, \quad S_j^2 = \frac{1}{n_j} \sum z_{ij} (x_i - \bar{x}_j)^T \bar{x}_j$$

   3.2. Sample new covariance (Multivariate normal distribution with unknown mean and error):
   
   $$\hat{n}_j = n_j + \alpha_0 + 1$$
   $$\hat{\Sigma}_j = n_j S_j^2 + \frac{n_j}{\hat{n}_j} (x_i - \mu_j) (x_i - \mu_j)^T + \alpha_0 (\mu_j - \mu_0) (\mu_j - \mu_0)^T + \Sigma_0$$
   $$\Sigma_j \sim W^{-1}(\hat{n}_j, \hat{\Sigma}_j)$$

   3.2. Sample new mean (Multivariate normal distribution with unknown mean and error):
   
   $$\hat{\mu}_j = \frac{n_j \bar{x}_j + \alpha_0 \mu_0}{\hat{n}_j + \alpha_0}$$
   $$\hat{\tau}_j = (\hat{n}_j + \alpha_0) \tau_j$$
   $$\mu_j \sim \mathcal{N}(\hat{\mu}_j, \Sigma)$$

Infinite normal mixture distribution

The approach in this case has the following steps:

0. Initialize the parameters.
   0.1. Initialize the indicator variables ($z$): Assign to each cell one of the $K$ normal distributions using any initialization procedure (random assignment, k-means...).
   0.2. Compute the statistics of each base:
   
   $$n_j = \sum z_{ij}, \quad \bar{x}_j = \frac{1}{n_j} \sum z_{ij} x_i, \quad S_j^2 = \frac{1}{n_j} \sum z_{ij} (x_i - \bar{x}_j)^T \bar{x}_j$$

0.3. Initialize parameters of each base:
   
   $$\mu_j = \bar{x}_j, \quad \Sigma_j = S_j^2, \quad \omega_j = \frac{1}{n_j} \sum z_{ij}$$

where $n = \sum n_j$ is the size of the dataset.

1. Sampling. For $N$ iterations, do
   1.1. Reassign samples. For each sample $i \in \{1, \ldots, n\}$
1.1.1 Remove sample $i$. Consider sample $z_i = 1$.

If $n_i = 1$ (the only sample assigned to that base distribution), remove the distribution from the active basis. Otherwise recompute the summary statistics for base distribution $j$ removing one sample as described in Efficient updating of summary statistics.

1.1.2 Compute weights (not-normalized) for reassigning the sample as described in Multivariate normal distribution with unknown mean and error and Infinite mixture distributions.

For the active bases $j \in \{1, \ldots, K\}$,

\[
\begin{align*}
    m_j^0 &= \frac{n_j + \kappa_0}{n_j + \kappa_0 + 1} \\
    m_j^1 &= \frac{n_j + \kappa_0}{n_j + \kappa_0 + 1} \\
    \mu_j &= \frac{(n_j \bar{x}_j + \kappa_0 \mu_0)}{n_j + \kappa_0} \\
    \Sigma_j' &= \frac{1}{m'} \left( m' (\mu_0 - \bar{x})(\mu_0 - \bar{x})^T + \eta j \Sigma_0 + \Sigma_0 \right) \\
    w_j &= \frac{n_j}{n + \alpha} - 1 t(y; \mu_j', \Sigma_j')
\end{align*}
\]

and for creating a new basis,

\[
    w_{K+1} = \frac{\alpha}{n + \alpha - 1} N(x; \mu_j', \Sigma_j')
\]

1.1.3 Sample new indicator: $z_i^{\text{new}} \sim \text{Multinomial}(1, w)$.

1.1.4 Update statistics.

If $z_{K+1} = 1$ create a new basis distribution and assign new statistics to it,

\[
    n_{K+1} = 1 \quad x_{K+1} = \bar{x} \quad \Sigma_{K+1} = 0
\]

else, update the statistics adding a term as described in Efficient updating of summary statistics.

1.2 Sample basis parameters. For each basis $j \in \{1, \ldots, K\}$:

1.2.1 Sample the weights (Finite mixture distributions):

\[
    \tilde{n}_j = n_j + \alpha \quad \omega \sim \text{Dirichlet}(\tilde{n})
\]

1.2.2 Sample the covariance matrix (Multivariate normal distribution with unknown mean and error):

\[
    \tilde{n}_j = n_j + \nu_0 + 1 \quad \Sigma_j = n \Sigma_j^2 + \frac{\nu_0 \eta_j}{\nu_0 + \eta_j} (x_j - \mu_0) (x_j - \mu_0)^T + \Sigma_0 \\
    \Sigma_j \sim W^{-1}(\tilde{\Sigma}_j, \tilde{n}_j)
\]

1.2.3 Sample the mean (Multivariate normal distribution with unknown mean and error):

\[
    \tilde{\mu}_j = \frac{n \bar{x}_j + \kappa_0 \mu_0}{n_j + \kappa_0} \quad \Sigma_j = \Sigma_j / (n_j + \kappa_0) \\
    \mu_j \sim N(\tilde{\mu}_j, \Sigma_j)
\]

**Modifications to the convolved distribution**

The preceding algorithms are the sampling algorithms for non-convolved finite and infinite mixture sampling. The convolved cases are exactly the same, but changing the equations with the modifications described in Modifications of the convolution distribution. In addition to this, it is necessary to add a step in the sampling loop to sample new parameters from the already fitted noise distribution $\{\theta_j^j\}$. 


Efficiency assessment

So far, the efficiency of deconvolution methods has been assessed using quantifiers applicable to point estimates, which are the ones proposed in the literature to date. To compare the efficiency of our model with previous methods that do not obey the positivity nor the normalization conditions, we can use the Mean Integrated Squared Estimation (MISE) measure:

$$MISE = \int_{-\infty}^{\infty} (p_{\text{true}}(x) - p_{\text{est}}(x))^2 \, dx,$$  
(Equation 107)

where $p_{\text{true}}$ is the real target distribution and $p_{\text{est}}$ is the point estimate of the deconvolved distribution. This is the traditional measure of convergence, but it is hard to interpret as it only has a lower bound. To address this issue, we introduce the mean integrated overlap (MIO):

$$MIO = 1 - \frac{1}{2} \int_{-\infty}^{\infty} |p_{\text{true}}(x) - p_{\text{est}}(x)| \, dx,$$  
(Equation 108)

This measure has the property that it is bounded in the interval $[0, 1]$ if the true and the estimated distributions are normalized, with 0 corresponding to the case of no overlap between distributions, and 1 to the case of complete overlap. Values below zero can be obtained if the estimated distribution does not follow the positivity requirement nor the normalization condition, as it is the case in FFT-based deconvolutions.

Synthetic data

For the target distribution we generated samples from three distributions: symmetric bimodal, asymmetric bimodal and skew symmetric distributions (Figure S1). This choice of distributions intends to capture the features present in real datasets, such as the presence of multiple peaks, different cluster sizes, and the generally non-Gaussian character of the data.27 As for the noise, we generated a set of three different noise datasets containing normal, skewed, and Student’s t distributions (the last two of which exhibit fat tails) (Figure S1), in order to test the flexibility of the method against very dissimilar autofluorescence profiles. In order to check the impact of the noise strength, the convolutions between target and noise were generated at two signal-to-noise ratios (SNR). In our context, we define the SNR as the ratio between target and signal variances for the whole dataset. We chose a case with negligible noise (SNR = 10), and a difficult case where the noise is of the same magnitude as the signal (SNR = 1). We generated sample datasets of different sizes, with 100, 1000 and, 10000 samples. The high range of these values is representative of typical single-cell flow cytometry experiments. The combination of the different target distribution types, noise distribution types, SNRs, and sample sizes generates a collection of 54 datasets with known ground truth.

Software

The software pipeline presented here, including MCMC for finite and infinite normal mixtures, has been implemented as a Julia package (scBayesDeconv.jl). The source code, with manual compilation and installation instructions, as well as full documentation and a notebook with examples of use, can be obtained publicly from GitHub (https://github.com/dsb-lab/scBayesDeconv.jl).