HHS Public Access

Author manuscript

Ann Appl Stat. Author manuscript; available in PMC 2018 August 29.

Published in final edited form as:

Ann Appl Stat. 2018 March ; 12(1): 609–632. doi:10.1214/17-AOAS1110.

A UNIFIED STATISTICAL FRAMEWORK FOR SINGLE CELL AND BULK RNA SEQUENCING DATA

Lingxue Zhu[*], **Jing Lei**[*,1], **Bernie Devlin**[†,2], and **Kathryn Roeder**[*,2,3]

[*]Carnegie Mellon University

[†]University of Pittsburgh

Abstract

Recent advances in technology have enabled the measurement of RNA levels for individual cells. Compared to traditional tissue-level bulk RNA-seq data, single cell sequencing yields valuable insights about gene expression profiles for different cell types, which is potentially critical for understanding many complex human diseases. However, developing quantitative tools for such data remains challenging because of high levels of technical noise, especially the "dropout" events. A "dropout" happens when the RNA for a gene fails to be amplified prior to sequencing, producing a "false" zero in the observed data. In this paper, we propose a Unified RNA-Sequencing Model (URSM) for both single cell and bulk RNA-seq data, formulated as a hierarchical model. URSM borrows the strength from both data sources and carefully models the dropouts in single cell data, leading to a more accurate estimation of cell type specific gene expression profile. In addition, URSM naturally provides inference on the dropout entries in single cell data that need to be imputed for downstream analyses, as well as the mixing proportions of different cell types in bulk samples. We adopt an empirical Bayes' approach, where parameters are estimated using the EM algorithm and approximate inference is obtained by Gibbs sampling. Simulation results illustrate that URSM outperforms existing approaches both in correcting for dropouts in single cell data, as well as in deconvolving bulk samples. We also demonstrate an application to gene expression data on fetal brains, where our model successfully imputes the dropout genes and reveals cell type specific expression patterns.

keywords and phrases

Single cell RNA sequencing; hierarchical model; empirical Bayes; Gibbs sampling; EM algorithm

[1]Supported by NSF Grants DMS-1553884 and DMS-1407771.
[2]Supported by Simons Foundation Grants SF402281, SFARI124827, and National Institute of Mental Health Grant R37MH057881.
[3]Supported by National Institute of Mental Health Grant R01MH109900.

L. Zhu, J. Lei, K. Roeder, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213, USA, lzhu@cmu.edu, jinglei@andrew.cmu.edu, roeder@andrew.cmu.edu.
B. Devlin, Department of Psychiatry and Human Genetics, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, Pennsylvania 15213, USA, devlinbj@upmc.edu

SUPPLEMENTARY MATERIAL

Supplement to "A unified statistical framework for single cell and bulk RNA sequencing data." (DOI: 10.1214/17-AOAS1110SUPP;.pdf). This supplement provides additional information on the Gibbs sampling and EM algorithm.

1. Introduction

A biological organism is made up of individual cells, which work in concert in tissues to constitute functioning organs. Biologists have long thought that the key to understanding most human diseases lies in understanding the normal and abnormal function of cells. Yet, until very recently, our view of what molecules are expressed and where and when was limited to the level of tissues. Indeed RNA sequencing (RNA-seq) was introduced as a critical tool to answer these questions, but the RNA itself was collected from tissues. This bulk RNA-seq data provides reliable measurements of gene expression levels throughout the genome for bulk samples. With sufficient sequencing depth, even weakly expressed transcripts can be accurately captured by RNA-seq data. This technology has led to breakthroughs in various fields. For example, Fromer et al. (2016) use bulk data, obtained from prefrontal cortex of post-mortem subjects, to gain insight into how genetic risk variation for schizophrenia affects gene expression and likely generates risk for this severe psychiatric disorder.

Still bulk RNA-seq data inevitably ignores the heterogeneity of individual cells because the measurements are summed over the population of cells in the tissue. Yet it is reasonable to predict that diseases like schizophrenia do not arise from malfunctioning brain tissue, per se, but rather certain malfunctioning cells within that tissue. A leading hypothesis is that schizophrenia arises from synaptic dysfunction, and synapses are fundamental to neurons, so should neurons alone be targeted for analyses into schizophrenia? Actually, brain tissue is composed of a remarkably heterogeneous set of cell types, which have vastly different functions and expression profiles. While many are different types of neurons, many others support and alter the function of those neurons and their synapses. Thus, the different gene expression profiles for distinct cell types can have profound functional consequences. These likely are critical for the development of tissues and human diseases, and will be especially important as we aspire to fix such complex diseases as schizophrenia.

It is also of interest to link gene expression with genetic variation, particularly damaging variants associated with risk of disease. Until recently, researchers have assumed that most cells express both copies of a gene equally; however, new findings suggest an even more complex situation motivating single cell measurements. Apparently, some neurons preferentially express the copy of a gene inherited from one parent over the other and this can shape how mutated genes are expressed at the cellular level [Huang et al. (2017a)].

One approach to characterize cell type specific gene expression profiles is to perform deconvolution on bulk RNA-seq data. Consider an observed gene expression matrix $X \in \mathbb{R}^{N \times M}$ for N genes in M bulk samples, each containing K different cell types. The goal of deconvolution is to find two nonnegative matrices $\tilde{A} \in \mathbb{R}^{N \times K}$ and $W \in \mathbb{R}^{K \times M}$, such that

$$X \approx \tilde{A} W, \quad (1.1)$$

where each column of W represents the mixing proportion of K cell types in each bulk sample, and each column of $\tilde{A}$ represents the average gene expression levels in each type of

cells. If the "signature" matrix $\tilde{A}$ is available for a set of "barcode genes" in each cell type, deconvolution reduces to a regression-type problem that aims at estimating W. Several algorithms have been proposed under this setting, including Cibersort [Newman et al. (2015)] and csSAM [Shen-Orr et al. (2010)]. However, without knowing the signature matrix, deconvolution is highly nontrivial, and this line of methods includes the Deconf algorithm [Repsilber et al. (2010)], semi-supervised Nonnegative Matrix Factorization algorithm (ssNMF) Gaujoux and Seoighe (2012) and Digital Sorting Algorithm (DSA) [Zhong et al. (2013)].

A fundamental challenge of the NMF-based methods is the nonuniqueness of the factorization [Donoho and Stodden (2003)]. Therefore, to obtain a biologically meaningful result, both ssNMF [Gaujoux and Seoighe (2012)] and DSA [Zhong et al. (2013)] use a set of "marker genes" to guide the factorization. A marker gene is a gene that only expresses in one cell type. In other words, there are several rows of $\tilde{A}$ that are priorly known to be non-zero at only one column. This is equivalent to the separability assumption introduced by Donoho and Stodden (2003) for the uniqueness of NMF. Unfortunately, marker genes are rarely known in practice. In fact, extracting high-quality marker genes is a challenging step, which is often approached by analyzing purified cells [Abbas et al. (2009)].

On the other hand, single cell RNA sequencing provides gene expression measurements in individual cells, yielding a high-resolution view of cellular states that are uncharacterized in bulk data. Recent advances in high-throughput technologies have made it possible to profile hundreds and thousands of cells [Fan, Fu and Fodor (2015), Kolodziejczyk et al. (2015)]. With several extra pre-processing steps including reverse transcription and amplification, the single cell mRNA library goes through similar sequencing procedures as the bulk samples, and the gene expression levels are measured by the number of mapped reads. With single cell RNA-seq data, one can investigate distinct subpopulations of cells, gain better understanding of the developmental features of different cell types [Grün et al. (2015)], identify cellular differences between healthy and diseased tissues [Kharchenko, Silberstein and Scadden (2014)] and infer gene-regulatory interactions [Padovan-Merhar and Raj (2013)].

The challenges of modeling single cell RNA-seq data come from high cell-to-cell variation, as well as high levels of technical noise during sequencing due to the low amounts of starting mRNAs in individual cells. One important bias comes from the so-called "dropout" events. A dropout happens when a transcript is not detected due to failure of amplification prior to sequencing, leading to a "false" zero in the observed data [Kolodziejczyk et al. (2015)]. Given the excessive amount of zero observations in single cell RNA-seq data, it is critical to distinguish between (i) the dropout genes where transcripts are missed in sequencing and (ii) the "structural zeros" where the genes are truly un-expressed. Modeling the dropout events is especially challenging because of their complicated dependency on gene expression levels and cell characteristics. Specifically, dropouts are more likely to occur in genes expressed at low levels, and certain cells may have systematically higher dropout probabilities than others. In addition to dropout events, other challenges in modeling single cell data include the over-dispersion due to both cellular and technical variation, as

well as high magnitude outliers due to bursts and fluctuations of gene expression levels. We refer the readers to Haque et al. (2017) for a more comprehensive review.

Despite the success of many early single-cell studies, statistical tools that account for the technical noise in single cell RNA-seq data, especially the dropout events, are limited. There have been efforts to analyze single cell data for various purposes. Many methods propose to quantify and account for technical noise using spike-ins [Brennecke et al. (2013), Vallejos, Marioni and Richardson (2015), Vallejos, Richardson and Marioni (2016)]. However, spike-ins are usually unavailable in single cell data due to its expenses in practice. For differential expression analysis, SCDE [Kharchenko, Silberstein and Scadden (2014)] is based on a Bayesian hypothesis testing procedure using a three-component mixture model to capture technical noise; subsequently, MAST [Finak et al. (2015)] uses a hurdle model that can adjust for various covariates; more recently, Vu et al. (2016) construct a beta-poisson mixture model, integrated within a generalized linear model framework. Various relevant problems have also been studied, including inferring the spatial localization of single cells in complex tissues [Satija et al. (2015)], dimension reduction using Zero-Inflated Factor Analysis (ZIFA) [Pierson and Yau (2015)], and clustering unlabeled single cells while accounting for technical variation [Prabhakaran, Azizi and Pe'er (2016)]. All of these aforementioned methods have been successfully applied to different single cell data sets. However, analytical methods that aim at the fundamental problem of imputing dropout genes and estimating the cell-type-specific gene expression profiles remain underdeveloped.

In this paper, we propose to jointly analyze single cell and bulk RNA-seq data using the Unified RNA-Sequencing Model (URSM), which simultaneously corrects for the dropout events in single cell data and performs deconvolution in bulk data. We point out that URSM only requires consistent cell types between both data sources, preferably measured on the same tissue from subjects with similar ages. It does not require the single cell and bulk data being measured on the same subjects, nor does it assume the same proportions of cell types in both data sets. Given a single cell data set, usually there are existing bulk data measured on the same tissue that can be modeled jointly using URSM. For example, BrainSpan provides extensive gene expression data on adult and developing human brains [Sunkin et al. (2013)], and GTex establishes a human RNA-seq gene expression database across 43 tissues [GTEx Consortium (2013)].

By integrating single cell and bulk RNA-seq data, URSM borrows the strength from both data sources, and is able to (i) obtain reliable estimation of cell type specific gene expression profiles, (ii) infer the dropout entries in single cell data and (iii) infer the mixing proportions of different cell types in bulk samples. Our framework explicitly models the dropout events in single cell data, and captures the relationship between dropout probability and expected gene expression levels. By involving high-quality bulk data, URSM achieves more accurate estimation of cellular expression profiles than using only single cell data. By incorporating the single cell data, URSM provides, for the first time, deconvolution of the bulk samples without going through the error-prone procedure of estimating marker genes. To the best of our knowledge, this is the first model that jointly analyzes these two types of RNA-seq data. We will illustrate in simulation (Section 4) and real-world data (Section 5) that URSM

successfully corrects for the dropouts in single cell data, and provides reliable deconvolution for bulk samples.

2. A unified statistical model

Suppose RNA-sequencing is conducted on N genes and K types of cells are of interest. Then bulk and single cell RNA-seq data can be linked together by a common profile matrix $A \in \mathbb{R}^{N\times K}$, where the kth column $A_{\cdot k}$ represents the expected *relative* expression levels of N genes in the kth type of cells, such that each column sums to one. Note that by considering the *relative* expression levels, the profile matrix A does not depend on sequencing depths, and thus remains the same in both data sources. The two data sources provide two different views on the profile matrix A. In single cell data, the observations are independent realizations of different columns of A with extra noise due to dropout events. In bulk data, the expected relative expression levels for a mixture sample are weighted sums of columns of A, where the weights correspond to mixing proportions of different cell types. Here, we propose URSM to analyze the bulk and single cell RNA-seq data together, which borrows the strength from both data sets and achieves more accurate estimation on the profile matrix. This further enhances the performance of deconvolving bulk samples, as well as inferring and imputing the dropout genes in single cells.

The plate model of URSM for generating single cell and bulk RNA-seq data is given in Figure 1. Specifically, for single cell data, let $Y \in \mathbb{R}^{N\times L}$ represent the measured expression levels of N genes in L single cells, where the entries are RNA-seq counts. To model the dropout events, we introduce the binary observability variable $S \in \{0, 1\}^{N\times L}$, where $S_{il} = 0$ if gene i in cell l is dropped out, and $S_{il} = 1$ if it is properly amplified. For each cell l, let $G_l \in \{1, \ldots, K\}$ denote its type, then the vector of gene expression $Y_{\cdot l} \in \mathbb{R}^N$ is assumed to follow a Multinomial distribution with probability vector $p_{\cdot l}$ and the sequencing depth $R_l = \sum_{i=1}^{N} Y_{il}$ is the number of trials. Without dropout events, $p_{\cdot l}$ would be the corresponding column of the profile matrix, $A_{\cdot G_l}$, which is the true relative expression levels for cell type G_l. With the existence of dropouts, $p_{\cdot l}$ becomes the element-wise product of $A_{\cdot G_l}$ and $S_{\cdot l}$, which is then normalized to sum to one. To capture the dependency between dropout probabilities and gene expression levels, the observation probability $\pi_{il} = \mathbb{P}(S_{il} = 1)$ is modeled as a logistic function of A_{i, G_l},

$$\pi_{il} = \text{logistic}(\kappa_l + \tau_l A_{i, G_l}), \quad (2.1)$$

so that lowly expressed genes have high probabilities of being dropped out, where the coefficients (κ_l, τ_l) are cell-dependent that capture the cellular heterogeneity. Under this model, the set of dropout entries and structural zeros are defined as

$$\text{dropouts} = \{(i,l): S_{il} = 0\}, \qquad (2.2)$$
$$\text{structural zeros} = \{(i,l): S_{il} = 1, Y_{il} = 0\}.$$

For bulk data, let $X \in \mathbb{R}^{N \times M}$ represent the RNA-seq counts of N genes in M bulk samples. For the j th bulk sample, let $W_{\cdot j} \in \mathbb{R}^K$ denote the mixing proportions of K cell types in the sample, satisfying $\sum_{k=1}^{K} W_{kj} = 1$. Then the gene expression vector $X_{\cdot j} \in \mathbb{R}^N$ is assumed to also follow a Multinomial distribution, where the probability vector is the weighted sum of K columns of A with the weights being $W_{\cdot j}$, and the number of trials is the sequencing depth for sample j, defined as $R_j = \sum_{i=1}^{N} X_{ij}$.

For the hierarchical model setting, we assign the conjugate Dirichlet prior for the mixing proportions $W_{\cdot j}$, and Gaussian priors for the cell-dependent dropout parameters (κ_l, τ_l). Here, we adopt an empirical Bayes' approach, where the parameters are estimated by maximum-likelihood-estimations (MLE) using the expectation-maximization (EM) algorithm. Using this framework, our goal is threefold: (i) learn the profile matrix A as part of the model parameters, which characterizes the cellular gene expression profiles, (ii) make posterior inference on the dropout status S for single cell data, which can be used to identify dropout entries and (iii) make posterior inference on the mixing proportions W in bulk samples. Finally, the inferred dropout entries in single cell data can be imputed by their expected values using the estimated A and sequencing depths R_l.

Full model specification

- Bulk data
  - $W_{\cdot j} \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\alpha)$ for $j = 1, \ldots, M$, where $\alpha \in \mathbb{R}^K$, $\alpha \geq 0$.
  - $X_{\cdot j} \mid W_{\cdot j} \overset{\text{indep.}}{\sim} \text{Multinomial}(R_j, AW_{\cdot j})$ for $j = 1, \ldots, M$, where $R_j = \sum_{i=1}^{N} X_{ij}$.
- Single cell data
  - $\kappa_l \overset{\text{i.i.d.}}{\sim} N(\mu_\kappa, \sigma_\kappa^2), \tau_l \overset{\text{i.i.d.}}{\sim} N(\mu_\tau, \sigma_\tau^2)$ for $l = 1, \ldots, L$.
  - $\pi_{il} = \text{logistic}(\kappa_l + \tau_l A_{i, G_l})$, where $G_l \in \{1, \ldots, K\}$ is the type of the lth cell.
  - $S_{il} \mid \kappa_l, \tau_l \overset{\text{indep.}}{\sim} \text{Bernoulli}(\pi_{il})$ for $i = 1, \ldots, N$; $l = 1, \ldots, L$.
  - $Y_{\cdot l} \mid S_{\cdot l}, \overset{\text{indep.}}{\sim} \text{Multinomial}(R_l, p_l)$ for $l = 1, \ldots, L$, where $R_l = \sum_{i=1}^{N} Y_{il}$,

$$p_l = (p_{il})_{i=1,\ldots,N}, \quad \text{where } p_{il} = \frac{A_{i, G_l} S_{il}}{\sum_{n=1}^{N} A_{n, G_l} S_{nl}}.$$

Remark 1—We assume all entries in A to be strictly positive. In principle, one can allow some entries A_{ik} to be exactly zero, but this will lead to a degenerate multinomial distribution and complicate the likelihood function. In addition, making inference on S_{il} when $A_{i,\,G_l} = 0$ is an ill-defined problem. If $A_{ik} = 0$, then we will have $X_{il} = 0$ for all type-k cells, but such structure rarely appears in real data. In practice, it is usually helpful to use some small positive numbers rather than exact zeros to capture the background signal in sequencing processes [Kharchenko, Silberstein and Scadden (2014)].

Remark 2—It is straightforward to use one part of URSM when only one data source is available. In Section 4, we will show the performance of the submodel for single cell data. It is also possible to use the submodel for bulk data when only bulk data are available, but extra information about marker genes needs to be incorporated in this scenario to avoid the non-identifiability issue, as explained in Section 1.

3. Inference and estimation: EM algorithm

This section presents an expectation-maximization (EM) algorithm [Dempster, Laird and Rubin (1977)] for fitting the maximum likelihood estimation (MLE) of the parameters $\theta = (A, \alpha, \mu_\kappa, \sigma_\kappa^2, \mu_\tau, \sigma_\tau^2)$, as well as a Gibbs sampling algorithm for posterior inference on latent variables $H = (W, S, \kappa, \tau)$. As illustrated in Section 2, the key values of scientific interests include (i) an estimate of the profile matrix A that characterizes the cellular gene expression profiles, (ii) $\mathbb{E}[S|Y, \theta]$, the inferred dropout probability at each entry in single cell data and (iii) $\mathbb{E}[W|X, \theta]$, the inferred mixture proportion of bulk samples.

The main difficulty of handling our model is the intractable posterior distributions due to nonconjugacy. Therefore, approximate inference needs to be performed. One of the main methods for approximate inference in Bayesian modeling is Monte Carlo Markov Chain (MCMC) sampling [Gelfand and Smith (1990)], where a Markov chain on latent variables is constructed, with stationary distribution being the true posterior. After obtaining a long enough chain, the posterior can be approximated with empirical estimation. Gibbs sampling [Casella and George (1992), Geman and Geman (1984)] is one of the most widely used forms of MCMC algorithms given its simplicity and efficiency. On the other hand, variational methods form an alternative line for approximate inference, where the posterior is approximated analytically by a family of tractable distributions [Blei, Kucukelbir and McAuliffe (2017), Jordan et al. (1999), Wainwright and Jordan (2008)]. While being computationally scalable in many large-scale problems, variational methods are inherently less accurate due to the inevitable gap between the variational distributions and the true posterior distribution.

In this paper, we present a Gibbs sampling algorithm for approximate inference on latent variables using the data augmentation trick. This algorithm can also be used in the E-step of the EM procedure, leading to a Gibbs-EM (GEM) algorithm for obtaining MLEs of model parameters [Dupuy and Bach (2016)]. The specific steps are outlined in Section 3.1 and Section 3.2, and more details can be found in the supplement [Zhu et al. (2018)]. Finally, we point out that one can also proceed with variational inference, but due to space limitation, we do not pursue this approach in detail.

3.1. E-step: Gibbs sampling

The latent variables for bulk data and single cell data are conditionally independent given observed data X, Y and parameters. Therefore, Gibbs sampling can be performed on the two data sources in parallel. In this section, we describe the sampling procedure for the two parts separately.

Bulk data—To obtain the posterior inference of W (the mixing proportions) in bulk data, we rewrite the model to be mixture of multinomials by introducing the augmented latent variables Z and d as follows:

$$W_{\cdot j} \overset{\text{i.i.d.}}{\sim} \text{Dirichlet}(\alpha), \quad j = 1, \ldots, M, \tag{3.1}$$

$$Z_{rj} \overset{\text{i.i.d.}}{\sim} \text{Multinomial}(1, W_{\cdot j}), \quad r = 1, \ldots, R_j,$$

$$d_{rj} \overset{\text{indep.}}{\sim} \text{Multinomial}(1, A_{\cdot Z_{rj}}), \quad r = 1, \ldots, R_j,$$

$$X_{ij} = \sum_{r=1}^{R_j} I_{\{d_{rj} = i\}}, \quad i = 1, \ldots, N, j = 1, \ldots, M.$$

Note that this model is closely related to the Latent Dirichlet Allocation (LDA) model [Blei, Ng and Jordan (2003)] in topic modeling, if we view a gene as a word, a cell type as a topic and a bulk sample as a document. Although the Gibbs sampling algorithm has been developed for LDA in Griffiths and Steyvers (2004), there are two difficulties that prevent us from directly applying this algorithm to our model. First, the LDA model assumes observations of d_{rj}, which are the actual words in an document, but in RNA-seq data, only the final counts X_{ij} are observed. Second, the sequencing depths R_j's are typically large in real data, so it will be extremely computationally demanding to keep track of Z_{rj} and d_{rj}. Therefore, we propose a modified algorithm by defining another set of augmented latent variables

$$\widetilde{Z}_{ij,k} := \sum_{r: d_{rj} = i} I_{\left\{Z_{rj} = k\right\}} \quad \text{and} \quad \widetilde{Z}_{ij} := (\widetilde{Z}_{ij,k}) \in \mathbb{R}^K, \tag{3.2}$$

and it can be shown that

$$W_{\cdot j} \mid W_{\cdot(-j)}, \widetilde{Z}, X \sim \text{Dirichlet}\left(\alpha + \sum_{i=1}^{N} \widetilde{Z}_{ij}\right), \qquad (3.3)$$
$$\widetilde{Z}_{ij} \mid \widetilde{Z}_{(-ij)}, W, X \sim \text{Multinomial}\left(X_{ij}, \frac{A_{i\cdot} \odot W_{\cdot j}}{\sum_{k=1}^{K} A_{ik} W_{kj}}\right),$$

where $\odot$ denotes element-wise multiplication, and the index $(-i)$ denotes everything else other than i.

Single cell data—As for posterior inference of S, κ, τ in single cell data, note that the first part of the model can be rewritten as

$$(\kappa_l, \tau_l) \sim N(\mu, \Sigma), \quad \text{where } \mu = (\mu_\kappa, \mu_\tau), \Sigma = \text{Diag}(\sigma_\kappa^2, \sigma_\tau^2), \qquad (3.4)$$
$$S_{il} \mid \kappa_l, \tau_l \sim \text{Bernoulli}(\text{logistic}(\psi_{il})), \quad \text{where } \psi_{il} = \kappa_l + \tau_l A_{i, G_l},$$

which has the same form as a Bayesian logistic regression, with covariates being $(1, A_{i, G_l})$. Therefore, following the recent development of Gibbs' sampling technique in this area [Polson, Scott and Windle (2013)], we introduce a set of augmented latent variables ω, and the conditional complete posteriors can be shown to be

$$\omega_{il} \mid \omega_{(-il)}, S, Y, \kappa, \tau \sim \text{Polya-Gamma}(1, \psi_{il}), \qquad (3.5)$$
$$(\kappa_l, \tau_l) \mid \kappa_{(-l)}, \tau_{(-l)}, \omega, S, Y \sim N(m_{\omega l}, V_{\omega l}^{-1}),$$
$$S_{il} \mid S_{(-il)}, \omega, S, \kappa, \tau, Y \sim \text{Bernoulli}(b_{il}),$$

where

$$\psi_{il} = \kappa_l + \tau_l A_{i,G_l},$$

$$V_{\omega l} = \begin{pmatrix} \sum_{i=1}^{N} \omega_{il} + \sigma_\kappa^{-2} & \sum_{i=1}^{N} \omega_{il} A_{i,G_l} \\ \sum_{i=1}^{N} \omega_{il} A_{i,G_l} & \sum_{i=1}^{N} \omega_{il} A_{i,G_l}^2 + \sigma_\tau^{-2} \end{pmatrix},$$

$$m_{\omega l} = V_{\omega l}^{-1} \begin{pmatrix} \sum_{i=1}^{N} S_{il} - N/2 + \mu_\kappa/\sigma_\kappa^2 \\ \sum_{i=1}^{N} S_{il} A_{i,G_l} - 1/2 + \mu_\tau/\sigma_\tau^2 \end{pmatrix},$$

$$b_{il} = \begin{cases} 1, & \text{if } Y_{il} > 0, \\ \text{logit}\left(\psi_{il} + R_l \log\left(\frac{\sum_{n \neq i} A_{n,G_l} S_{nl}}{A_{i,G_l} + \sum_{n \neq i} A_{n,G_l} S_{nl}}\right)\right), & \text{if } Y_{il} > 0. \end{cases}$$

3.2. M-step

In the M-step of GEM algorithm, the parameters are updated to maximize a lower bound on the expected complete log likelihood function, or the so-called Evidence Lower BOund (ELBO), where the posterior expectation $\mathbb{E}_Q$ is estimated using Gibbs samples obtained in the E-step. The optimal dropout parameters $(\mu_\kappa, \sigma_\kappa^2, \mu_\tau, \sigma_\tau^2)$ have the following closed forms:

$$\begin{aligned} \hat{\mu}_\kappa &= \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_Q(\kappa_l), \ \hat{\sigma}_\kappa^2 = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_Q[(\kappa_l - \hat{\mu}_\kappa)^2], \\ \hat{\mu}_\tau &= \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_Q(\tau_l), \ \hat{\sigma}_\tau^2 = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}_Q[(\tau_l - \hat{\mu}_\tau)^2]. \end{aligned} \tag{3.6}$$

For A and α, there are no closed form solutions, and we use the projected gradient ascent algorithm:

$$A_{\cdot k}^{\text{new}} \leftarrow \text{Proj}(A_{\cdot k}^{\text{old}} + t \cdot \nabla \text{ELBO}(A_{\cdot k}^{\text{old}})), \ \alpha^{\text{new}} \leftarrow \text{Proj}(\alpha^{\text{old}} + t \cdot \nabla \text{ELBO}(\alpha^{\text{old}})), \tag{3.7}$$

where the step size t is determined by backtracking line search, and the *Proj* function is the projection onto the feasible set:

$$A_{ik} \geq \varepsilon_A, \quad \sum_{i=1}^{N} A_{ik} = 1, \quad \alpha_k \geq \varepsilon_\alpha, \tag{3.8}$$

where ε_A, $\varepsilon_a > 0$ are some small predetermined constants. The gradients are computed as

$$\frac{\partial \text{ELBO}}{\partial A_{ik}} = \sum_{j=1}^{M} \frac{\mathbb{E}_Q[\widetilde{Z}_{ij,k}]}{A_{ik}} + \sum_{l: G_l = k} \left[\frac{Y_{il}\mathbb{E}_Q(S_{il})}{A_{ik}} - \mathbb{E}_Q[\omega_{il}\tau_l^2]A_{ik} - \frac{\mathbb{E}_Q(S_{il})R_l}{u_l} \right. \tag{3.9}$$

$$\left. + \mathbb{E}_Q\left[\left(S_{il} - \frac{1}{2}\right)\tau_l - \omega_{il}\tau_l\kappa_l\right]\right], \quad \frac{\partial \text{ELBO}}{\partial \alpha_k} = \sum_{j=1}^{M} \mathbb{E}_Q[\log W_{kj}] + M\left[\Psi\left(\sum_{k=1}^{K}\alpha_k\right) - \Psi(\alpha_k)\right],$$

where $u_l = \sum_{i=1}^{N} A_{i,G_l}\mathbb{E}_Q(S_{il})$ and $\Psi(\cdot)$ is the digamma function. More detailed derivations can be found in the supplement.

4. Simulation results

In this section, we evaluate the performance of URSM in synthetic datasets. We focus on the accuracy of recovering the profile matrix A and mixing proportions W, as well as the ability of distinguishing between dropout entries and structural zeros using the posterior inference of S.

4.1. Settings

Let N be the number of genes. The sequencing depths for bulk samples are independently generated from Poisson($50N$). To account for the fact that the sequencing depths of single cell data are usually much lower and highly variable, they are generated from the Negative Binomial with mean $2N$ and dispersion parameter 2.

The cell-type specific profile matrix A is generated as follows: (i) simulate all entries independently from log-normal with $\mu = 0$, $\sigma = 1$, (ii) for each cell type k, let $N_m = 10$ genes be marker genes, that is, set $A_{il} = 0$ for $l \neq k$, (iii) for each cell type k, let $N_a = 10$ genes be anti-marker genes, that is, set $A_{ik} = 0$, (iv) let another set of $N_h = 30$ genes be housekeeping genes that have the same expression levels in all cell types and (v) finally, properly normalize A so that each column sums to 1. Specifically, in each column, we normalize the N_h housekeeping genes such that they sum to N_h/N, and the remaining genes sum to $1 - N_h/N$.

Finally, the observation status $\{S_{il}\}_{il}$ for each gene i in each single cell l is simulated independently from Bernoulli(π_{il}). Recall that $S_{il} = 0$ indicates a dropout, and the dropout probability is determined by

$$1 - \pi_{il} = 1 - \text{logistic}(\kappa_l + \tau_l A_{i,G_l}), \tag{4.1}$$

where $G_l \in \{1, \ldots, K\}$ is the type of cell l. In the following sections, κ_l's are independently generated from Normal(-1, 0.5^2), and τ_l's are independently generated from Normal($1.5N$, $(0.15N)^2$). Note that by construction, the mean of each column of A, $\bar{A}_{\cdot,k}$, is always $1/N$.

Therefore, $\mathbb{E}[\kappa_l + \tau_l \bar{A}_{\cdot, G_l}] = 0.5$ for each cell, which corresponds to an average dropout probability of 37.8%, and the maximal dropout probability is 73.1% when $A_{ik} = 0$.

4.2. Estimation of profile matrix

In this section, we illustrate that URSM provides accurate estimation on the profile matrix A after correcting for dropouts and utilizing bulk samples. Following the simulation settings specified in Section 4.1, we generate $L = 100$ single cells and $M = 150$ bulk samples on $N = 200$ genes. We consider $K = 3$ cell types. For single cells, 30%, 30% and 40% of the cells are assigned to the 3 different types, respectively. For bulk samples, the hyperparameter of the mixing proportions is set to $\alpha = (1, 2, 3)$. The dropout probability curves, simulated following equation (4.1), are shown in Figure 2(a). The simulated single cell data has 64.6% entries being zero.

A naive method to estimate the profile matrix A is to use the sample means of single cell expression levels, after normalizing by their sequencing depths. Specifically, recall that $Y \in \mathbb{R}^{N \times L}$ represents the observed expression levels in single cells, $\{G_l\}_{l=1,\ldots,L}$ represent the cell types and $\{R_l\}_{l=1,\ldots,L}$ are the sequencing depths, defined as $R_l = \Sigma_i Y_{il}$. Then an entry A_{ik} can be estimated by

$$\hat{A}_{ik}^{\text{naive}} = \frac{1}{\#\{l : G_l = k\}} \sum_{l : G_l = k} \frac{Y_{il}}{R_l}. \quad (4.2)$$

However, due to the presence of dropout events and the dependency between π_{il} and A, this naive sample mean estimation is biased, with L_1 loss 0.81 [Figure 2(b)], where the L_1 loss is computed as $\Sigma_{i,k} |\hat{A}_{ik} - A_{ik}|$. On the other hand, by explicitly modeling the occurrence of dropout events and capturing the relationship between dropout probability and expected expression level, a submodel of URSM that only uses single cell data successfully corrects for the bias, and substantially reduces the loss to 0.27 [Figure 2(c)]. Finally, by integrating the bulk data, URSM further improves the estimation and further reduces the L_1 loss to 0.17 [Figure 2(d)].

4.3. Deconvolution of bulk samples

Now we further examine the model performance on inferring the mixing proportions W in bulk samples, using the same simulation setting as in Section 4.2. We compare the performance of URSM to three widely used deconvolution methods: Digital Sorting Algorithm (DSA) [Zhong et al. (2013)], semi-supervised Nonnegative Matrix Factorization (ssNMF) [Gaujoux and Seoighe (2012)] and Cibersort [Newman et al. (2015)].

Both DSA and ssNMF rely heavily on a set of *given* marker genes as input to guide the matrix factorization, where a “marker gene” is only expected to express in one cell type. Unfortunately, marker genes are rarely known in practice, and a widely adopted procedure is to estimate the list of marker genes from purified cells by selecting those with the most different expression levels across cell types. Here, we mimic this procedure by estimating a

list of marker genes from single cell data to guide DSA and ssNMF. Specifically, we adopt the method in Abbas et al. (2009), which calculates a p-value of each gene by comparing its expression level in the highest and second-highest types of cells, then selects the group of genes with the smallest p-values. Figure 3 shows the L_1 loss of estimating A and W using DSA and ssNMF with different sets of estimated marker genes with p-values smaller than $\{10^{-8}, \ldots, 10^{-3}\}$, and the number of selected marker genes is listed in Table 1. It is clear that these two algorithms are sensitive to the input marker genes. For comparison, we also evaluate the performances of DSA and ssNMF when the oracle information of true marker genes is available.

On the other hand, Cibersort requires a "signature" matrix containing the expression levels of a group of "barcode" genes that collectively distinguish between different cell types. Note that this essentially requires knowing part of the profile matrix A, which contains much more information than the marker gene list. Here, we use the estimated $\hat{A}$ from our unified model as the signature matrix for Cibersort. We report the L_1 loss of estimating W when Cibersort only takes the expression levels of the selected marker genes, as well as when Cibersort uses the entire $\hat{A}$. Figure 3(b) suggests that Cibersort prefers larger number of barcode genes as input.

Finally, URSM automatically utilizes the information in single cell data to guide deconvolution. Figure 3 illustrates that URSM and Cibersort usually outperform DSA and ssNMF using estimated marker genes, and achieve comparable L_1 loss even when DSA and ssNMF have the oracle information of marker genes.

4.4. Inference of dropout entries in single cell data

Next, we present the inference on dropout entries in single cell data, again using the same setting as in Section 4.2. Here, our goal is to distinguish between dropout entries and structural zeros, as defined in equation (2.2). Note that we only need to make inference for locations where the observed expression levels are zero, that is, on the set $\{(i, l) : Y_{il} = 0\}$. Recall that $S_{il} = 0$ if gene i is dropped out in cell l, and our model provides the estimated posterior mean of S:

$$\tilde{\pi}_{il} = \mathbb{E}(S_{il} \mid X, Y, \theta), \quad (4.3)$$

where θ denotes the model parameters. Hence a natural approach is to predict the entries with small $\tilde{\pi}_{il}$ to be dropouts.

A potential competitor for imputing dropout entries is the Nonnegative Matrix Factorization (NMF) [Lee and Seung (2000)]. One can construct a low-rank approximation to the single cell expression matrix $Y \in \mathbb{R}^{N \times L}$ using NMF. Intuitively, the approximated values tend to be higher at dropout entries, and closer to zero at structural-zero entries. As shown in Figure 4(a), if the rank is properly specified, this simple NMF-based method demonstrates certain ability to distinguish between dropout genes and structural zeros, but not as well as URSM. In addition, in order to further impute the dropout entries, a good estimation of the profile matrix A is also needed. Figure 4(b) shows the estimation of A by taking sample average as

in equation (4.2), with Y substituted by the NMF approximation. It is clear that the NMF approach fails to correct for the bias introduced by the dropout events, while URSM succeeds in both identifying dropout entries and obtaining an unbiased estimation of A [recall Figure 2(d)].

4.5. Robustness

Finally, we demonstrate the robustness of our model. We apply URSM under the scenario where the number of cell types in single cell data K_{sc} is not equal to the number of cell types in bulk data K_{bk}, as well as when the number of genes N grows. URSM always takes $K = \max\{K_{sc}, K_{bk}\}$ as input, and estimates $\hat{A}_{unif} \in \mathbb{R}^{N\times K}$ and $\hat{W}_{unif} \in \mathbb{R}^{K\times M}$. When $K_{sc} > K_{bk}$, it is straightforward to directly apply URSM, and ideally the estimated $\hat{W}_{unif}$ will assign zero proportions to the missing cell types in bulk samples. However, when $K_{sc} < K_{bk}$, without extra information, deconvolution is an ill-defined problem because of the nonidentifiability issue (see Section 1 for more details). In order to find a biological meaningful solution, we follow the idea in ssNMF [Gaujoux and Seoighe (2012)] and use a set of marker genes to initialize the parameters for the EM algorithm by setting the corresponding entries in A to be zero. We consider the scenario where for each cell type, 5 true marker genes and 3 imperfect marker genes are used for initialization. The imperfect marker genes are selected from the nonmarker genes, where we pick the ones with the largest difference between the highest and second highest expression levels across cell types in A.

Following Section 4.1, we simulate $M = 150$ bulk samples, where the mixing proportions in bulk samples are generated from Dir(α) with $\alpha = (1, \ldots, K_{bk})$. For single cell data, we generate 40 cells in the majority cell type, and 30 cells in each of the remaining $K_{sc} - 1$ types. To reduce the computation load and enhance stability, we use the maximum a posteriori estimation for W in the E-step for bulk samples. More details are included in the supplement.

Again, we compare URSM to DSA, ssNMF and Cibersort. Both DSA and ss-NMF require a set of marker genes as input, and we report their performances under two scenarios: (i) the oracle scenario where 5 true marker genes are provided for each cell type and (ii) a more realistic scenario as used by our uniform model, where 5 true marker genes and 3 imperfect marker genes are provided for each cell type. Note that when $K_{sc} > K_{bk}$, bulk samples contain no information of the expression patterns for the missing cell types, so we allow DSA and ssNMF to only deconvolve K_{bk} cell types in these cases. We point out that this strategy favors the DSA and ssNMF methods by providing them extra information of the missing cell types in bulk samples. For Cibersort, as in the previous sections, we use the estimated profile matrix obtained from our uniform model as the input signature matrix.

Figure 5(a) summarizes the performance of different models under various choices of K_{sc} and K_{bk} when $N = 200$ in 10 repetitions. In order to make a comparable comparison across different K's, we report the *average per cell type* L_1 loss, that is, the average L_1 loss $\|\hat{A}_{\cdot,k} - A_{\cdot,k}\|_1$ and $\|\hat{W}_{\cdot,k} - W_{\cdot,k}\|_1$ across all columns k. We see that the performance of URSM remains robust under different settings, and is usually comparable to DSA and ssNMF algorithms even when the latter two algorithms have the oracle marker gene information.

Not surprisingly, Cibersort has similar performance as URSM because it uses our estimated $\hat{A}_{\text{unif}}$ as input. We point out that when the sample mean estimation $\hat{A}_{\text{naive}}$ [equation (4.2)] is given to Cibersort as the signature matrix, the performance is unstable and it cannot provide deconvolution when $K_{\text{sc}} < K_{\text{bk}}$. Finally, we also demonstrate the performance of different models when $N = \{200, 500, 1000\}$, where we set $K_{\text{sc}} = K_{\text{bk}} = 3$. Figure 5(b) verifies that URSM remains robust with larger numbers of genes.

5. Application to fetal brain data

5.1. Data preprocessing

In this section, we apply URSM to gene expression measured on fetal brains. The single cell RNA-seq data comes from Camp et al. (2015), where 226 cells from fetal brains are sequenced on 18,927 genes. The authors have removed endothelial cells and interneurons, and the remaining 220 single cells are labeled into three types: 40 apical progenitors (APs), 19 basal progenitors (BPs) and 161 neurons (Ns). In addition, the authors have normalized the RNA-seq counts using FPKM (Fragments Per Kilobase of exon per Million fragments mapped) and performed log-transformation by $\log_2(x + 1)$. We refer the readers to Camp et al. (2015) for more details of the single cell data preprocessing. On the other hand, microarray bulk gene expression data on fetal brains is provided by the BrainSpan atlas Kang et al. (2011). Within the same window of development, 12 to 13 post-conception week, 72 bulk samples from prefrontal cortex are measured on 16,947 genes. To apply our model, the single cell RNA-seq data are transformed back to linear scale by $2^x - 1$, and all measurements are truncated to integers. To approximate the RNA-seq counts in bulk samples, we transform the BrainSpan microarray data in the same way and treat them as pseudo-RNA-seq counts. The resulting bulk samples have an average pseudo sequencing depth of 5.5×10^6, which is 26 times larger than the average effective sequencing depth in single cells, 2.1×10^5, where the effective sequencing depth is calculated as the sum of FPKM across all genes in each single cell.

To reduce computational load, we only focus on genes with significantly different expression levels among the three cell types. Specifically, we use the 315 so-called PC genes proposed in Camp et al. (2015), which have the largest loadings in a Principal Component Analysis (PCA) and account for the majority of cellular variation. After restricting to the overlapping genes that are also sequenced in BrainSpan bulk samples, a final list of 273 PC genes are obtained and used in the following analyses. When restricting to these 273 genes, the average effective sequencing depth (i.e., the sum of RNA-seq counts in each sample) is 3.2×10^5($sd = 1.6 \times 10^4$) in BrainSpan tissues, and 1.4×10^4($sd = 4.3 \times 10^3$) in single cells.

Due to the nature of active cell development from APs and BPs to Neurons in fetal brains, we expect to have a few cells that are actively transitioning between two cell types, whose labels are ambiguous. We first remove these ambiguously labeled cells from our analysis. Specifically, we project the single cells to the leading 2-dimensional principal subspace, where the pseudo developing time is constructed using the Monocle algorithm [Trapnell et al. (2014)]. Based on the results, the 3 BPs that are close to AP or Neuron clusters are removed, so are the 4 Neurons that are close to AP or BP clusters [Figure 6(b)]. The

remaining 213 single cells are retained for analysis, and their gene expression levels on the 273 PC genes are visualized in Figure 6(a).

5.2. Imputation of single cell data

Here, we apply URSM to identify and impute the dropout entries in single cell data. Note that in order to distinguish between dropout entries and structural zeros in single cell data [equation (2.2)], we only need to focus on the entries where the observed gene expression levels are zero. The inference of dropout entries is based on the estimated posterior expectation of $\mathbb{E}(S_{il}|X, Y, \Theta)$. As a result, among the 37,771 zero-observation entries, 45.7% are inferred to be dropouts with probability one [Figure 6(c)]. These entries are then imputed by their expected values, calculated using the corresponding entries in the estimated profile matrix A multiplied by the sequencing depths of the corresponding cells. To illustrate the impact of imputation, we apply PCA again on the imputed data. Figure 6(d) visualizes the cells in the first two principal components, and the clusters for different cell types are more clearly separated.

5.3. Deconvolution of bulk samples

Finally, we present the deconvolution results of bulk samples using URSM. According to the prior knowledge that the proportions in bulk samples should be roughly consistent with that in single cell data, the mixing parameter α is initialized at $(2 \times 10^4, 10^4, 7 \times 10^4)$ for AP, BP and Neurons. The scale of α is chosen to be comparable to the average effective sequencing depths of 1.4×10^4 among all single cells. Figure 7(a) shows the inferred mixing proportions of APs, BPs and Neurons in each of the 72 bulk samples, with an average of 17.7% AP cells, 8.7% BP cells and 73.6% Neurons.

For comparison, we also apply the Digital Sorting Algorithm (DSA) [Zhong et al. (2013)], semi-supervised Nonnegative Matrix Factorization (ssNMF) [Gaujoux and Seoighe (2012)] and Cibersort [Newman et al. (2015)] on the BrainSpan bulk samples. The marker genes for DSA and ssNMF are selected by comparing each gene's expression level in the highest and second-highest types of cells in the single cell data, and genes with p-value $< 10^{-5}$ are treated as markers [Abbas et al. (2009)]. This procedure leads to 21 AP markers, 6 BP markers and 28 Neuron markers, which serve as input to DSA and ssNMF. For Cibersort, the input signature matrix is provided by the estimated $\hat{A}$ from URSM. Figure 7(b)–(d) suggest that the proportions estimated by ssNMF tend to have too large variations, while DSA overestimates the neural composition, and Cibersort obtains similar results as URSM.

As another perspective to verify the deconvolution results, we use the intuition that the true proportions of a cell type should be correlated with the expression levels of its marker genes in bulk samples. To check whether this holds in the results, we first normalize each bulk sample by their effective sequencing depths, such that the normalized expressions sum to one in each sample. We focus on 7 genes based on biological knowledge, including the radial glia (RG) markers *PAX6* and *GLI3* that are expected to only express in AP and BP cells, the RG marker *HES1* that is mostly expressed in AP cells, the early BP marker *HES6*, as well as neuronal genes *NEUROD6*, *BCL11B* and *MYT1L* [Camp et al. (2015)]. Table 2 summarizes the correlations calculated by estimated proportions using different methods,

and we see that URSM and Cibersort usually achieve the highest correlations. Finally, we point out that if Cibersort uses the naive sample mean estimation from single cell data as the signature matrix, it will fail to identify BP cells and achieve much lower correlations.

6. Discussion

In this paper, we propose URSM, a unified framework to jointly analyze two types of RNA-seq data: the single cell data and the bulk data. URSM utilizes the strengths from both data sources, provides a more accurate estimation of cell type specific gene expression profiles, and successfully corrects for the technical noise of dropout events in single cell data. As a side product, URSM also achieves deconvolution of bulk data by automatically incorporating the cellular gene expression patterns.

Dropouts present one of the biggest challenges to modeling scRNA-seq data. URSM assumes a dependency between expression level and the probability of observing dropout and aims, probabilistically, to infer which observations are likely dropouts. There are a number of alternative approaches in the literature; for a discussion see Huang et al. (2017b) and Vallejos et al. (2017). The most common statistical approach is to explicitly model the zero-inflation process, for example, SCDE [Kharchenko, Silberstein and Scadden (2014)], MAST [Finak et al. (2015)] and ZIFA [Pierson and Yau (2015)]. Some methods assess the fraction of dropouts per gene, other methods, such as CIDR [Lin, Troup and Ho (2017)], take this process to the next step by imputing the dropout values. SAVER [Huang et al. (2017b)] avoids trying to determine which observations are dropouts and aims to impute any poorly measured value using the gene-to-gene correlation pattern, and other features in the cell-type specific samples.

We apply URSM to two gene expression data sets from fetal brains, and obtain promising results on imputing single cell RNA-seq data and deconvolving bulk samples. With more upcoming single cell data on fetal brains, it would be of great scientific interest to apply URSM to specimen from different brain developing periods, which will aid our understanding on gene expression patterns during early brain development and their impact on many complex human disorders. In practice, the degrees of heterogeneity can vary for different tissues. For example, liver tissues may contain more homogeneous cell types. In all cases, URSM can be applied to obtain an accurate estimate of the cell type specific profile.

There are many existing bulk RNA-seq data sets for various human and nonhuman tissues that can be paired with different single cell data and jointly modeled using this unified framework. We also conduct simulation studies to demonstrate that as long as most cell types are consistent across the two data sources, URSM is robust to subtle mismatched cell types.

As for computation, the bottleneck is the Gibbs sampling step, which scales linearly with N, M, L and K. In practice, we find that a few hundred Gibbs samples and 50–100 EM iterations are usually enough to obtain sensible results. In our experiment, for 100 single cells and 150 bulk samples, one EM iteration with 150 Gibbs samples takes about 3 minutes for 200 genes and 12 minutes for 1000 genes using a single core on a computer equipped

with an AMD Opteron(tm) Processor 6320 @ 2.8 GHz. It is straightforward to further reduce the computation time by utilizing the conditional independency to parallelize the Gibbs sampling procedure.

Many downstream analyses can be conducted with this framework. In particular, URSM provides accurate estimates of the cell type specific profile matrix, which can be used for differential expression analysis between diseased and control samples. One can also apply URSM to single cells sequenced at different developmental periods to study the developmental trajectories of the cellular profiles.

As technologies improve and costs decline, single cell analysis can move to the new level by incorporating differential expression by maternal or paternal source of the chromosome. Such information can be captured if there are genetic differences between parents in the genes. Moreover, genetic variation can affect expression of genes; already experiments are being performed to determine which genetic variants are associated with changes in single cell expression. This would allow analysis of expression based on parental origin of each copy of the gene. These sources of variation are ignored in our model. Refining and extending scRNA-seq analytical tools to accommodate these sources of variation is one of the challenges for the future.

In this paper, we present our model assuming a given number of cell types K. In the situation where K is not known a priori, one can first run the model using a larger value of K, examine the clustering of single cells after imputation, and then reduce to a reasonable choice of K by combining cells from similar clusters.

Finally, we point out that the current model is developed under the setting of supervised learning where the labels for single cells are known. One can extend this framework to conduct unsupervised cell clustering by introducing extra latent variables for cell labels in the hierarchical model. In addition, by the nature of the Multinomial distribution, the current model is fully determined by its first moment. Therefore, the imputation of single cell data may be further improved by introducing gene-gene correlations to the model. We leave the exploration in these directions to future work.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the anonymous reviewers and Editor for their constructive comments and suggestions.

References

Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF. Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. PLoS ONE. 2009; 4:e6098. [PubMed: 19568420]

Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. J Amer Statist Assoc. 2017; 112:859–877.

Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003; 3:993–1022.

Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, Baying B, Benes V, Teichmann SA, Marioni JC, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013; 10:1093–1095. [PubMed: 24056876]

Camp JG, Badsha F, Florio M, Kanton S, Gerber T, Wilsch-Bräuninger M, Lewitus E, Sykes A, Hevers W, Lancaster M, et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. Proc Natl Acad Sci USA. 2015; 112:15672–15677. [PubMed: 26644564]

Casella G, George EI. Explaining the Gibbs sampler. Amer Statist. 1992; 46:167–174.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Statist Soc Ser B. 1977; 39:1–38. With discussion.

Donoho D, Stodden V. When does non-negative matrix factorization give a correct decomposition into parts? Advances in Neural Information Processing Systems. 2003

Dupuy C, Bach F. Online but accurate inference for latent variable models with local Gibbs sampling. J Mach Learn Res. 2016:1.

Fan HC, Fu GK, Fodor SPA. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015; 347:1258367. [PubMed: 25657253]

Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, Slichter CK, Miller HW, McElrath MJ, Prlic M, Linsley PS, Gottardo R. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015; 16:278. [PubMed: 26653891]

Fromer M, Roussos P, Sieberts SK, Johnson JS, Kavanagh DH, Perumal TM, Ruderfer DM, Oh EC, Topol A, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci. 2016; 19:1442–1453. [PubMed: 27668389]

Gaujoux R, Seoighe C. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: A case study. Infect Genet Evol. 2012; 12:913–921. [PubMed: 21930246]

Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. J Amer Statist Assoc. 1990; 85:398–409.

Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell. 1984; 6:721–741. [PubMed: 22499653]

Griffiths TL, Steyvers M. Finding scientific topics. Proc Natl Acad Sci USA. 2004; 101:5228–5235. [PubMed: 14872004]

Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, Clevers H, van Oudenaarden A. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015; 525:251–255. [PubMed: 26287467]

GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet. 2013; 45:580–585. [PubMed: 23715323]

Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Gen Med. 2017; 9:75.

Huang W-C, Ferris E, Cheng T, Hörndli CS, Gleason K, Tamminga C, Wagner JD, Boucher KM, Christian JL, Gregg C. Diverse non-genetic, allele-specific expression effects shape genetic architecture at the cellular level in the mammalian brain. Neuron. 2017a; 93:1094–1109.e7. [PubMed: 28238550]

Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray J, Raj A, Li M, Zhang NR. Gene expression recovery for single cell RNA sequencing. BioRxiv. 2017b; doi: 10.1101/138677

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Mach Learn. 1999; 37:183–233.

Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AM, Pletikos M, Meyer KA, Sedmak G, et al. Spatio-temporal transcriptome of the human brain. Nature. 2011; 478:483–489. [PubMed: 22031440]

Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014; 11:740–742. [PubMed: 24836921]

Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC, Teichmann SA. The technology and biology of single-cell RNA sequencing. Mol Cell. 2015; 58:610–620. [PubMed: 26000846]

Lee DD, Seung HS. Advances in Neural Information Processing Systems. Vol. 13. MIT Press; Cambridge, MA: 2000. Algorithms for non-negative matrix factorization; 556–562.

Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biol. 2017; 18:59. [PubMed: 28351406]

Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12:453–457. [PubMed: 25822800]

Padovan-Merhar O, Raj A. Using variability in gene expression as a tool for studying gene regulation. Wiley Interdiscip Rev, Syst Biol Med. 2013; 5:751–759. [PubMed: 23996796]

Pierson E, Yau C. ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. Genome Biol. 2015; 16:241. [PubMed: 26527291]

Polson NG, Scott JG, Windle J. Bayesian inference for logistic models using Pólya–gamma latent variables. J Amer Statist Assoc. 2013; 108:1339–1349.

Prabhakaran S, Azizi E, Pe'er D. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. Proceedings of the 33rd International Conference on Machine Learning; 2016. 1070–1079.

Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SHE, Jacobsen M. Biomarker discovery in heterogeneous tissue samples-taking the in-silico deconfounding approach. BMC Bioinform. 2010; 11:27.

Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015; 33:495–502. [PubMed: 25867923]

Shen-Orr SS, Tibshirani R, Khatri P, Bodian DL, Staedtler F, Perry NM, Hastie T, Sarwal MM, Davis MM, Butte AJ. Cell type-specific gene expression differences in complex tissues. Nat Methods. 2010; 7:287–289. [PubMed: 20208531]

Sunkin SM, Ng L, Lau C, Dolbeare T, Gilbert TL, Thompson CL, Hawrylycz M, Dang C. Allen brain atlas: An integrated spatio-temporal portal for exploring the central nervous system. Nucleic Acids Res. 2013; 41:D996–D1008. [PubMed: 23193282]

Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014; 32:381–386. [PubMed: 24658644]

Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. 2015; 11:e1004333. [PubMed: 26107944]

Vallejos CA, Richardson S, Marioni JC. Beyond comparisons of means: Understanding changes in gene expression at the single-cell level. Genome Biol. 2016; 17:1. [PubMed: 26753840]

Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: Challenges and opportunities. Nat Methods. 2017; 14:565–571. [PubMed: 28504683]

Vu TN, Wills QF, Kalari KR, Niu N, Wang L, Rantalainen M, Pawitan Y. Beta-Poisson model for single-cell RNA-seq data analyses. Bioinformatics. 2016; 32:2128–35. [PubMed: 27153638]

Wainwright MJ, Jordan MI. Graphical models, exponential families, and variational inference. Foundations and Trends® in Machine Learning. 2008; 1:1–305.

Zhong Y, Wan Y-W, Pang K, Chow LM, Liu Z. Digital sorting of complex tissues for cell type-specific gene expression profiles. BMC Bioinform. 2013; 14:1.

Zhu L, Lei J, Devlin B, Roeder K. 2018; Supplement to "A unified statistical framework for single cell and bulk RNA sequencing data.". doi: 10.1214/17-AOAS1110SUPP

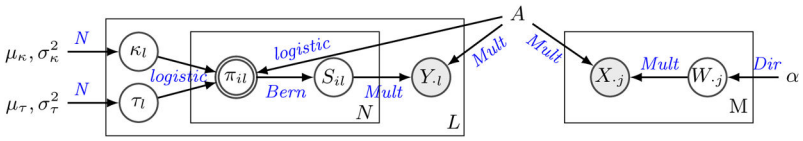


Fig. 1.

Plate model of URSM, with both single cell data (on the left) and bulk samples (on the right). The two greyed nodes X and Y represent observed gene expression levels. Node S is a binary variable representing dropout status in single cells, and node W represents the mixing proportions in bulk samples. The node **π** representing observation probability is double-circled because it is deterministic, and all model parameters are shown without circles, including the profile matrix A that links the two data sources.

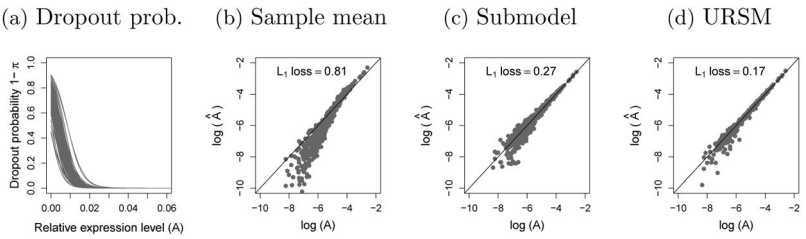


Fig. 2.

(a) Simulated logistic dropout probability curves for 100 single cells, as defined in equation (4.1). (b)–(d) True profile matrix A versus the estimated Â, plotted in the log scale, using (i) the naive sample mean estimation [equation (4.2)], (ii) a submodel using only single cell data and (iii) URSM with both single cell and bulk data. The L_1 loss $\Sigma_{i,k} |\hat{A}_{ik} - A_{ik}|$ is reported on the top.

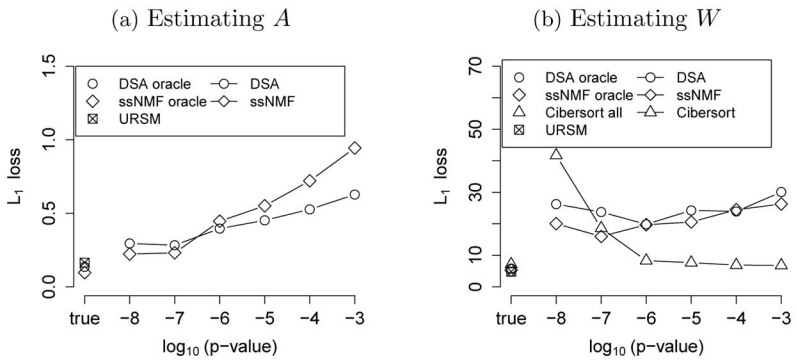


Fig. 3.

The L_1 loss of recovering (a) the profile matrix, $\Sigma_{i,k} |\hat{A}_{ik} - A_{ik}|$, and (b) mixing proportions, $\Sigma_{k,j} |\hat{W}_{kj} - W_{kj}|$. We evaluate DSA and ssNMF when the marker genes are extracted from single cell data using different thresholds of p-values, as well as under the oracle condition where the true marker genes are given. We evaluate Cibersort on estimating W when the input signature matrix is based on the estimated $\hat{A}$ from URSM. We report its performance when the entire $\hat{A}$ is used ("Cibersort all"), as well as when only the estimated marker genes are used ("Cibersort"). The performance of URSM is plotted with a square in both panels, which does not depend on thresholding p-values.

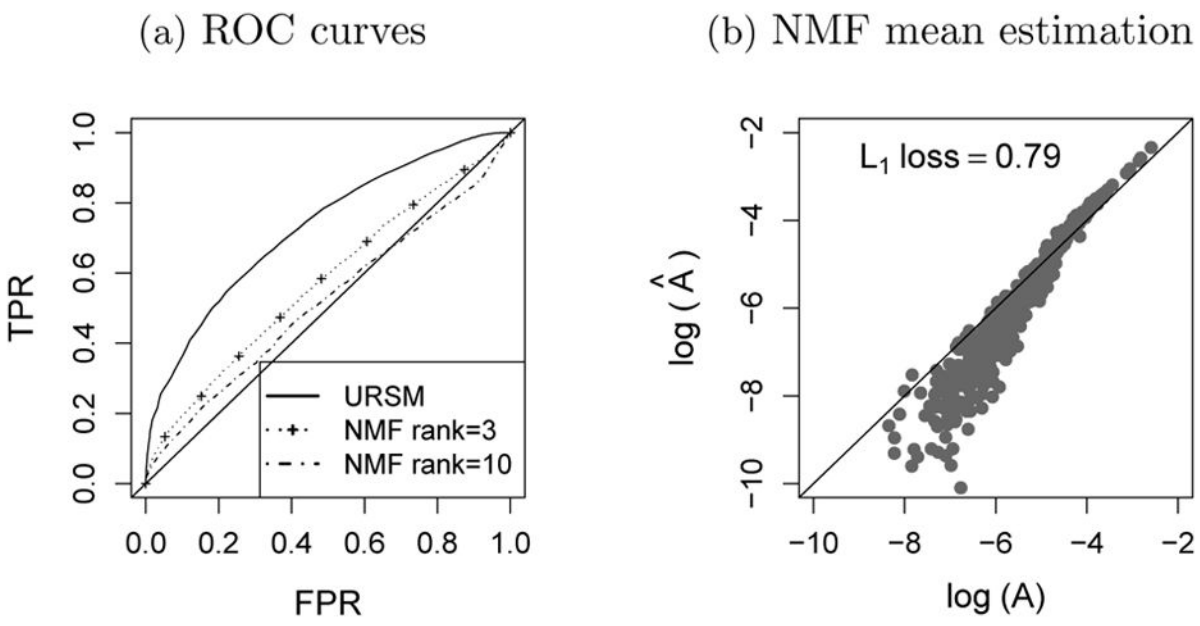
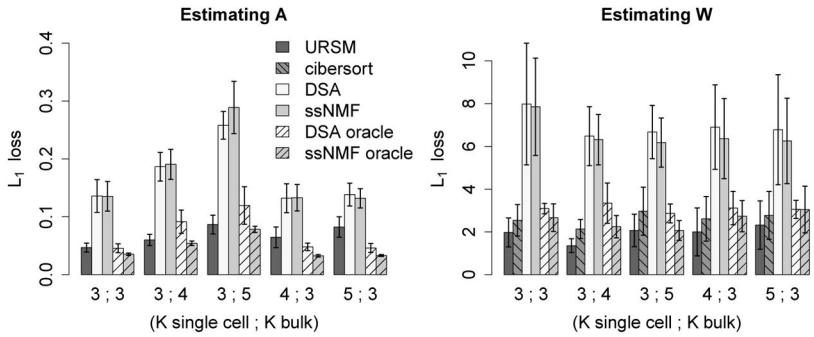
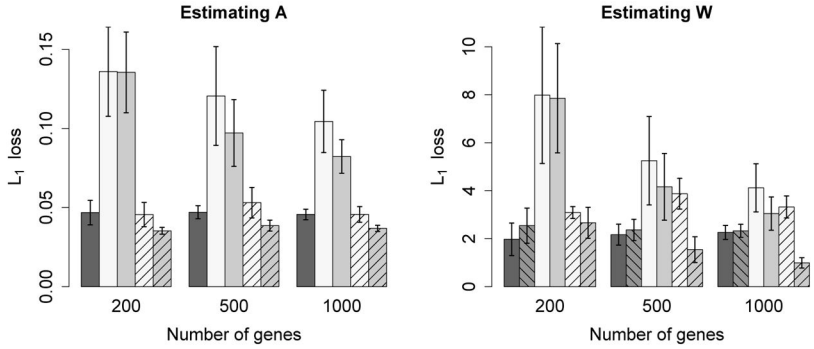


Fig. 4.

(a) ROC curves of identifying dropout entries in single cell data. (b) True profile matrix A versus the sample average of a rank-3 NMF approximation, plotted in the log scale. The L_1 loss $\Sigma_{i,k} |\hat{A}_{ik} - A_{ik}|$ is reported on the top.



(a) $N = 200$, varying K_{sc} and K_{bk}



(b) $K_{sc} = K_{bk} = 3$, varying N

Fig. 5.

The average per cell type L_1 loss of recovering the profile matrix A and the mixing proportions W in 10 repetitions, with the standard deviations shown by the error bars, when (a) $K_{sc}, K_{bk} \in \{3, 4, 5\}$ with N = 200 genes; (b) N = {200, 500, 1000} with $K_{sc} = K_{bk} = 3$. Each figure shows the performance of (i) URSM, (ii) DSA and ssNMF with 5 true marker genes and 3 imperfect marker genes per cell type as input and (iii) DSA and ssNMF under the oracle scenario where 5 true marker genes per cell type are provided. We also report the performance of Cibersort for estimating W using the estimated $\hat{A}_{unif}$ from URSM as the input signature matrix.

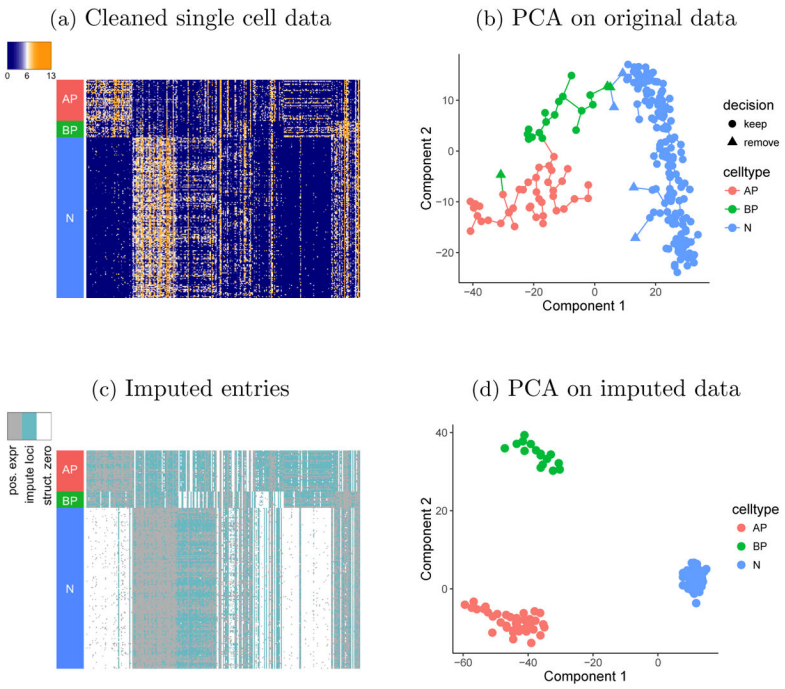


Fig. 6.

(a) Single cell gene expressions ($\log_2$(FPKM + 1)) after removing 7 ambiguously labeled cells. Rows are 213 cells and columns are 273 genes. (b) PCA applied on the original single cell data with 220 labeled cells using 273 PC genes, where the Monocle algorithm is applied to construct pseudo developmental times. 7 cells are identified to be ambiguously labeled and are removed from our analyses (marked as triangles). (c) Entries in cleaned single cell data that are inferred to be dropout and imputed (marked in blue) versus the entries that are inferred to be structural zeros (marked in white) in cleaned single cell data. The entries with positive expression levels have no need for posterior inference, and are marked in grey. (d) After imputing dropout genes, PCA is conducted on the 213 cells using 273 PC genes, and the three different types of cells are more clearly separated.

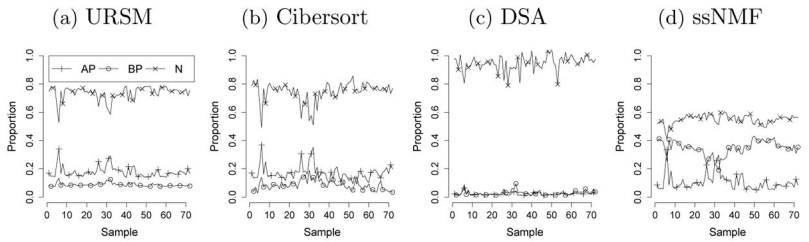


Fig. 7.

Deconvolution of bulk samples into three cell types, using (a) URSM, (b) Cibersort, (c) Digital Sorting Algorithm (DSA) and (d) semi-supervised Nonnegative Matrix Factorization (ssNMF).

Table 1

Number of selected marker genes using different thresholding p-values

	$\log_{10}$ (*p*-value)						
# of markers	**−8**	**−7**	**−6**	**−5**	**−4**	**−3**	**True markers**
cell type 1	5	8	11	16	19	27	10
cell type 2	2	2	8	11	16	23	10
cell type 3	1	2	7	10	17	21	10

Table 2

Correlation between the estimated proportions of a cell type k in bulk samples, $(W_{kj})_j$, and the normalized expression levels $(X_{ij}/R_j)_j$ of its marker gene i in bulk samples. For genes marking both AP and BP, the sum of proportions is used

Gene	Marked cell type	URSM	Cibersort	DSA	ssNMF
HES1	AP	0.73	0.62	**0.80**	0.68
HES6	BP	**0.66**	0.58	0.53	−0.72
PAX6	AP.BP	**0.91**	0.80	0.80	0.61
GLI3	AP.BP	**0.90**	0.80	0.83	0.54
NEUROD6	N	0.28	**0.37**	0.02	−0.36
BCL11B	N	0.45	**0.57**	0.23	0.02
MYT1L	N	**0.44**	0.37	0.32	0.80