Non-negative matrix factorization algorithms greatly improve topic model fits

Peter Carbonetto
Research Computing Center
University of Chicago
pcarbo@uchicago.edu

Abhishek Sarkar
Department of Human Genetics
University of Chicago
aksarkar@uchicago.edu

Zihao Wang
Department of Statistics
University of Chicago
wangzh@uchicago.edu

Matthew Stephens
Departments of Statistics and Human Genetics
University of Chicago
mstephens@uchicago.edu

Abstract

We report on the potential for using algorithms for non-negative matrix factorization (NMF) to improve parameter estimation in topic models. While several papers have studied connections between NMF and topic models, none have suggested leveraging these connections to develop new algorithms for fitting topic models. Importantly, NMF avoids the "sum-to-one" constraints on the topic model parameters, resulting in an optimization problem with simpler structure and more efficient computations. Building on recent advances in optimization algorithms for NMF, we show that first solving the NMF problem then recovering the topic model fit can produce remarkably better fits, and in less time, than standard algorithms for topic models. While we focus primarily on maximum likelihood estimation, we show that this approach also has the potential to improve variational inference for topic models. Our methods are implemented in the R package `fastTopics`, available at `https://github.com/stephenslab/fastTopics`.

1 Introduction

Since their introduction more than twenty years ago, topic models have been widely used to identify structure in large collections of documents [7, 32, 36, 37, 49]. In brief, topic models analyze a term-document count matrix—in which entry (i, j) records the number of occurrences of term j in document i—to learn a representation of each document as a non-negative, linear combination of "topics," each of which is a vector of term frequencies. More recently, topic models have been successfully applied to another setting involving large count matrices, analysis of single-cell genomics data [5, 18, 29, 66]. While there have been many different approaches to topic modeling, with different fitting procedures, prior distributions, and/or names (e.g., the aspect model [38]; probabilistic latent semantic indexing, or pLSI [35, 36]; latent Dirichlet allocation, or LDA [7]), most of these approaches are based on the same basic model: a multinomial distribution of the counts [36], which we refer to as the "multinomial topic model" (see eq. 3).

Fitting topic models is computationally challenging, and has attracted considerable attention. Early work used a simple expectation maximization (EM) algorithm to obtain maximum likelihood estimates (MLEs) under the multinomial topic model [36]. This approach exploits a natural data augmentation representation of the topic model in which latent variables are introduced to represent which topic gave rise to each term in each document. This same data augmentation representation was subsequently used to fit topic models in Bayesian frameworks by using a mean-field variational approximation

Preprint. Under review.

[7], and by using a Markov chain Monte Carlo method called the "collapsed" Gibbs sampler [32]. Teh *et al* [62] later combined ideas from both methods in a collapsed variational inference algorithm. These algorithms are summarized and compared in [3], who note their similarity in structure and performance. The speed of these algorithms on large data sets can be improved by the use of "online" versions that exploit stochastic approximation techniques [35, 57].

In parallel to this work on topic models, the last twenty years have also seen considerable progress in the closely related problem of non-negative matrix factorization (NMF), which attempts to find a factorization of a non-negative data matrix $\mathbf{X}$ into two non-negative matrices $\mathbf{L}, \mathbf{F}$ such that $\mathbf{X} \approx \mathbf{L}\mathbf{F}^T$ [16, 43]. While the potential for NMF to discover structure in text documents was recognized early in its development [43], a full recognition of the formal connection between NMF and topic models has emerged more gradually [9–11, 20, 21, 25, 27, 69], and perhaps remains underappreciated. A key result is that the objective function for NMF—specifically, NMF based on maximizing a Poisson log-likelihood [43], which we refer to as "Poisson NMF"—is equal, up to a constant, to the log-likelihood for the multinomial topic model (Lemma 1).

In revisiting this close connection, we make two key contributions. First, we suggest a new strategy to fit the multinomial topic model: first fit the Poisson NMF model, then recover the equivalent topic model by a simple transformation (eq. 4). This strategy has several benefits over fitting the topic model directly; in particular, the Poisson NMF optimization problem has simpler constraints and a structure that can be exploited for more efficient computation. While several papers have noted close connections between Poisson NMF and the multinomial topic model, as far as we know none have previously suggested this approach or argued for its benefits.

Second, we show that our proposed strategy can yield substantial improvements to topic model fits compared with the EM-like algorithms that are most commonly used. To achieve these improvements, we leverage recent advances in optimization methods for Poisson NMF [2, 34, 39, 45]. While the improvements we demonstrate are primarily for maximum likelihood (ML) estimation—which, for large data sets, has the benefit of being simple and fast—our experiments also suggest that these same ideas could be used to improve on the "EM-like" algorithms used to fit mean-field variational approximations for LDA. In a new R package, `fastTopics`, we provide implementations of our approach that exploit data sparsity to facilitate efficient analysis of large data sets [12].

2 Poisson NMF and the multinomial topic model

Here we provide side-by-side descriptions of Poisson NMF [16, 27, 43, 44] and the multinomial topic model [7, 9, 32, 36, 37, 49] to highlight their close connection. While formal and informal connections between these two models have been made previously [9–11, 20, 21, 25, 69], these previous papers draw connections between the algorithms and/or stationary points of the objective functions. By contrast, we state a simple and more general result relating the likelihoods of the two models (Lemma 1), which we view as a more fundamental result underlying previous results.

Let $\mathbf{X}$ denote an observed $n \times m$ matrix of counts x_{ij}. For example, in analysis of text documents, x_{ij} would denote the number of times term j occurs in document i. Both Poisson NMF and the multinomial topic model can be thought of as fitting different, but closely related, models for $\mathbf{X}$.

The Poisson NMF model has parameters that are non-negative matrices, $\mathbf{L} \in \mathbf{R}_+^{n \times K}$ and $\mathbf{F} \in \mathbf{R}_+^{m \times K}$, where $\mathbf{R}_+^{m \times n}$ denotes the set of non-negative, real $m \times n$ matrices. Given $K \geq 1$, the model is

$$\begin{aligned} x_{ij} \mid \mathbf{L}, \mathbf{F} &\sim \text{Poisson}(\lambda_{ij}), \\ \lambda_{ij} &= (\mathbf{L}\mathbf{F}^T)_{ij} = \textstyle\sum_{k=1}^K l_{ik} f_{jk}. \end{aligned} \tag{1}$$

Poisson NMF can be viewed as a matrix factorization by noting that (1) implies $E[\mathbf{X}] = \mathbf{L}\mathbf{F}^T$, so fitting a Poisson NMF essentially seeks values of $\mathbf{L}$ and $\mathbf{F}$ such that $\mathbf{X} \approx \mathbf{L}\mathbf{F}^T$.[1]

The multinomial topic model also has matrix parameters, $\mathbf{L}^\star \in \mathbf{R}_+^{n \times K}$ and $\mathbf{F}^\star \in \mathbf{R}_+^{m \times K}$ whose elements $l_{ik}^\star, f_{jk}^\star$ must satisfy additional "sum-to-one" constraints,

$$\textstyle\sum_{j=1}^m f_{jk}^\star = 1, \quad \sum_{k=1}^K l_{ik}^\star = 1. \tag{2}$$

[1] In this formulation, $\mathbf{L}, \mathbf{F}$ are not uniquely identifiable; for example, multiplying the kth column of $\mathbf{L}$ by $a_k > 0$ and dividing the kth column of $\mathbf{F}$ by a_k does not change $\mathbf{L}\mathbf{F}^T$. One can avoid this non-identifiability by imposing constraints or introducing priors on either $\mathbf{L}$ or $\mathbf{F}$, but this is not needed for our aims here.

Given $K \geq 2$, the multinomial topic model is

$$\begin{aligned} &x_{i1}, \ldots, x_{im} \mid \mathbf{L}^\star, \mathbf{F}^\star, t_i \sim \text{Multinomial}(t_i; \pi_{i1}, \ldots, \pi_{im}), \\ &\pi_{ij} = \left(\mathbf{L}^\star (\mathbf{F}^\star)^T\right)_{ij} = \textstyle\sum_{k=1}^K l_{ik}^\star f_{jk}^\star, \end{aligned} \tag{3}$$

where $t_i \equiv \sum_{j=1}^m x_{ij}$. The topic model can also be viewed as a matrix factorization [36, 59, 60], $\mathbf{\Pi} = \mathbf{L}^* (\mathbf{F}^*)^T$, where $\mathbf{\Pi} \in \mathbf{R}_+^{n \times m}$ is the matrix of multinomial probabilities π_{ij}.

Now we state an equivalence between Poisson NMF and the multinomial topic model: we define a mapping between the parameter spaces for the two models (Definition 1), then we state an equivalence between their likelihoods (Lemma 1), which leads to an equivalence in the MLEs (Corollary 1).

Definition 1 (Poisson NMF–multinomial topic model reparameterization)**.** Given $K \geq 2$, $\mathbf{L} \in \mathbf{R}_+^{n \times K}$, $\mathbf{F} \in \mathbf{R}_+^{m \times K}$, define the mapping $\psi : \mathbf{L}, \mathbf{F} \mapsto \mathbf{L}^\star, \mathbf{F}^\star, \boldsymbol{s}, \boldsymbol{u}$ by the following procedure:

$$\begin{aligned} u_k &\leftarrow \textstyle\sum_{j=1}^m f_{jk}, & k &= 1, \ldots, K \\ f_{jk}^\star &\leftarrow f_{jk}/u_k, & j &= 1, \ldots, m,\ k = 1, \ldots, K \\ s_i &\leftarrow \textstyle\sum_{k=1}^K l_{ik} u_k, & i &= 1, \ldots, n \\ l_{ik}^\star &\leftarrow l_{ik} u_k / s_i, & i &= 1, \ldots, n,\ k = 1, \ldots, K. \end{aligned} \tag{4}$$

This is a one-to-one mapping, and its inverse $\psi^{-1} : \mathbf{L}^\star, \mathbf{F}^\star, \boldsymbol{s}, \boldsymbol{u} \mapsto \mathbf{L}, \mathbf{F}$ is $f_{jk} \leftarrow u_k f_{jk}^\star$, $l_{ik} \leftarrow s_i l_{ik}^\star / u_k$. The mapping ψ yields $\mathbf{L}^\star \in \mathbf{R}_+^{n \times K}$, $\mathbf{F}^\star \in \mathbf{R}_+^{m \times K}$ that satisfy the sum-to-one constraints.

Lemma 1 (Equivalence of Poisson NMF and multinomial topic model likelihoods)**.** Denote the Poisson NMF model (1) probability density by $p_{\mathsf{PNMF}}(\mathbf{X} \mid \mathbf{L}, \mathbf{F})$ and denote the multinomial topic model (3) probability density by $p_{\mathsf{topic}}(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star)$. Assume $\mathbf{L} \in \mathbf{R}_+^{n \times K}$ and $\mathbf{F} \in \mathbf{R}_+^{m \times K}$, and let $\mathbf{L}^\star, \mathbf{F}^\star, \boldsymbol{s}, \boldsymbol{u}$ be the result of applying ψ to $\mathbf{L}, \mathbf{F}$. Then we have

$$p_{\mathsf{PNMF}}(\mathbf{X} \mid \mathbf{L}, \mathbf{F}) = p_{\mathsf{topic}}(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star) \times \textstyle\prod_{i=1}^n \text{Poisson}(t_i; s_i). \tag{5}$$

Proof. The result follows from the well-known relationship between the multinomial and Poisson,

$$\textstyle\prod_{j=1}^m \text{Poisson}(x_j; \lambda_j) = \text{Multinomial}(x; t, \lambda_1/s, \ldots, \lambda_m/s) \times \text{Poisson}(t; s), \tag{6}$$

which holds for any $\lambda_1, \ldots, \lambda_m \in \mathbf{R}_+$, in which $s = \sum_{j=1}^m \lambda_j$ and $t = \sum_{j=1}^m x_j$ [23, 30, 69]. □

Lemma (1) is more general than previous results [10, 20, 25, 27] because it applies to *any* $\mathbf{L}, \mathbf{F}, \mathbf{L}^\star, \mathbf{F}^\star$ from Definition 1, not just a stationary point of the likelihood. See [9, 21, 69] for related results.

Corollary 1 (Relation between MLEs for Poisson NMF and multinomial topic model)**.** *Let* $\hat{\mathbf{L}} \in \mathbf{R}_+^{n \times K}, \hat{\mathbf{F}} \in \mathbf{R}_+^{m \times K}$ *denote MLEs for the Poisson NMF model,* $\hat{\mathbf{L}}, \hat{\mathbf{F}} \in \text{argmax}_{\mathbf{L},\mathbf{F}}\, p_{\mathsf{PNMF}}(\mathbf{X} \mid \mathbf{L}, \mathbf{F})$. [2] *If* $\hat{\mathbf{L}}^\star, \hat{\mathbf{F}}^\star$ *are obtained by applying* ψ *to* $\hat{\mathbf{L}}, \hat{\mathbf{F}}$*, these are MLEs for the multinomial topic model,* $\hat{\mathbf{L}}^\star, \hat{\mathbf{F}}^\star \in \text{argmax}_{\mathbf{L}^\star,\mathbf{F}^\star}\, p_{\mathsf{topic}}(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star)$. *Conversely, let* $\hat{\mathbf{L}}^\star \in \mathbf{R}_+^{n \times K}, \hat{\mathbf{F}}^\star \in \mathbf{R}_+^{m \times K}$ *denote multinomial topic model MLEs, set* $s_i = t_i$*, for* $i = 1, \ldots, n$*, and choose any* $u_1, \ldots, u_K \in \mathbf{R}^+$*. If* $\hat{\mathbf{L}}, \hat{\mathbf{F}}$ *are obtained by applying* ψ^{-1} *to* $\hat{\mathbf{L}}^\star, \hat{\mathbf{F}}^\star, \boldsymbol{s}, \boldsymbol{u}$*, these are Poisson NMF MLEs.*

See Appendix A.1 for a generalization of Corollary 1 to *maximum a posteriori* estimation with Gamma priors on $\mathbf{F}$ or Dirichlet priors on $\mathbf{F}^\star$, and with uniform priors on $\mathbf{L}$ or $\mathbf{L}^\star$.

3 Optimization algorithms for Poisson NMF

Corollary 1 implies that *any algorithm for ML estimation in Poisson NMF is also an algorithm for ML estimation in the multinomial topic model.* For a given K, fitting the Poisson NMF model involves solving the following optimization problem:

$$\begin{aligned} &\text{minimize } \ell(\mathbf{L}, \mathbf{F}) \equiv \sum_{i=1}^n \sum_{j=1}^m \boldsymbol{l}_i^T \boldsymbol{f}_j - x_{ij} \log(\boldsymbol{l}_i^T \boldsymbol{f}_j) \\ &\text{subject to } \mathbf{L} \geq 0, \mathbf{F} \geq 0, \end{aligned} \tag{7}$$

[2] The notation $\hat{\theta} \in \text{argmax}_\theta\, h(\theta)$ means $h(\hat{\theta}) \geq h(\theta)$ for all θ, and accounts for the fact that an MLE may not be unique due to non-identifiability.

Require: data $\mathbf{X} \in \mathbf{R}_+^{n \times m}$, initial estimates $\mathbf{L}^{(0)} \in \mathbf{R}_+^{n \times K}$, $\mathbf{F}^{(0)} \in \mathbf{R}_+^{m \times K}$, and a function FIT-POIS-REG$(\mathbf{A}, \boldsymbol{y})$ that returns a MLE of $\boldsymbol{b}$ in (9).

for $t = 1, 2, \ldots$ **do**
  for $i = 1, \ldots, n$ **do** {iterations can be performed in parallel}
    $\boldsymbol{l}_i \leftarrow$ FIT-POIS-REG$(\mathbf{F}^{(t-1)}, \boldsymbol{x}_i)$
    Store $\boldsymbol{l}_i$ in ith row of $\mathbf{L}^{(t)}$
  end for
  for $j = 1, \ldots, m$ **do** {iterations can be performed in parallel}
    $\boldsymbol{f}_j \leftarrow$ FIT-POIS-REG$(\mathbf{L}^{(t-1)}, \boldsymbol{x}_j)$
    Store $\boldsymbol{f}_j$ in jth row of $\mathbf{F}^{(t)}$
  end for
end for

Algorithm 1: Alternating Poisson Regression for Poisson NMF ($\boldsymbol{x}_i$ is a row of $\mathbf{X}$, $\boldsymbol{x}_j$ is a column).

where $\boldsymbol{l}_i, \boldsymbol{f}_j$ are column vectors containing the ith row of $\mathbf{L}$ and the jth row of $\mathbf{F}$, respectively. The loss function $\ell(\mathbf{L}, \mathbf{F})$ is equal to $-\log p_{\mathsf{PNMF}}(\mathbf{X} \mid \mathbf{L}, \mathbf{F})$ after removing terms that do not depend on $\mathbf{L}$ or $\mathbf{F}$. This loss function can also be derived as a divergence measure [15, 19, 22]. Since $\ell(\mathbf{L}, \mathbf{F})$ is not convex when $K \geq 2$, we seek only to find a local minimum. (The special case of $K = 1$ has a closed-form solution, but only factorizations with $K \geq 2$ yield topic model fits.)

A key motivation for our proposed strategy is that the Poisson NMF optimization problem (7) is simpler than the usual formulation for multinomial topic models because it lacks the sum-to-one constraints (2). However, not all approaches to (7) exploit this benefit. Indeed, the traditional way to solve (7)—the "multiplicative updates" of Lee and Seung [43]—is equivalent to EM [14], and is closely related to the EM-like algorithms traditionally used in topic models. In contrast, other recently developed algorithms for solving (7) based on co-ordinate descent (CD) do not have an existing counterpart in the topic model literature (CD cannot obviously deal with the sum-to-one constraints). These CD algorithms can greatly outperform EM-based methods for NMF [34], and here we argue that they should also be the approach of choice for fitting topic models.

To facilitate comparison of the EM and CD algorithms, in Sec. 3.1 we introduce an "Alternating Poisson Regression" framework for fitting Poisson NMF that includes both EM and CD as special cases. Next we describe the algorithms we have implemented, drawing on recent work [2, 34, 39, 45] and our own experimentation. See [34] for a more comprehensive review of Poisson NMF methods.

3.1 Alternating Poisson Regression for Poisson NMF

Alternating Poisson Regression arises from solving (7) by alternating between optimizing over $\mathbf{L}$ with $\mathbf{F}$ fixed, and optimizing over $\mathbf{F}$ with $\mathbf{L}$ fixed. This is an example of a block-coordinate descent algorithm [65] (also known as nonlinear Gauss-Seidel [4]), where the "blocks" are $\mathbf{L}$ and $\mathbf{F}$. It is analogous to "alternating least squares" for matrix factorization with Gaussian errors.

We highlight two simple but important points. First, by symmetry of (7), optimizing $\mathbf{F}$ given $\mathbf{L}$ has exactly the same form as optimizing $\mathbf{L}$ given $\mathbf{F}$, which simplifies implementation. Second, because of the separability of the sum in (7), optimizing $\mathbf{F}$ given $\mathbf{L}$ breaks down into m independent K-dimensional subproblems of the form

$$\begin{array}{l} \text{minimize } \phi_j(\boldsymbol{f}_j) \equiv \sum_{i=1}^n \boldsymbol{l}_i^T \boldsymbol{f}_j - x_{ij} \log(\boldsymbol{l}_i^T \boldsymbol{f}_j) \\ \text{subject to } \boldsymbol{f}_j \geq 0, \end{array} \tag{8}$$

for $j = 1, \ldots, m$ (and similarly for optimizing $\mathbf{L}$ given $\mathbf{F}$). Because the m subproblems (8) are independent, their solutions can be pursued in parallel. While both of these observations are simple, *neither of them hold in the multinomial topic model due to the additional sum-to-one constraints* (2).

The subproblem (8) is itself a well-studied ML estimation problem [63, 50]; it is equivalent to computing an MLE of $\boldsymbol{b} = (b_1, \ldots, b_K)^T \geq 0$ in the following additive Poisson regression model,

$$\begin{array}{l} y_i \sim \text{Poisson}(\mu_i), \\ \mu_i = a_{i1} b_1 + \cdots + a_{iK} b_K, \end{array} \tag{9}$$

in which the data are $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbf{R}_+^n$ and $\mathbf{A} \in \mathbf{R}_+^{n \times K}$. (We call the regression "additive" to distinguish it from the more common formulation in which the effects are multiplicative

[47].) For convenience, we introduce a function that returns an MLE of $\boldsymbol{b}$, FIT-POIS-REG$(\mathbf{A}, \boldsymbol{y}) \equiv \operatorname{argmax}_{\boldsymbol{b}} p_{\text{PR}}(\boldsymbol{y} \mid \mathbf{A}, \boldsymbol{b})$, where $p_{\text{PR}}(\boldsymbol{y} \mid \mathbf{A}, \boldsymbol{b}) = \prod_{i=1}^{n} p(y_i \mid \mu_i)$ denotes the likelihood for (9).

Any algorithm that solves FIT-POIS-REG be can applied iteratively to solve the Poisson NMF problem (7). This idea is formalized in Algorithm 1. Frobenius-norm NMF can be solved in a similar way by iteratively solving a series of non-negative least squares problems [26, 41].

3.2 Specific algorithms

We now consider two different approaches to fitting the Poisson regression model (9), which, when inserted into Algorithm 1, produce two different Poisson NMF algorithms. These algorithms are closely connected to several existing Poisson NMF methods, and we describe these connections.

3.2.1 Expectation maximization

There is a long history of solving the Poisson regression problem by EM [17, 42, 46, 48, 54, 58]. The EM updates for this problem simply iterate

$$\bar{z}_{ik} = y_i a_{ik} b_k / \mu_i \tag{10}$$

$$b_k = \textstyle\sum_{i=1}^{n} \bar{z}_{ik} / \sum_{i=1}^{n} a_{ik}, \tag{11}$$

where $\bar{z}_{ik}$ represents a posterior expectation in an augmented model (see Appendix A.2.1).

Combining the E (10) and M (11) steps with the substitutions given in Algorithm 1 yields the "multiplicative" update rules of Lee and Seung [44] (see Appendix A.2.2),

$$l_{ik}^{\text{new}} \leftarrow l_{ik} \times \frac{\sum_{j=1}^{m} x_{ij} f_{jk} / \lambda_{ij}}{\sum_{j=1}^{m} f_{jk}} \tag{12}$$

$$f_{jk}^{\text{new}} \leftarrow f_{jk} \times \frac{\sum_{i=1}^{n} x_{ij} l_{ik} / \lambda_{ij}}{\sum_{i=1}^{n} l_{ik}}. \tag{13}$$

Additionally, applying the reparameterization (4) to the multiplicative updates (12, 13) recovers the EM updates for the multinomial topic model [3, 9, 25, 37] (see Appendix A.2.3). Therefore, when FIT-POIS-REG is implemented using EM, Algorithm 1 can be viewed as a "stepwise" variant of the multiplicative updates for Poisson NMF, or EM for the multinomial topic model. By "stepwise," we mean that the update order suggested by Algorithm 1 is to iterate the E and M steps for $\boldsymbol{l}_1$, then for $\boldsymbol{l}_2$, and so on, followed by updates to rows of $\mathbf{F}$, whereas in a typical EM implementation, the E step is performed for all latent variables, followed by the M step for all parameters.

3.2.2 Co-ordinate descent

Co-ordinate descent is a simple alternative to EM that iteratively optimizes a single co-ordinate b_k while the remaining co-ordinates are fixed. An advantage of CD is that each 1-d optimization is easy to implement via Newton's method:

$$b_k^{\text{new}} \leftarrow \max\{0, b_k - \alpha_k g_k / h_k\}, \tag{14}$$

where g_k and h_k are partial derivatives with respect to $\ell_{\text{PR}}(\boldsymbol{b}) \equiv -\log p_{\text{PR}}(\boldsymbol{y} \mid \mathbf{A}, \boldsymbol{b})$,

$$g_k \equiv \frac{\partial \ell_{\text{PR}}}{\partial b_k} = \sum_{i=1}^{n} a_{ik} \left(1 - \frac{y_i}{\mu_i}\right), \qquad h_k \equiv \frac{\partial^2 \ell_{\text{PR}}}{\partial b_k^2} = \sum_{i=1}^{n} \frac{y_i a_{ik}^2}{\mu_i^2}, \tag{15}$$

and $\alpha_k \geq 0$ is a step size that can be determined by a line search [52] or some other method.

Several recent algorithms for Poisson NMF—the cyclic co-ordinate descent (CCD) [39], sequential co-ordinate descent (SCD) [45] and scalar Newton (SN) [34] methods—can be viewed as variants of this approach (see also [8]). The CCD and SCD methods appear to be independent developments of essentially the same algorithm; they both take a full (feasible) Newton step, setting $\alpha_k = 1$ when $b_k - \alpha_k g_k / h_k > 0$. By foregoing a line search to determine α_k, they are not guaranteed to decrease $\ell_{\text{PR}}(\boldsymbol{b})$ when $\boldsymbol{b}$ is far from a solution [52]. The SN method was developed to remedy this, with a step size scheme that always produces a decrease while avoiding the extra expense of a line search. Hien & Gillis [34] compared CCD, SN and other algorithms for Poisson NMF (they did not compare SCD), and found that, despite the absence of a line search, CCD usually performed best in real data sets.

3.3 Implementation details and enhancements

To accelerate convergence of the algorithms, we used the extrapolation method of [2]. Although this method was developed for NMF with the Frobenius norm objective, our initial investigations showed that it also worked well for Poisson NMF, and was more effective for this problem than other acceleration schemes we tried, including the damped Anderson [33] and quasi-Newton methods [68].

We implemented the algorithms, together with extrapolation and other enhancements, in an R package, `fastTopics` [12]. To make these algorithms scale well to the large data sets typically used in topic modeling, we reworked the computations to exploit sparsity of $\mathbf{X}$; for example, by precomputing $\sum_{i=1}^{n} \boldsymbol{l}_i$, the remaining computations needed to solve subproblem (8) with EM or CD scale linearly with the number of nonzeros in column j of $\mathbf{X}$. The result is that, for sparse $\mathbf{X}$, the complexity of the Poisson NMF algorithms is $O((N + n + m)K)$, where N is the number of nonzeros in $\mathbf{X}$, whereas for non-sparse $\mathbf{X}$ the complexity is $O(nmK)$. See Appendix A.4 for further details.

4 Fitting multinomial topic models using EM and CD: an illustration

To summarize, we have described two variants of Algorithm 1 for Poisson NMF: the first fits the Poisson regression model (9) using EM, and is equivalent to existing EM algorithms for Poisson NMF and the multinomial topic model (aside from differences in order of the updates); the second uses coordinate descent (CD) to fit (9), and has no equivalent among existing algorithms for the multinomial topic model. In the remainder, we refer to these two variants as the EM and CD algorithms.

Here we compare the performance of the EM and CD algorithms for fitting the multinomial topic model in two simulated data sets, one in which both algorithms perform similarly, and a second where CD greatly outperforms EM. R code implementing these simulations is provided in the companion repository [13].

We simulated two small 100×400 data sets from the correlated topic model [6], which uses a logistic normal distribution [1], $l_{ik} = \exp(\eta_{ik}) / \sum_{k'=1}^{K} \exp(\eta_{ik'})$, $(\eta_{i1}, \ldots, \eta_{iK}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, to model correlations among the topic proportions l_{ik}. For the first data set ("Scenario A"), we chose $\boldsymbol{\Sigma}$ with negative correlations, $\mu_k = 0$, $\sigma_{kk}^2 = 11, \sigma_{kk'} = -2, k \neq k'$, so that most documents had a single predominant topic (Fig. 1, Panel A). For the second data set ("Scenario B"), we made a small change to these settings, introducing a strong positive correlation, $\sigma_{56} = 8$, between topics 5 and 6. This results in many more documents showing non-trivial membership in both topic 5 and topic 6 (Panel B). This kind of positive correlation is not unexpected in real data; two related topics may appear together more often than their overall frequency would predict.

We ran the EM and CD algorithms on both data sets as described in Section 3.3 and Appendix A.5.3. We assessed convergence by examining the change in the log-likelihood and the residuals of the first-order Karush-Kuhn-Tucker (KKT) conditions, which should vanish near a solution. To reduce the possibility that different runs might find different local maxima, we initialized the algorithms by first running 50 EM updates. (By "update" or "iteration," we mean one iteration of the outer loop in Algorithm 1.) The results are shown in Figure 1.

In Scenario A, although CD converged faster, both algorithms progressed rapidly toward the same solution (Panels A1–A3). Thus, in this data set, use of EM or CD makes little practical difference. Scenario B shows a different story. EM initially made good progress—in the initialization phase, which is not shown in the plots, the 50 EM updates improved the the log-likelihood by >5,000—but after performing an 750 additional EM updates, the estimates remained far from the solution achieved by CD (Panels B1, B2). To be clear, EM continued to make progress even after 750 iterations, but this progress was slow; after the 750 iterations, the log-likelihood was increasing by about 0.1 at each iteration. Furthermore, the estimates of topic proportions differed substantially (Panel B3), showing that the large difference in likelihood reflects consequential differences in fit. In particular, EM often overestimated membership in topic 6, and consequently underestimated membership in other topics.

To rule out the possibility that EM had settled on a different local maximum, we performed an additional 200 CD updates from the final EM state (dashed, red line in Panels B1, B2). Doing so recovered the better CD solution.

While the EM algorithm we use here performs ML estimation, similar "variational EM" (VEM) algorithms are frequently used to fit variational approximations for topic models, notably in the

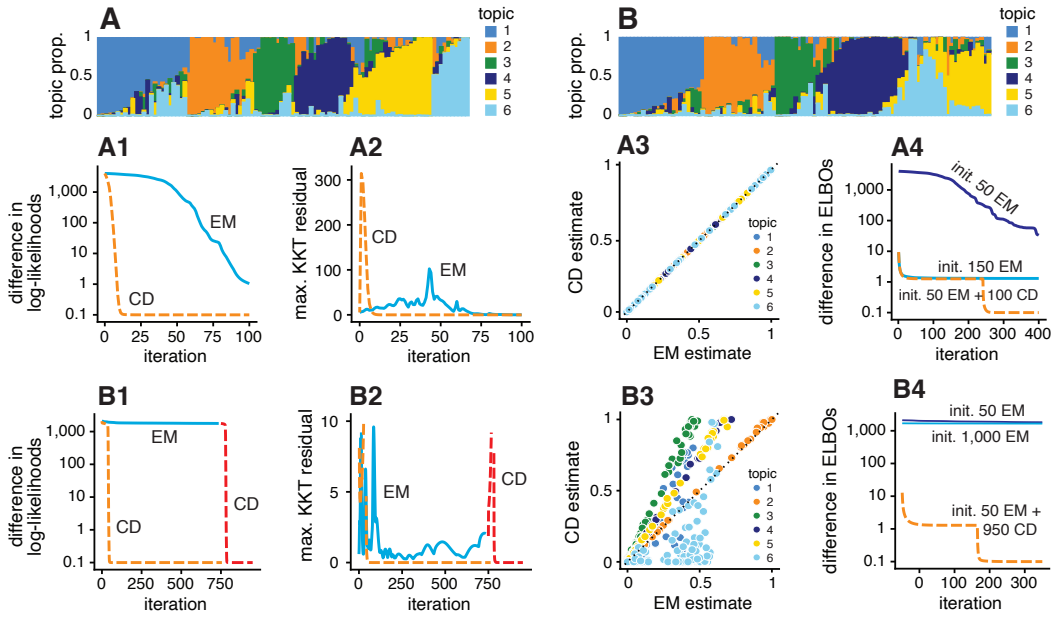

Figure 1: Results of fitting $K = 6$ topic models to data sets A and B. Panels A and B show the topic proportions used to simulate the two data sets. Panels A1, A2, B1, B2 show the improvement in the EM and CD estimates over time, as measured by the log-likelihood difference with respect to best $\log p_{\text{topic}}(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star)$ achieved by all fits, and by the maximum KKT residual. Plots A3, B3 compare the final EM and CD estimates of the topic proportions. Panels A4, B4 show the improvement in the LDA fit over time, in which VEM is initialized to the estimates from performing different numbers of EM and/or CD updates. Log-likelihood and ELBO differences less than 0.1 are shown as 0.1.

LDA approach of [7] in which point estimates $\hat{l}_{i1}, \ldots, \hat{l}_{iK}$ are replaced by an estimated posterior, $\text{Dirichlet}(\gamma_{i1}, \ldots, \gamma_{iK})$ [7]. Given the similarity of the algorithms [3, 10, 21], one might expect VEM to have similar convergence issues to EM. To check this, we ran VEM for LDA, initializing f_{jk} and γ_{ik} using either the EM or CD estimates. To replicate the ML estimation setting as closely as possible, we used a fixed, uniform prior for the topic proportions. In both scenarios, the CD initialization produced a better fit (higher Evidence Lower Bound, or ELBO) than the EM initializations (Panels A4 and B4). In Scenario B, the difference is particularly striking; the CD initialization produced a solution with ELBO that was over 1,000 units higher. While not systematic, these results nonetheless illustrate the potential for CD-based NMF methods to improve variational inference for LDA.

Note that the EM convergence problems in Scenario B are not easily solved using existing acceleration schemes for multinomial topic models; the quasi-Newton-accelerated EM algorithm, implemented in the `maptpx` R package [61], also shows slow convergence in this data set (Fig. A1 in the Appendix).

5 Numerical experiments

Having hinted at the advantages of CD over EM for fitting topic models, we now perform a more systematic comparison in real data sets. In particular, we compare Algorithm 1, implemented with EM or CD updates, with or without extrapolation [2]. We run these comparisons on four data sets (Table 1): two text data sets [28, 53] that have been used to evaluate topic modeling methods (e.g., [3, 64]); and two data sets from single-cell RNA sequencing (scRNA-seq) experiments [51, 67]. While the second application may be less familiar to readers, recent papers have illustrated the potential for topic models to uncover structure from scRNA-seq data, in particular structure that is not well captured by (hard) clustering [18, 66, 5, 40]. See Appendix A.5 for more details on the data sets.

To reduce the possibility that multiple optimizations converge to different local maxima of the likelihood, which could complicate the comparisons, we first ran 1,000 EM updates—that is, 1,000 iterations of the outer loop of Algorithm 1—then we examined the performance of the algorithms *after* this initialization phase. Therefore, in our comparisons we assessed the extent to which the different

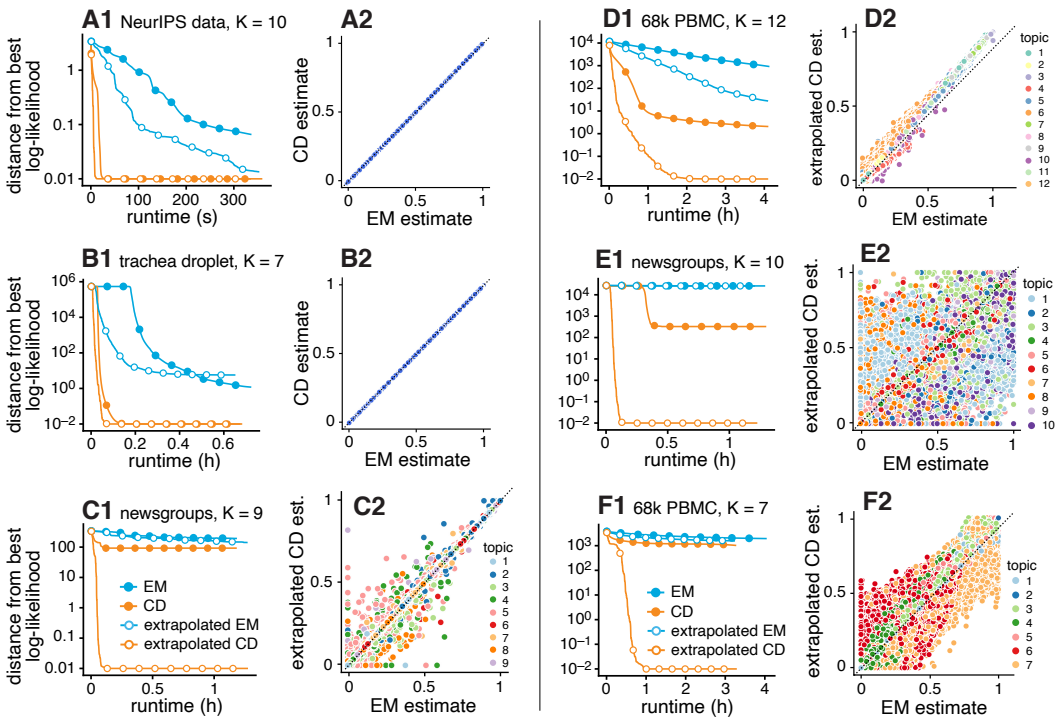

Figure 2: Selected results on fitting topic models using Poisson NMF algorithms. Panels A1, ..., F1 show the improvement in the topic model fit over time. (The 1,000 EM iterations performed during the initialization phase are not shown.) Log-likelihoods $\log p_{\text{topic}}(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star)$ are shown relative to the best log-likelihood obtained among the four algorithms compared. Log-likelihood differences less than 0.01 are shown as 0.01. Circles are plotted at every 100 iterations. Plots A2, ..., F2 compare final estimates of the topic proportions for all topics. For extended results, see Figures A2–A9.

algorithms improve upon this initial fit. In addition to these quantitative performance comparisons, we also assessed whether different estimates qualitatively impact interpretation of the topics. The Appendix A.5 gives additional details on the experiment setup, and the companion repository [13] contains the source code used to generate the results presented here and in the Appendix.

Table 1: Data sets used in the experiments.

name	rows	columns	nonzeros	ref.
NeurIPS	2,483	14,036	3.7%	[28]
newsgroups	18,774	55,911	0.2%	[53]
trachea droplet	7,193	18,388	9.3%	[51]
68k PBMC	68,579	20,387	2.7%	[67]

In summary, our results showed that the extrapolated CD updates usually produced the best fit, and converged to a solution at least as fast as other algorithms, and often faster. Selected comparisons are shown in Fig. 2, with more comprehensive results (all four data sets, with K ranging from 2 to 12) included in the Appendix (Figures A2–A9). The extrapolation consistently helped convergence of CD updates, and often helped convergence of EM. The runtime per iteration was roughly the same in all algorithms (circles are plotted at every 100 iterations). Beyond these general patterns, there was considerable variation in the algorithms' performance among the different data sets, and within each data set at different settings of K. To make sense of the diverse results, we distinguish three main patterns.

Plots A1 and B1 in Fig. 2 illustrate the first pattern: EM achieved a reasonably good fit, so any improvement over EM was small regardless of the algorithm used. Indeed, the final EM and CD estimates in these two examples are virtually indistinguishable (Fig. 2, Panels A2 and B2).

Plots C1 and D1 illustrate the second pattern: 1,000 iterations of EM was insufficient to obtain a good fit, and running additional EM or CD updates substantially improved the fit. Among the four algorithms, the extrapolated CD updates provided the greatest improvement. And yet the large difference in log-likelihood produced only modest differences in estimated topic proportions (Plots C2

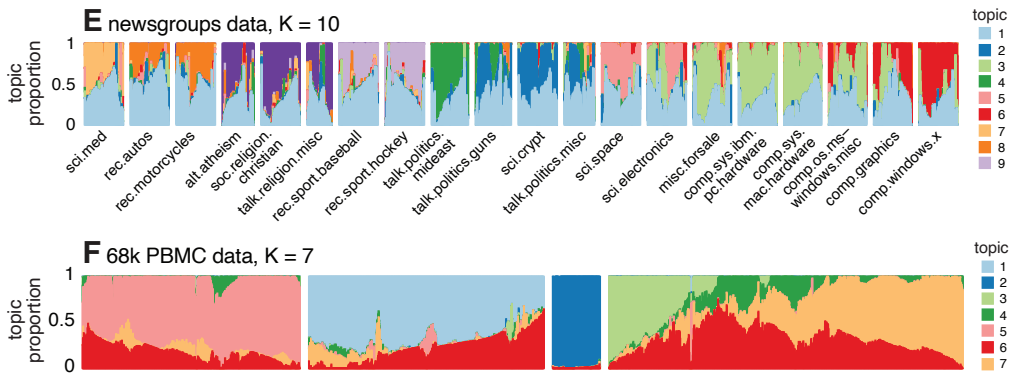

Figure 3: Extrapolated CD topic model fits visualized using Structure plots [55, 56]. Each bar represents a data sample, and bar heights represent topic proportions. In Panel E, the documents are arranged by newsgroup. To create the Structure plot in Panel F, the cells are split into 4 clusters based on their topic proportions; the clusters, from left to right, contain 3733, 4028, 163 and 60,655 cells. To better visualize the smaller clusters in Panel F, the larger clusters are subsampled in the plot.

and D2). Therefore, while the CD updates with extrapolation produced large gains in computational performance, these gains did not meaningfully impact the topic modeling results.

Plots E1 and F1, corresponding to the newsgroups word count and 68k PBMC RNA-seq data sets, are examples of the third pattern: the extrapolated CD updates outperformed the other updates, and produced estimated topic proportions strikingly different from the EM estimates (Plots E2, F2). To understand why, we examined the CD estimates of the topic proportions in Structure plots [55, 56] (Fig. 3). The topics with the greatest discrepancies between the EM and CD estimates are topic 1 in the newsgroups data and topics 3, 6 and 7 in the 68k PBMC data. Topic 1 in the newsgroups data and topic 6 in 68k PBMC data seem to capture global trends as they are present to some degree in most data samples. Topic 1 of the newsgroups is likely capturing "off-topic" discussion; none of the words appearing more frequently in topic 1 obviously point to any specific newsgroup (Fig. A10). In topic 6 of the 68k PBMC data, the genes with the largest expression increases are ribosomal protein genes (Fig. A11); abundance of ribosomal protein genes is known to be a major contributor to variation in PBMC scRNA-seq data sets [24]. Topics 3, 4, 6 and 7 capture heterogeneity in the largest cluster—the right-most cluster in Fig. 3F, mainly consisting of natural killer cells and T cells)—which brings to mind Scenario B in Sec. 4. By contrast, topics 1, 2 and 5, which identify myeloid cells, CD34+ cells and B cells, respectively, are mostly independent, and the EM and CD estimates for these topics do not differ much (Fig. 2, Panel F2). These results suggest that interdependent topics impede EM's convergence, and in these settings CD can offer large improvements.

6 Conclusions and discussion

Here, we suggested a simple new strategy for fitting topic models: first fit a Poisson NMF, then recover the corresponding topic model. We showed that this strategy works particularly well when Poisson NMF is optimized via co-ordinate descent.

We focussed on ML estimation, but the ideas and algorithms presented here also apply to MAP estimation with Dirichlet priors on $\mathbf{F}^\star$. Extending these ideas to improve variational EM (VEM) for topic models (*i.e.*, LDA) may also be of interest. As our examples demonstrate, simply initializing VEM from the solution found by our approach may already improve fits. However, given the success of the CD approach, it may also be fruitful to develop CD-based alternatives to VEM for Poisson NMF [31], which could then be translated into topic model fits. That said, in many applications topic models are mainly used for dimension reduction—the goal being to learn compact representations of complex patterns—and in these applications, ML or MAP estimation may suffice.

Finally, although we improved topic model fits compared with EM-like algorithms, some topic modeling applications involve massive data sets that require an "online" approach [35, 57]. It may be therefore fruitful to develop online versions of our algorithms. This is straightforward in principle, although online learning brings additional practical challenges, such as the choice of learning rates.

Acknowledgments and Disclosure of Funding

We are grateful for the many people who have contributed their ideas and have given feedback on our work, including Mihai Anitescu, Kushal Dey, Adam Gruenbaum, Anthony Hung, Youngseok Kim, Kaixuan Luo, John Novembre, Sebastian Pott, Alan Selewa and Jason Willwerscheid. We thank the staff at the University of Chicago Research Computing Center for providing the high-performance computing resources used to implement the numerical experiments. This work was supported by the NHGRI at the National Institutes of Health under award number 5R01HG002585.

References

[1] J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B*, 44(2):139–160, 1982.

[2] A. M. S. Ang and N. Gillis. Accelerating nonnegative matrix factorization algorithms using extrapolation. *Neural Computation*, 31(2):417–439, 2019.

[3] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 27–34, 2009.

[4] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.

[5] P. Bielecki, S. J. Riesenfeld, J.-C. Hütter, E. Torlai Triglia, M. S. Kowalczyk, R. R. Ricardo-Gonzalez, M. Lian, M. C. Amezcua Vesely, L. Kroehling, H. Xu, M. Slyper, C. Muus, L. S. Ludwig, E. Christian, L. Tao, A. J. Kedaigle, H. R. Steach, A. G. York, M. H. Skadow, P. Yaghoubi, D. Dionne, A. Jarret, H. M. McGee, C. B. M. Porter, P. Licona-Limón, W. Bailis, R. Jackson, N. Gagliani, G. Gasteiger, R. M. Locksley, A. Regev, and R. A. Flavell. Skin-resident innate lymphoid cells converge on a pathogenic effector state. *Nature*, 592:128–132, 2021.

[6] D. M. Blei and J. D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] C. Bouman and K. Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Transactions on Image Processing*, 5(3):480–492, 1996.

[9] W. Buntine. Variational extensions to EM and multinomial PCA. In *Proceedings of the 13th European Conference on Machine Learning*, pages 23–34, 2002.

[10] W. Buntine and A. Jakulin. Discrete component analysis. In M. Saunders, Craigand Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, latent structure and feature selection*, pages 1–33, Berlin, Heidelberg, 2006. Springer.

[11] J. Canny. GaP: a factor model for discrete data. In *Proceedings of the 27th Annual International ACM SIGIR Conference*, pages 122–129, 2004.

[12] P. Carbonetto, K. Luo, K. Dey, J. Hsiao, and M. Stephens. *fastTopics: fast algorithms for fitting topic models and non-negative matrix factorizations to count data*, 2021. URL `https://github.com/stephenslab/fastTopics`. R package version 0.5-24.

[13] P. Carbonetto, A. Sarkar, Z. Wang, and M. Stephens. Code and data accompanying this manuscript, May 2021. URL `https://github.com/stephenslab/fastTopics-experiments`.

[14] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009:785152, 2009.

[15] A. Cichocki, S. Cruces, and S.-I. Amari. Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization. *Entropy*, 13(1):134–170, 2011.

[16] J. E. Cohen and U. G. Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190(1977):149–168, 1993.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–22, 1977.

[18] K. K. Dey, C. J. Hsiao, and M. Stephens. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics*, 13(3):e1006599, 2017.

[19] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in Neural Information Processing Systems*, volume 18, pages 283–290, 2005.

[20] C. Ding, T. Li, and W. Peng. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, 52(8): 3913–3927, 2008.

[21] T. Faleiros and A. Lopes. On the equivalence between algorithms for non-negative matrix factorization and latent Dirichlet allocation. In *Proceedings of the 24th European Symposium on Artificial Neural Networks*, pages 171–176, 2016.

[22] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[23] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[24] S. Freytag, L. Tian, I. Lönnstedt, M. Ng, and M. Bahlo. Comparison of clustering tools in R for medium-sized 10x genomics single-cell RNA-sequencing data. *F1000Research*, 7:1297, 2018.

[25] E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *Proceedings of the 28th Annual International ACM SIGIR Conference*, pages 601–602, 2005.

[26] N. Gillis. The why and how of nonnegative matrix factorization. *arXiv*, 1401.5226, 2014.

[27] N. Gillis. *Nonnegative matrix factorization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.

[28] A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8:2265–2295, 2007.

[29] C. González-Blas, L. Minnoye, D. Papasokrati, S. Aibar, G. Hulselmans, V. Christiaens, K. Davie, J. Wouters, and S. Aerts. cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nature Methods*, 16(5):397–400, 2019.

[30] I. J. Good. Some statistical applications of Poisson's work. *Statistical Science*, 1(2):157–170, 1986.

[31] P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with hierarchical Poisson factorization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 326–335, 2015.

[32] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5228–5235, 2004.

[33] N. C. Henderson and R. Varadhan. Damped Anderson acceleration with restarts and monotonicity control for accelerating EM and EM-like algorithms. *Journal of Computational and Graphical Statistics*, 28(4):834–846, 2019.

[34] L. T. K. Hien and N. Gillis. Algorithms for nonnegative matrix factorization with the Kullback-Leibler divergence. *arXiv*, 1104.3889, 2020.

[35] M. Hoffman, F. Bach, and D. Blei. Online learning for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23, pages 856–864, 2010.

[36] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference*, pages 50–57, 1999.

[37] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[38] T. Hofmann, J. Puzicha, and M. Jordan. Learning from dyadic data. In *Advances in Neural Information Processing Systems*, volume 11, pages 466–472, 1999.

[39] C.-J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD International Conference*, pages 1064–1072, 2011.

[40] A. Hung, G. Housman, E. A. Briscoe, C. Cuevas, and Y. Gilad. Characterizing gene expression responses to biomechanical strain in an in vitro model of osteoarthritis. *bioRxiv*, 2021. doi: 10.1101/2021.02.22.432314.

[41] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58 (2):285–319, 2014.

[42] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography*, 8(2):306–316, 1984.

[43] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[44] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562, 2001.

[45] X. Lin and P. C. Boutros. Optimization and expansion of non-negative matrix factorization. *BMC Bioinformatics*, 21:7, 2020.

[46] L. Lucy. An iterative algorithm for the rectification of observed distributions. *The Astronomical Journal*, 79:745–754, 1974.

[47] P. McCullagh. *Generalized linear models*. Chapman and Hall, New York, 2nd edition, 1989.

[48] X.-L. Meng and D. Van Dyk. The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, 59(3):511–567, 1997.

[49] T. Minka and J. Lafferty. Expectation propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 352–359, 2002.

[50] R. Molina, J. Nunez, F. Cortijo, and J. Mateos. Image restoration in astronomy: a Bayesian perspective. *IEEE Signal Processing Magazine*, 18(2):11–29, 2001.

[51] D. T. Montoro, A. L. Haber, M. Biton, V. Vinarsky, B. Lin, S. E. Birket, F. Yuan, S. Chen, H. M. Leung, J. Villoria, N. Rogel, G. Burgin, A. M. Tsankov, A. Waghray, M. Slyper, J. Waldman, L. Nguyen, D. Dionne, O. Rozenblatt-Rosen, P. R. Tata, H. Mou, M. Shivaraju, H. Bihler, M. Mense, G. J. Tearney, S. M. Rowe, J. F. Engelhardt, A. Regev, and J. Rajagopal. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature*, 560(7718):319–324, 2018.

[52] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, New York, NY, 2nd edition, 2006.

[53] J. Rennie. 20 newsgroups data set. URL `http://qwone.com/~jason/20Newsgroups`.

[54] W. H. Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972.

[55] N. A. Rosenberg. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.

[56] N. A. Rosenberg. distruct: a program for the graphical display of population structure. *Molecular Ecology Notes*, 4(1):137–138, 2004.

[57] M.-A. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13 (7):1649–1681, 2001.

[58] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1(2):113–122, 1982.

[59] A. P. Singh and G. J. Gordon. A unified view of matrix factorization models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, volume 2, pages 358–373, 2008.

[60] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent semantic analysis: a road to meaning*, pages 427–448. Lawrence Erlbaum, 2006.

[61] M. Taddy. On estimation and selection for topic models. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 1184–1193, La Palma, Canary Islands, 2012. PMLR.

[62] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 19, pages 1353–1360, 2007.

[63] Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985.

[64] H. M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 977–984, 2006.

[65] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015.

[66] H. Xu, J. Ding, C. B. Porter, A. Wallrapp, M. Tabaka, S. Ma, S. Fu, X. Guo, S. J. Riesenfeld, C. Su, D. Dionne, L. T. Nguyen, A. Lefkovith, O. Ashenberg, P. R. Burkett, H. N. Shi, O. Rozenblatt-Rosen, D. B. Graham, V. K. Kuchroo, A. Regev, and R. J. Xavier. Transcriptional atlas of intestinal immune cells reveals that neuropeptide α-CGRP modulates group 2 innate lymphoid cell responses. *Immunity*, 51(4):696–708, 2019.

[67] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.

[68] H. Zhou, D. Alexander, and K. Lange. A quasi-Newton acceleration for high-dimensional optimization algorithms. *Statistics and Computing*, 21(2):261–273, 2011.

[69] M. Zhou, L. Hannah, D. Dunson, and L. Carin. Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 1462–1471, 2012.

A Appendix

The Appendix contains additional methods, results and discussion supporting the paper, "Non-negative matrix factorization algorithms greatly improve topic model fits," including derivations of the EM algorithms (Sec. A.2); derivation of the KKT conditions for Poisson NMF (Sec. A.3); enhancements and implementation details for the Poisson NMF algorithms (Sec. A.4); details of the numerical experiments (Sec. A.5); and extended results of the numerical experiments (Sec. A.6).

A.1 Extension of Corollary 1 to MAP estimation

Corollary 2 (Relation between MAP estimates for Poisson NMF and multinomial topic model)**.** *Let* $\hat{\mathbf{L}} \in \mathbf{R}_+^{n\times K}, \hat{\mathbf{F}} \in \mathbf{R}_+^{m\times K}$ *denote* maximum a posteriori *(MAP) estimates for the Poisson NMF model, in which the elements of* $\mathbf{F}$ *are assigned independent gamma priors,* $f_{jk} \sim \mathrm{Gamma}(\alpha_{jk}, \beta_k)$*, with* $\alpha_{jk} > 1$ *for all* $j = 1, \ldots, m$*,* $k = 1, \ldots, K$*, and* $\mathbf{L}$ *is assigned an (improper) uniform prior,*

$$\hat{\mathbf{L}}, \hat{\mathbf{F}} \in \underset{\mathbf{L},\mathbf{F}}{\mathrm{argmax}}\, p_{\mathsf{PNMF}}(\mathbf{X} \mid \mathbf{L}, \mathbf{F}) \times \prod_{j=1}^{m}\prod_{k=1}^{K} \mathrm{Gamma}(f_{jk}; \alpha_{jk}, \beta_k), \tag{16}$$

where $\mathrm{Gamma}(\theta; \alpha, \beta)$ *denotes the probability density of the gamma distribution with shape* α *and inverse scale* β*. If* $\hat{\mathbf{L}}^\star, \hat{\mathbf{F}}^\star$ *are obtained by applying* ψ *to* $\hat{\mathbf{L}}, \hat{\mathbf{F}}$*, these will be MAP estimates for the multinomial topic model with independent Dirichlet priors on the columns of* $\mathbf{F}$*,* $f_{1k}, \ldots, f_{mk} \sim \mathrm{Dirichlet}(\alpha_{1k}, \ldots, \alpha_{mk})$*, and a uniform prior on* $\mathbf{L}^\star$*,*

$$\hat{\mathbf{L}}^\star, \hat{\mathbf{F}}^\star \in \underset{\mathbf{L}^\star,\mathbf{F}^\star}{\mathrm{argmax}}\, p_{\mathsf{topic}}(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star) \times \prod_{k=1}^{K} \mathrm{Dirichlet}(f_{1k}, \ldots, f_{mk}; \alpha_{1k}, \ldots, \alpha_{mk}). \tag{17}$$

Conversely, let $\hat{\mathbf{L}}^\star \in \mathbf{R}_+^{n\times K}, \hat{\mathbf{F}}^\star \in \mathbf{R}_+^{m\times K}$ *denote MAP estimates for the multinomial topic model* (17)*, set* $s_i = t_i = \sum_{i=1}^m x_{ij}$*, for* $i = 1, \ldots, n$*, and set* $u_k = \sum_{j=1}^m (\alpha_{jk} - 1)/\beta_k$*, for* $k = 1, \ldots, K$*. Then if* $\hat{\mathbf{L}}, \hat{\mathbf{F}}$ *are obtained by applying the inverse transformation* ψ^{-1} *to* $\hat{\mathbf{L}}^\star, \hat{\mathbf{F}}^\star, \boldsymbol{s}, \boldsymbol{u}$*, these will be MAP estimates for Poisson NMF* (16)*.*

A.2 EM algorithms

A.2.1 EM for the additive Poisson regression model

Here we derive the standard EM algorithm [6, 20, 24, 19, 25, 26, 30, 33] for fitting the additive Poisson regression model (9). First, we introduce latent variables $z_{ik} \sim \mathrm{Poisson}(a_{ik}b_k)$ such that $\sum_{k=1}^K z_{ik} = y_i$, $i = 1, \ldots, n$. Under the model augmented with the z_{ik}'s, the expected complete log-likelihood is

$$E[\log p(\boldsymbol{y}, \boldsymbol{z} \mid \mathbf{A}, \boldsymbol{b})] = \sum_{i=1}^{n}\sum_{k=1}^{K} \phi_{ik} \log(a_{ik}b_k) - \sum_{i=1}^{n}\sum_{k=1}^{K} a_{ik}b_k + \text{const}, \tag{18}$$

where "const" includes additional terms in the likelihood that do not depend on $\boldsymbol{b}$, and $\bar{z}_{ik} = E[z_{ik} \mid \boldsymbol{b}]$ is the expectation of z_{ik} with respect to the posterior $p(\boldsymbol{z} \mid \mathbf{A}, \boldsymbol{b})$. The M step (11) is derived by taking the partial derivative of (18) with respect to b_k, and solving for b_k. The E step involves computing posterior expectations at the current $\boldsymbol{b} = (b_1, \ldots, b_K)$. The posterior distribution of $z_i = (z_{i1}, \ldots, z_{iK})$ is multinomial with y_i trials and multinomial probabilities $p_{ik} \propto a_{ik}b_k$. Therefore, the posterior expected value of z_{ik} is

$$\bar{z}_{ik} = y_i p_{ik} = y_i a_{ik} b_k / \mu_i. \tag{19}$$

The EM algorithm consists of iterating the E (10) and M (11) steps until some stopping criterion is met. Alternatively, the E and M steps can be combined, yielding the update

$$b_k^{\text{new}} \leftarrow b_k \times \frac{\sum_{i=1}^n a_{ik} y_i/\mu_i}{\sum_{i=1}^n a_{ik}}. \tag{20}$$

A.2.2 EM for Poisson NMF

Here we derive an EM algorithm [5, 11] for Poisson NMF by applying the EM updates for the additive Poisson regression model that were derived in Appendix A.2.1. Making substitutions $\mathbf{A} \to \mathbf{F}$, $\boldsymbol{y} \to \boldsymbol{x}_i$, $\boldsymbol{b} \to \boldsymbol{l}_i$ in (20), where $\boldsymbol{x}_i$ is a row of $\mathbf{X}$ and $\boldsymbol{l}_i$ is a row of $\mathbf{L}$, and appropriately changing the indices, we arrive at the following update for the loadings:

$$l_{ik}^{\text{new}} \leftarrow l_{ik} \times \frac{\sum_{j=1}^m x_{ij} f_{jk}/\lambda_{ij}}{\sum_{j=1}^m f_{jk}} \tag{21}$$

Similarly, making substitutions $\mathbf{A} \to \mathbf{L}$, $\boldsymbol{y} \to \boldsymbol{x}_j$, $\boldsymbol{b} \to \boldsymbol{f}_j$, where $\boldsymbol{x}_j$ is a column of $\mathbf{X}$ and $\boldsymbol{f}_j$ is a row of $\mathbf{F}$, the update for the factors is

$$f_{jk}^{\text{new}} \leftarrow f_{jk} \times \frac{\sum_{i=1}^{n} x_{ij} l_{ik} / \lambda_{ij}}{\sum_{i=1}^{n} l_{ik}}. \tag{22}$$

These are the Poisson NMF "multiplicative" updates [22].

A.2.3 EM for the multinomial topic model

Here we derive EM for the multinomial topic model [2, 15], and draw a connection between EM for the multinomial topic model and the multiplicative updates for Poisson NMF.

The EM algorithm is based on the following augmented model [3]:

$$\begin{aligned} p(z_{ir} = k \mid l_i^\star) &= l_{ik}^\star \\ p(w_{ir} = j \mid \mathbf{F}^\star, z_{ir} = k) &= f_{jk}^\star, \end{aligned} \tag{23}$$

where we have introduced latent topic assignments $z_{ir} \in \{1, \ldots, K\}$, and the data are encoded as $w_{ir} \in \{1, \ldots, m\}$, for $r = 1, \ldots, t_i$, with t_i being the size of the ith document. Summing over the topic assignments z_{ij}, the augmented model (23) recovers the multinomial likelihood (3) up to a constant of proportionality; that is,

$$p(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star) \propto \int p(\boldsymbol{w}, \boldsymbol{z} \mid \mathbf{L}^\star, \mathbf{F}^\star)\, d\boldsymbol{z} = \prod_{i=1}^{n} \prod_{r=1}^{t_i} \sum_{z_{ir}=1}^{K} p(w_{ir} \mid \mathbf{F}^\star, z_{ir})\, p(z_{ir} \mid \boldsymbol{l}_i^\star),$$

in which the word counts in each document are recovered as $x_{ij} = \sum_{r=1}^{t_i} \delta_j(w_{ir})$

The E step consists of computing posterior expectations for the latent topic assignments z_{ij},

$$p_{ijk} \equiv p(z_{ij} = k \mid \mathbf{X}, \mathbf{L}^\star, \mathbf{F}^\star) = l_{ik}^\star f_{jk}^\star / \pi_{ij}. \tag{24}$$

The M step updates for the topic proportions $l_{ik}^\star$ and word frequencies $f_{jk}^\star$ are

$$l_{ik}^\star = \sum_{j=1}^{m} x_{ij} p_{ijk} / t_i \tag{25}$$

$$f_{jk}^\star \propto \sum_{i=1}^{n} x_{ij} p_{ijk}. \tag{26}$$

Combining the E (24) and M (25, 26) steps, we arrive at the following updates:

$$(l_{ik}^\star)^{\text{new}} \leftarrow \frac{l_{ik}^\star}{t_i} \sum_{j=1}^{m} x_{ij} f_{jk}^\star / \pi_{ij} \tag{27}$$

$$(f_{jk}^\star)^{\text{new}} \leftarrow \frac{f_{jk}^\star}{\xi_k} \sum_{i=1}^{n} x_{ij} l_{ik}^\star / \pi_{ij}, \tag{28}$$

where ξ_k is a normalizing constant ensuring that $(f_{1k}^\star)^{\text{new}} + \cdots + (f_{mk}^\star)^{\text{new}} = 1$.

The same EM updates (27, 28) can also be derived in a different way, by applying the reparameterization (4) and its inverse to the Poisson NMF multiplicative updates (12, 13).

A.2.4 EM for the multinomial mixture model

The multinomial mixture model is

$$\begin{aligned} y_1, \ldots, y_n &\sim \text{Multinomial}(t, \pi_1, \ldots, \pi_n), \\ \pi_i &= a_{i1}^\star b_1^\star + \cdots + a_{iK}^\star b_K^\star, \end{aligned} \tag{29}$$

in which the data are $\mathbf{A}^\star \in \mathbf{R}_+^{n \times K}$ and $\boldsymbol{y} = (y_1, \ldots, y_n) \in \mathbf{R}_+^n$, and $t = \sum_{i=1}^{n} y_i$. To ensure that the π_i's are indeed probabilities, we require $b_k^\star \geq 0$, $a_{ik}^\star \geq 0$, $\sum_{k=1}^{K} b_k^\star = 1$ and $\sum_{i=1}^{n} a_{ik}^\star = 1$.

Omitting the derivation, the E and M steps for the multinomial mixture model (29) are

$$p_{ik} = \frac{a^\star_{ik} b^\star_k}{\sum_{k'=1}^K a^\star_{ik'} b^\star_{k'}} \tag{30}$$

$$b^\star_k = \sum_{i=1}^n y_i p_{ik}/t. \tag{31}$$

A.2.5 Poisson regression–multinomial mixture model reparameterization

The additive Poisson regression model (9) is equivalent to the multinomial mixture model (29) by a simple reparameterization. For practical implementation (with finite precision arithmetic), the EM updates for the multinomial mixture model are more convenient, so we use the following reparameterization to implement EM for the additive Poisson regression model.

The multinomial mixture model (29) is a reparameterization of the Poisson regression model (9) that preserves the likelihood; that is,

$$\text{Multinomial}(\boldsymbol{y}; t, \pi_1, \ldots, \pi_n) \times \text{Poisson}(t; s) = \prod_{i=1}^n \text{Poisson}(y_i; \mu_i), \tag{32}$$

in which the parameters on the left-hand side are recovered from the parameters on the right-hand side as

$$\begin{aligned} u_k &\leftarrow a_{1k} + \cdots + a_{nk} \\ s &\leftarrow b_1 u_1 + \cdots + b_K u_K \\ a^\star_{ik} &\leftarrow a_{ik}/u_k \\ b^\star_k &\leftarrow b_k u_k / s. \end{aligned} \tag{33}$$

Once the multinomial parameters $b^\star_1, \ldots, b^\star_K$ have been updated by performing one or more EM updates, the Poisson parameters $b_1, \ldots, b_K$ are recovered as $b_k = t b^\star_k / u_k$ using the MLE of s, $\hat{s} = t$.

A.3 Derivation of KKT conditions for Poisson NMF optimization problem

Here we derive the first-order optimality (Karush-Kuhn-Tucker) conditions for the Poisson NMF optimization problem (7). The first-order KKT conditions for this optimization problem are

$$\nabla_{\mathbf{F}} \hat{\ell}(\mathbf{L}, \mathbf{F}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) = 0 \tag{34}$$

$$\nabla_{\mathbf{L}} \hat{\ell}(\mathbf{L}, \mathbf{F}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) = 0 \tag{35}$$

$$\boldsymbol{\Omega} \odot \mathbf{F} = 0 \tag{36}$$

$$\boldsymbol{\Gamma} \odot \mathbf{L} = 0, \tag{37}$$

in which $\hat{\ell}(\mathbf{L}, \mathbf{F}, \boldsymbol{\Gamma}, \boldsymbol{\Omega})$ is the Lagrangian function,

$$\hat{\ell}(\mathbf{L}, \mathbf{F}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}) = \ell(\mathbf{L}, \mathbf{F}) - \sum_{i=1}^n \sum_{k=1}^K \gamma_{ik} l_{ik} - \sum_{j=1}^m \sum_{k=1}^K \omega_{jk} f_{jk}. \tag{38}$$

Here we have introduced matrices of Lagrange multipliers $\boldsymbol{\Gamma} \in \mathbf{R}_+^{n \times K}$ and $\boldsymbol{\Omega} \in \mathbf{R}_+^{m \times K}$ associated with the non-negativity constraints $\mathbf{L} \geq 0$ and $\mathbf{F} \geq 0$. Combining conditions (34) and (35), we obtain

$$\boldsymbol{\Omega} = (1 - \mathbf{U})^T \mathbf{L} \tag{39}$$

$$\boldsymbol{\Gamma} = (1 - \mathbf{U}) \mathbf{F}, \tag{40}$$

in which $\mathbf{U} \in \mathbf{R}_+^{n \times m}$ with entries $u_{ij} = x_{ij}/(\mathbf{L}\mathbf{F}^T)_{ij}$. See [7, 10] for generalizations of these conditions.

A.4 Additional enhancements and implementation details

Here we describe additional enhancements to the Poisson NMF algorithms and other implementation details.

To obtain a good initialization of $\mathbf{L}$ and $\mathbf{F}$, we used the Topic-SCORE algorithm [17] to estimate $\mathbf{F}$, then we performed 10 CD updates of $\mathbf{L}$. We found that the Topic-SCORE algorithm was very fast, and typically completed in less time than it took to run a single outer-loop iteration in Algorithm 1).

Following common practice in co-ordinate descent algorithms, we incompletely solved each subproblem FIT-POIS-REG$(\mathbf{L}, \boldsymbol{x}_j)$ and FIT-POIS-REG$(\mathbf{F}, \boldsymbol{x}_i)$; the intuition is that accurately solving each subproblem is wasted effort, particularly early on when the estimates change a lot from one iteration to the next. We found that 4 EM or 4 CD updates worked well when initialized to the estimate obtained in the previous (outer loop) iteration. Incompletely solving the subproblems has been similarly shown to work well for Frobenius-norm NMF [12, 16, 18].

Whenever we used CD to optimize FIT-POIS-REG$(\mathbf{A}, \boldsymbol{y})$, we performed a single EM update prior to running the CD updates. This nudged the estimate closer to the solution, and improved convergence of CD (in the absence of a line search).

To help with convergence of the updates to $\mathbf{L}$ and $\mathbf{F}$, any updated parameters that fell below 10^{-15} were set to this value. This step was motivated by Theorem 1 of [12].

To improve numerical stability of the updates, as others have done (e.g., [21]), we rescaled $\mathbf{F}$ and $\mathbf{L}$ after each full update. Specifically, we rescaled the matrices so that the column means of $\mathbf{F}$ were equal to the column means of $\mathbf{L}$. Note that the Poisson rates $\lambda_{ij} = (\mathbf{L}\mathbf{F}^T)_{ij}$ and the likelihood $p_{\mathsf{PNMF}}(\mathbf{X} \mid \mathbf{L}, \mathbf{F})$ are invariant to this rescaling.

We used Intel Threading Building Blocks (TBB) multithreading to optimize the FIT-POIS-REG$(\mathbf{F}, \boldsymbol{x}_i)$ and FIT-POIS-REG$(\mathbf{L}, \boldsymbol{x}_j)$ subproblems in parallel. We also used sparse matrix computation techniques to leverage sparsity of the count data. This latter substantially reduced computation for all data sets, which displayed high levels of sparsity (>90% of entries equal to zero). For sparse $\mathbf{X}$, the complexity of computing the Poisson NMF likelihood, for example, is $O((N + n + m)K)$, where N is the number of nonzeros in $\mathbf{X}$, whereas for non-sparse $\mathbf{X}$ the complexity is $O(nmK)$.

Finally, we used two measures to assess the quality of the computed solutions to (7): the change in the loss function $\ell(\mathbf{L}, \mathbf{F})$ or, equivalently, the change in the Poisson NMF log-likelihood, $\log p_{\mathsf{PNMF}}(\mathbf{X} \mid \mathbf{L}, \mathbf{F})$; and the maximum residual of the KKT conditions (39, 40).

A.5 Details of numerical experiments

These sections give additional details on preparation of the data sets (Sec. A.5.1), the computing setup (Sec. A.5.2), and software and code used (Sec. A.5.3) for the numerical experiments (Sec. 5).

A.5.1 Data sets

See Table 1 for summary of data sets used in the numerical experiments. The NeurIPS [13] and newsgroups [29] data sets are word counts extracted from, respectively, 1988–2003 NeurIPS (formerly NIPS) papers and posts to 20 different newsgroups. Both data sets have been used to evaluate topic modeling methods (e.g., [2, 34]). The trachea droplet [27] and 68k PBMC [35] data are UMI (unique molecular identifier) counts from droplet-based single-cell RNA sequencing experiments in trachea epithelial cells in C57BL/6 mice and in "unsorted" human peripheral blood mononuclear cells (PBMCs) [Fresh 68k PBMC Donor A], processed using the GemCode Single Cell Platform. The 68k PBMC data have been used to benchmark methods for single-cell RNA-seq data (e.g., [1, 8, 32]). The data sets were retrieved from `http://ai.stanford.edu/~gal/data.html` (NeurIPS), `http://qwone.com/~jason/20Newsgroups` (newsgroups), `https://support.10xgenomics.com/single-cell-gene-expression/datasets`, (68k PBMC) and the Gene Expression Omnibus (GEO), accession GSE103354 (trachea droplet).

The trachea droplet data are publicly available on the GEO website and as such follow the NCBI copyright and data usage policies (`https://www.ncbi.nlm.nih.gov/home/about/policies`); see also `https://www.ncbi.nlm.nih.gov/geo/info/disclaimer.html` for GEO data disclaimers. The 68k PBMC data are distributed under the CC BY 4.0 license. The 68k PBMC and trachea droplet data were obtained under IRB-approved protocols; see [27, 35] for details. The NeurIPS and newsgroups data sets are distributed without a license.[3]

[3] Private communications with Jason Rennie and Gal Chechik.

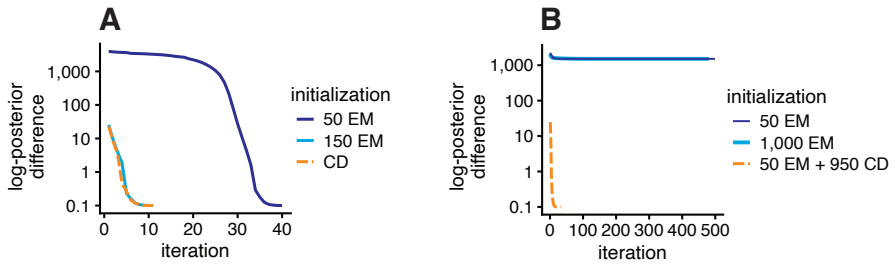

Figure A1: *Maximum a posteriori* (MAP) estimation of the multinomial topic model parameters, with $K = 6$, using the quasi-Newton-accelerated EM algorithm implemented in `maptpx` [31]. Plots A and B shows the improvement in the log-posterior over time in Scenarios A and B, in which the MAP estimates are initialized to MLEs obtained by performing different numbers of EM and/or CD updates.

All data sets were stored as sparse $n \times m$ count matrices $\mathbf{X}$, where n is the number of documents or cells, and m is the number of words or genes. The scripts used to prepare the data from the source files, as well as the prepared NeurIPS and newsgroups data sets, are included in the companion repository [13].

A.5.2 Computing environment

All computations on real data sets were run in R 3.5.1 [28], linked to the OpenBLAS 0.2.19 optimized numerical libraries, on Linux machines (Scientific Linux 7.4) with Intel Xeon E5-2680v4 ("Broadwell") processors. For performing the Poisson NMF optimization, which includes some multithreaded computations, 8 CPUs and 16 GB of memory were used.

A.5.3 Source code and software

The methods were implemented in the `fastTopics` R package [12] (the results were generated using version 0.5-24 of the package). The core optimization algorithms were developed in C++ and interfaced to R using `Rcpp` [9]. The CD updates were adapted from the C++ code included with R package NNLM 0.4-3 [23]. For the `maptpx` results, we used a slightly modified version of the `maptpx` package (version 1.9-8), available at `https://github.com/stephenslab/maptpx`, that allows initialization of both the topic proportions as well as the word frequencies. For the variational EM results, we used the C implementation by Blei *et al* (`http://www.cs.columbia.edu/~blei/lda-c`), which was interfaced to R using the `topicmodels` package [14]. The git repository [13] contains code implementing the numerical experiments, and includes a workflowr website [4] for browsing the results.

A.6 Extended results

Figure A1 shows the results from running `maptpx` on the two simulated data sets (Sec. 4).

Figures A2–A9 give more detailed results on the Poisson NMF algorithms' progress in fitting topic models to the real data sets (Sec. 5). The scatterplots in Figures A10 and A11 highlight words and genes that appear more frequently in one topic compared to the other topics.

References in Appendix

[1] T. Abdelaal, L. Michielsen, D. Cats, D. Hoogduin, H. Mei, M. J. T. Reinders, and A. Mahfouz. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biology*, 20:194, 2019.

[2] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 27–34, 2009.

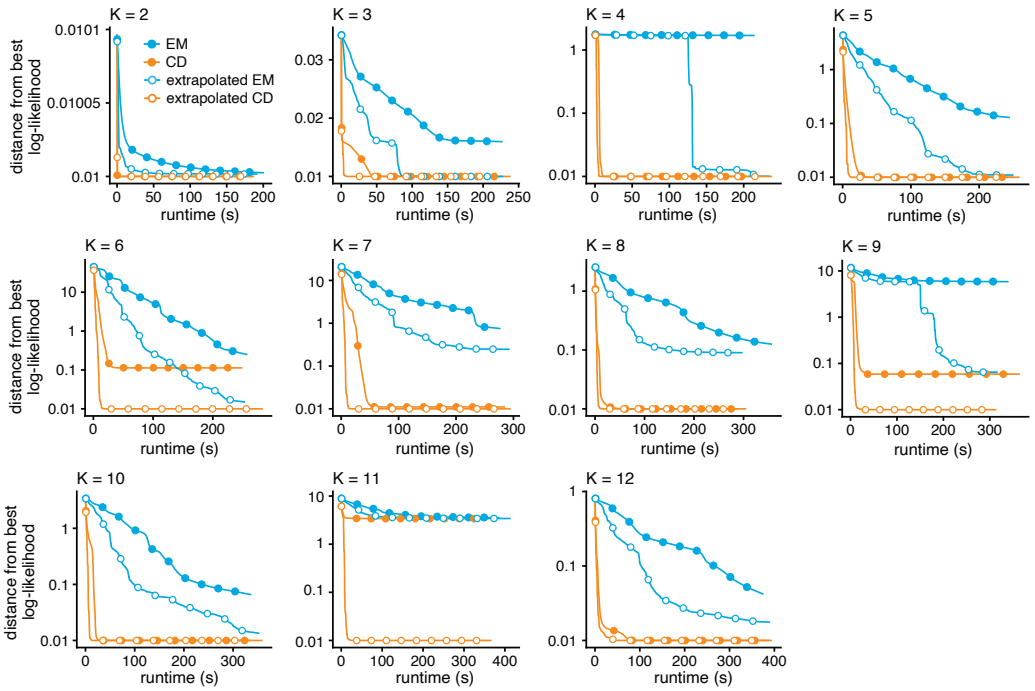

Figure A2: Improvement in model fit over time for the Poisson NMF algorithms applied to the NeurIPS data, with $K = 2, \ldots, 12$. Multinomial topic model log-likelihoods $\log p_{\text{topic}}(\mathbf{X} \mid \mathbf{L}^\star, \mathbf{F}^\star)$ are shown relative to the best log-likelihood recovered among the four algorithms compared (EM and CD, with and without extrapolation). The 1,000 EM iterations performed during the initialization phase are not shown. Log-likelihood differences less than 0.01 are shown as 0.01. Circles are drawn at intervals of 100 iterations.

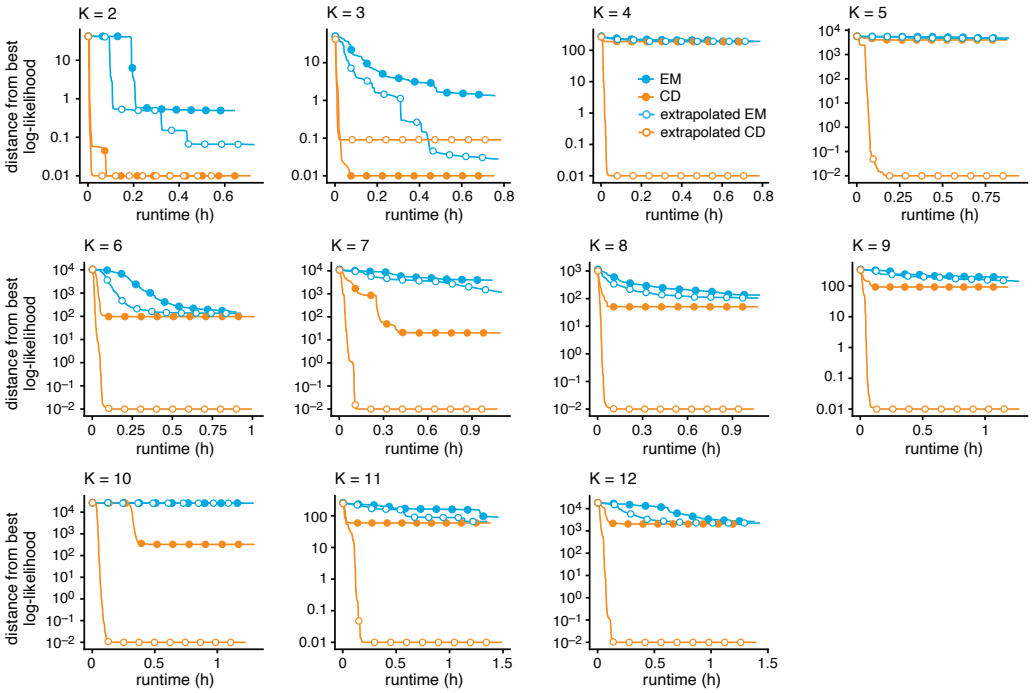

Figure A3: Improvement in model fit over time for the Poisson NMF algorithms applied to the newsgroups data, with $K = 2, \ldots, 12$. See the Fig. A2 caption for details.

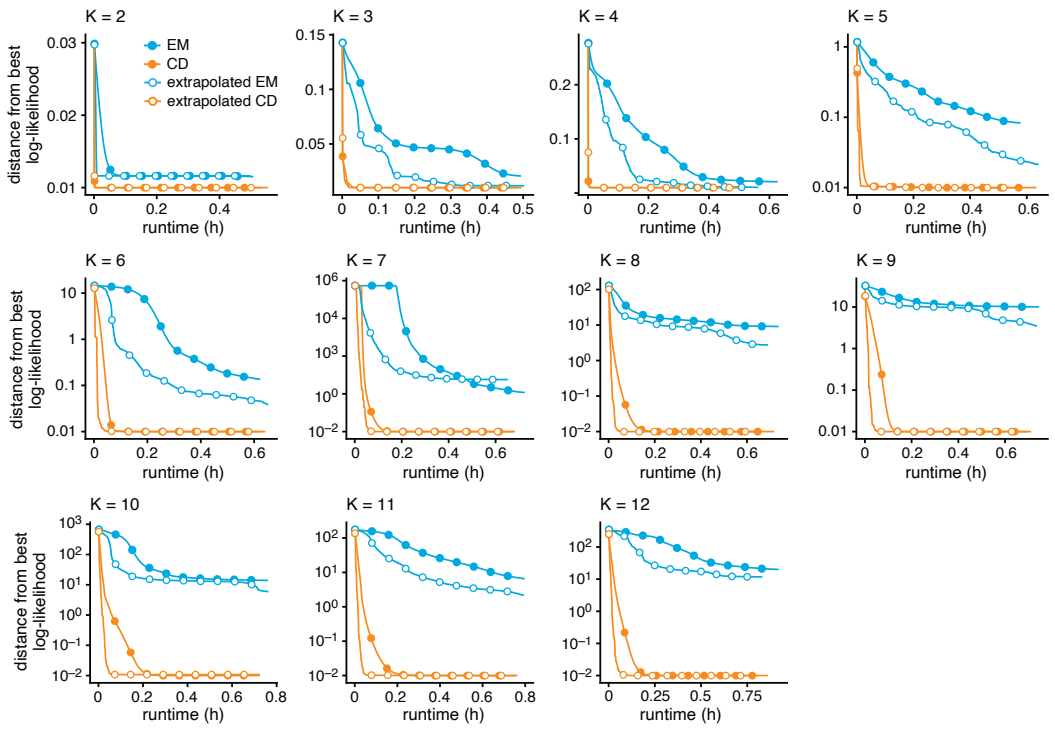

Figure A4: Improvement in model fit over time for the Poisson NMF algorithms applied to the trachea droplet data, with $K = 2, \ldots, 12$. See the Fig. A2 caption for details.

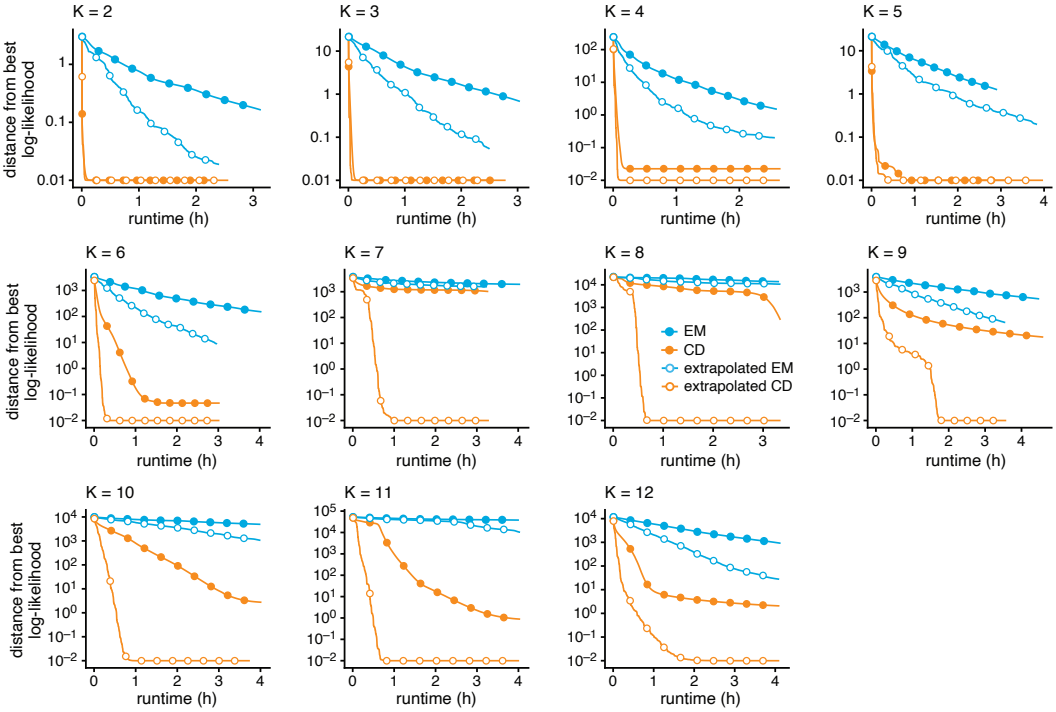

Figure A5: Improvement in model fit over time for the Poisson NMF algorithms applied to the trachea droplet data, with $K = 2, \ldots, 12$. See the Fig. A2 caption for details.

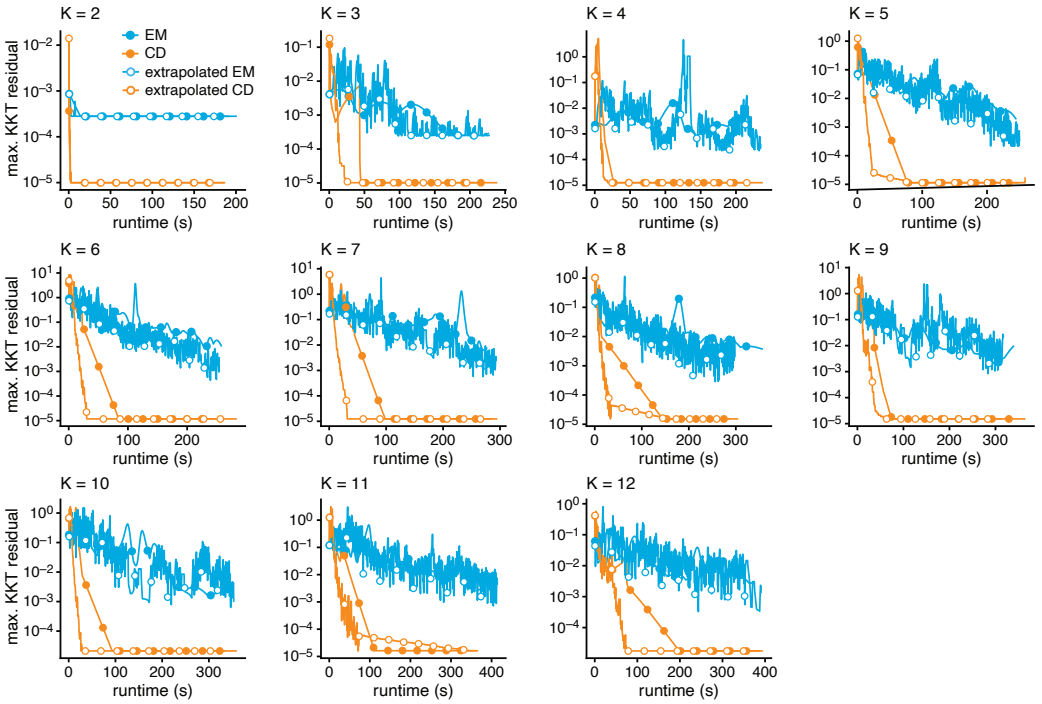

Figure A6: Evolution of the KKT residuals over time for the Poisson NMF algorithms applied to the NeurIPS data, with $K = 2, \ldots, 12$. The KKT residuals should vanish near a local maximum of the likelihood, so looking at the largest KKT residual can be used to assess how well the algorithm recovers an MLE. Note that, unlike the likelihood, the KKT residuals are not expected to decrease monotonically over time. Circles are drawn at intervals of 100 iterations.

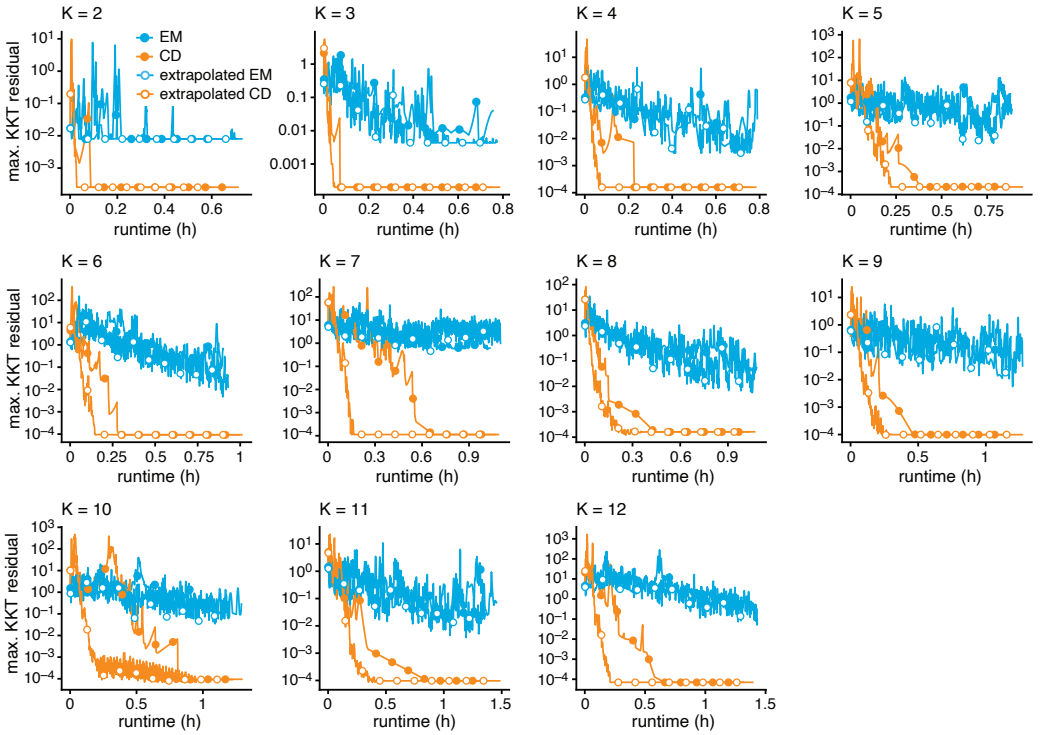

Figure A7: Evolution of the KKT residuals over time for the Poisson NMF algorithms applied to the newsgroups data, with $K = 2, \ldots, 12$. See the Fig. A6 caption for details.

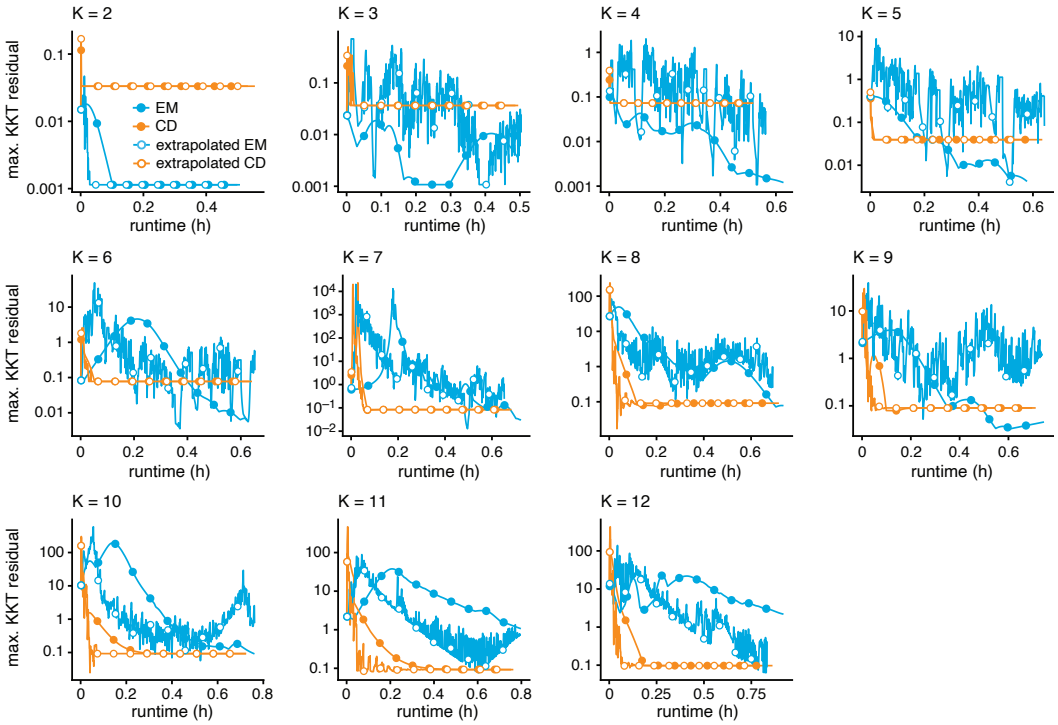

Figure A8: Evolution of the KKT residuals over time for the Poisson NMF algorithms applied to the trachea droplet data, with $K = 2, \ldots, 12$.

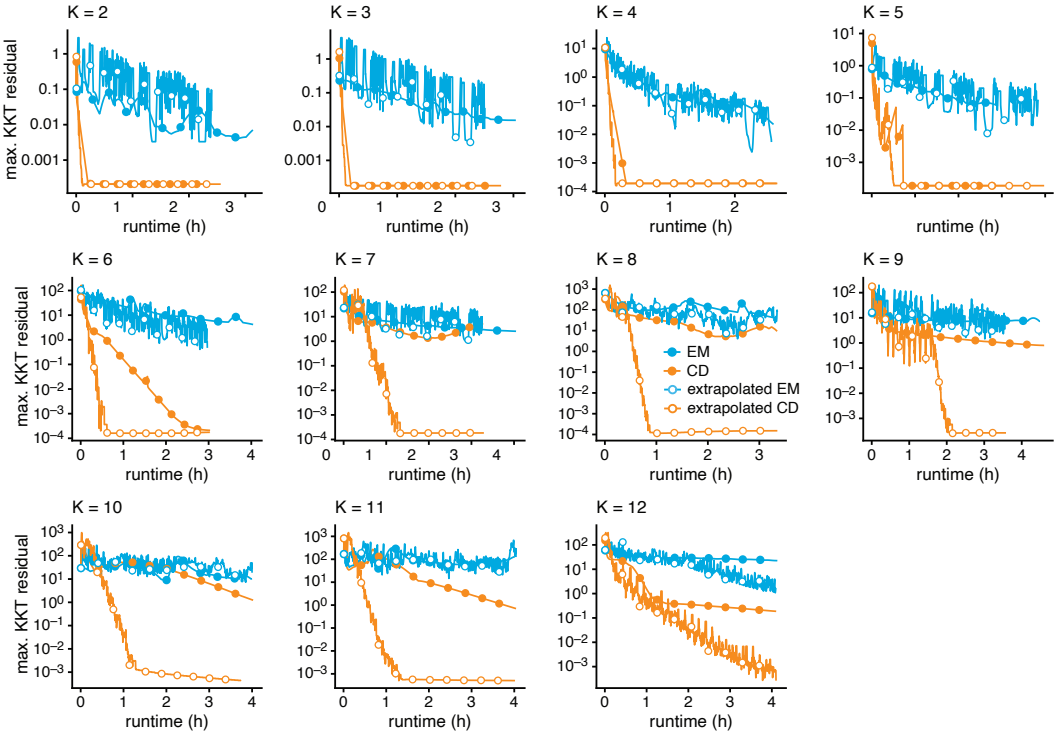

Figure A9: Evolution of the KKT residuals over time for the Poisson NMF algorithms applied to the 68k PBMC data, with $K = 2, \ldots, 12$.

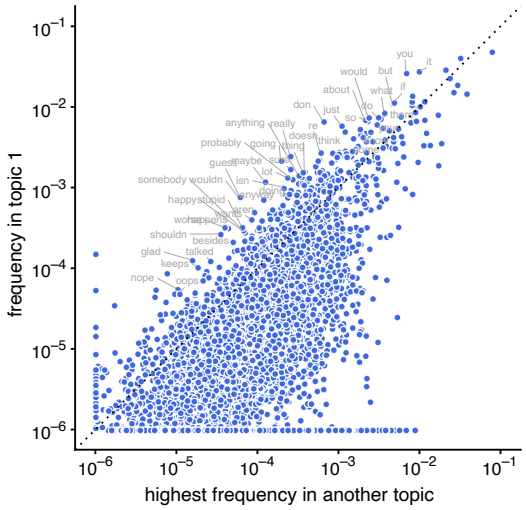

Figure A10: Word frequencies in topic 1 vs. other topics in the multinomial topic model estimated from the newsgroups data using extrapolated CD updates, with $K = 10$. Words that are more frequent in topic 1 compared to other topics lie above the diagonal. Frequencies less than 10^{-6} are shown as 10^{-6}.

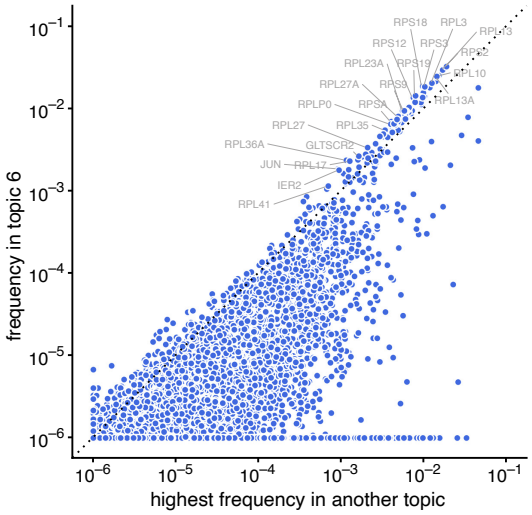

Figure A11: Gene frequencies in topic 6 vs. other topics in the multinomial topic model estimated from the 68k PBMC data using extrapolated CD updates, with $K = 7$. Genes that have higher frequencies in topic 6 compared to other topics lie above the diagonal. Frequencies less than 10^{-6} are shown as 10^{-6}.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. D. Blischak, P. Carbonetto, and M. Stephens. Creating and sharing reproducible research code the workflowr way [version 1; peer review: 3 approved]. *F1000Research*, 8(1749), 2019.

[5] A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009:785152, 2009.

[6] A. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Transactions on Medical Imaging*, 12(2):328–333, 1993.

[7] I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in Neural Information Processing Systems*, volume 18, pages 283–290, 2005.

[8] J. J. Diaz-Mejia, E. C. Meng, A. R. Pico, S. A. MacParland, T. Ketela, T. J. Pugh, G. D. Bader, and J. H. Morris. Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Research*, 8(296), 2019.

[9] D. Eddelbuettel and R. François. Rcpp: seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[10] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, 2011.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[12] N. Gillis and F. Glineur. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4):1085–1105, 2012.

[13] A. Globerson, G. Chechik, F. Pereira, and N. Tishby. Euclidean embedding of co-occurrence data. *Journal of Machine Learning Research*, 8:2265–2295, 2007.

[14] B. Grün and K. Hornik. topicmodels: an R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011.

[15] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[16] C.-J. Hsieh and I. S. Dhillon. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In *Proceedings of the 17th ACM SIGKDD International Conference*, pages 1064–1072, 2011.

[17] Z. T. Ke and M. Wang. A new svd approach to optimal topic estimation. *arXiv*, 1704.07016, 2019.

[18] J. Kim, Y. He, and H. Park. Algorithms for nonnegative matrix and tensor factorizations: a unified view based on block coordinate descent framework. *Journal of Global Optimization*, 58 (2):285–319, 2014.

[19] T. Krishnan. EM algorithm in tomography: a review and a bibliography. *Bulletin of Informatics and Cybernetics*, 27:5–22, 1995.

[20] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *Journal of Computer Assisted Tomography*, 8(2):306–316, 1984.

[21] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[22] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13, pages 556–562, 2001.

[23] X. Lin and P. C. Boutros. *NNLM: fast and versatile non-negative matrix factorization*, 2019. URL `http://CRAN.R-project.org/package=NNLM`.

[24] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley-Interscience, Hoboken, NJ, 2008.

[25] X.-L. Meng and D. Van Dyk. The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, 59(3):511–567, 1997.

[26] R. Molina, J. Nunez, F. Cortijo, and J. Mateos. Image restoration in astronomy: a Bayesian perspective. *IEEE Signal Processing Magazine*, 18(2):11–29, 2001.

[27] D. T. Montoro, A. L. Haber, M. Biton, V. Vinarsky, B. Lin, S. E. Birket, F. Yuan, S. Chen, H. M. Leung, J. Villoria, N. Rogel, G. Burgin, A. M. Tsankov, A. Waghray, M. Slyper, J. Waldman, L. Nguyen, D. Dionne, O. Rozenblatt-Rosen, P. R. Tata, H. Mou, M. Shivaraju, H. Bihler, M. Mense, G. J. Tearney, S. M. Rowe, J. F. Engelhardt, A. Regev, and J. Rajagopal. A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature*, 560(7718):319–324, 2018.

[28] *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org`.

[29] J. Rennie. 20 newsgroups data set. URL `http://qwone.com/~jason/20Newsgroups`.

[30] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging*, 1(2):113–122, 1982.

[31] M. Taddy. On estimation and selection for topic models. In N. D. Lawrence and M. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22, pages 1184–1193, La Palma, Canary Islands, 2012. PMLR.

[32] F. W. Townes, S. C. Hicks, M. J. Aryee, and R. A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20:295, 2019.

[33] Y. Vardi, L. A. Shepp, and L. Kaufman. A statistical model for positron emission tomography. *Journal of the American Statistical Association*, 80(389):8–20, 1985.

[34] H. M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 977–984, 2006.

[35] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017.