ALGORITHMS FOR FITTING TOPIC MODELS AND NON-NEGATIVE MATRIX FACTORIZATIONS TO COUNT DATA

PETER CARBONETTO*

1. Introduction. In this document, we study the problem of finding two non-negative matrices, F and L, minimizing the following loss function:

$$\ell(L, F) = -\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \log \lambda_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{ij}, \tag{1}$$

where $\lambda_{ij} = \sum_{k=1}^{K} l_{ik} f_{jk}$, and X, L and F are matrices of dimension $n \times m$, $n \times K$ and $m \times K$, respectively, with non-negative entries x_{ij}, l_{ik}, f_{jk}. (In the discussion and derivations below, K is assumed to be 2 or greater—the special case of $K = 1$, which has a relatively simple solution, is treated in Appendix **??**.) Since (1) is not convex when $K \geq 2$, we seek only to find a local minimum.

There are several ways to motivate minimizing (1). One motivation begins by modeling the counts x_{ij} using the Poisson distribution,

$$x_{ij} \sim \text{Poisson}(\lambda_{ij}), \tag{2}$$

where Poisson(λ) denotes the Poisson distribution with mean (and variance) λ. Next, by setting each of the Poisson rates to be a simple linear combination, $\lambda_{ij} = \sum_{k=1}^{K} l_{ik} f_{jk}$, the loss function (1) is recovered as the negative log-likelihood of this model (with terms that do not depend on either F or L omitted from the expression). From this point-of-view, solving (1) is equivalent to maximum-likelihood estimation (MLE) of F and L for the Poisson model (2), and, hence, it is sometimes called "Poisson non-negative matrix factorization."

A second point-of-view provides additional intuition behind (1). This loss function can be also viewed as measuring how well the low-rank factorization LF^T approximates X; that is, choices of L and F that achieve lower values of $\ell(L, F)$ reconstruct X more accurately. The precise quality measure that yields (1) is a Bregman divergence $D_\phi(X, LF^T)$ [1] with $\phi(x) = \sum_t x_t \log x_t$ as the choice of generating convex function [refs]. To draw an analogy to other matrix factorization methods such as principal components analysis (PCA) or factor analysis [2], we refer to L as the "loadings" matrix, and F the "factors" matrix. (In other papers, L and F are called the "activations" and "basis vectors," respectively).

Minizing (1) can also be motivated in a third way by connecting it to fitting topic models [refs]—we elaborate on this connection in the next section ("Connection to topic modeling").

As you can probably gather by these three motivations, the problem of minizing (1) is already very well-studied (e.g., [refs]). However, we have a particular focus on developing efficient algorithms for count data sets that are large and sparse—by *sparse*, we mean that most of the counts x_{ij} are zero. It turns out that existing algorithms do not seem to accommodate large, sparse data sets very well. Here we will look closely at the computational properties of this optimization problem, and

*Department of Human Genetics and the Research Computing Center, University of Chicago, Chicago, IL

use these investigations to help us design efficient algorithms for solving (1) in the setting of large, sparse data sets.

To illustrate why this matters, consider computing the loss function (1). This is easy to implement in R:

```
> # Code goes here.
```

Also explain the difficulty of this optimization problem, and why this motivates us to revisit algorithms for efficiently solving this problem. Despite the fact that this problem is well-studied, it turns out, as we will see, that existing solutions are inadequate for large, sparse data sets.

2. Connection to topic modeling. By optimizing (1) subject to the non-negativity constraints, we are also optimizing a topic model [refs]. Here we make this connection clear.

Give R code here implementing this transformation.

3. An EM algorithm. *Text goes here.*

4. Evaluating convergence. *Text goes here.*

5. EM, revisited. *Text goes here.*

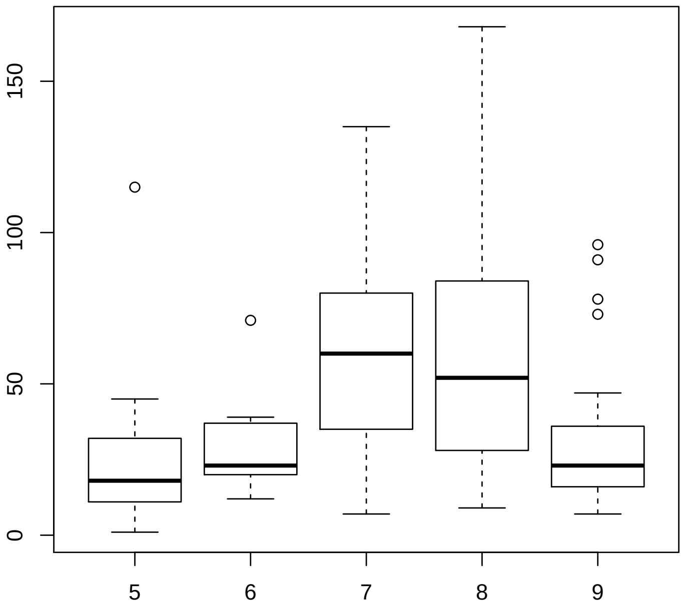
In this example we embed parts of the examples from the `kruskal.test` help page into a LaTeX document:

```
> data(airquality, package = "datasets")
> library("stats")
> kruskal.test(Ozone ~ Month,data = airquality)
        Kruskal-Wallis rank sum test

data:  Ozone by Month
Kruskal-Wallis chi-squared = 29, df = 4, p-value = 7e-06
```

which shows that the location parameter of the Ozone distribution varies significantly from month to month. Finally, we include a boxplot of the data, using

```
> boxplot(Ozone ~ Month, data = airquality)
```



Here is a citation: [3].

Acknowledgments. Thanks to Youngseok Kim, Abhishek Sarkar, Jason Willwersheid, Zihao Wang, Kushal Dey, Kevin Luo, Joyce Hsiao, Anthony Hung, Matthew Stephens, and many other members of the Stephens lab, former and past, for their valuable input. Also acknowledge the Research Computing Center.

Appendix A. The rank-one case. In the special case of $k = 1$, the Poisson rates can be expressed as a simple outer product of two vectors,

$$\Lambda = lf^T,$$

where $l = (l_1, \ldots, l_n)^T$ and $f = (f_1, \ldots, f_m)^T$, and the loss function (1) simplifies to

$$\ell(L, F) = -\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij} \log(l_i f_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} l_i f_j. \tag{3}$$

Any choice of l, f that minimizes (3) must therefore be solution to the following system of equations:

$$\begin{aligned} f_j &= \frac{\sum_{i=1}^n x_{ij}}{\sum_{i=1}^n l_i}, \quad j = 1, \ldots, m, \\ l_i &= \frac{\sum_{j=1}^m x_{ij}}{\sum_{j=1}^m f_j}, \quad i = 1, \ldots, n. \end{aligned}$$

Since the solution is only defined up to a scaling factor, we enforce the constraint that the mean of f is the same as the mean of l, yielding the following very simple solution:

$$
\begin{aligned}
f_j &= \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = 1, \ldots, m, \\
l_i &= \frac{1}{m}\sum_{j=1}^{m} x_{ij}, \quad i = 1, \ldots, n.
\end{aligned}
$$

In other words, the MLE for the rank-one matrix factorization is recovered as the row and column means of X.

REFERENCES

[1] L. M. Bregman, *The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming*, USSR Computational Mathematics and Mathematical Physics, 7 (1967), pp. 200–217.

[2] B. E. Engelhardt and M. Stephens, *Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis*, PLoS Genetics, 6 (2010), p. e1001117.

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