# Non-Parametric Discrete Mixture Model Recovery via Nonnegative Matrix Factorization

Stefan Karpinski, John R. Gilbert, Elizabeth M. Belding
Department of Computer Science
University of California, Santa Barbara
{*sgk,gilbert,ebelding*}*@cs.ucsb.edu*

Mixture modeling expresses probability densities as convex combinations of constituent probability distributions:

$$q_i(x) = \sum_{j=1}^{r} w_{ij} p_j(x), \tag{1}$$

Here $q_i$ and $p_j$ are density functions, and $w_{ij}$ are nonnegative weights, summing to unity for each $i$. In classical mixture modeling, the constituent density functions, $p_j$, are assumed to be from some class of parametric distributions. Various well established algorithms, typically using expectation minimization, can optimally recover the weights, $w_{ij}$, given an observed sample of values from the $q_i$ distributions [1].

In certain settings, however, mixture modeling is desirable, but the constituent distributions are neither known in advance, nor can they be assumed to be parametric. In this work, we demonstrate how, for discrete event spaces, nonnegative matrix factorization (NMF) can be effectively used to simultaneously recover both weights and constituent distributions, given a large collection of variably-sized samples from mixtures.

For discrete event spaces, Equation 1 is expressed succinctly as matrix multiplication. Letting $Q_{ik} = q_i(k)$, $W_{ij} = w_{ij}$, and $P_{jk} = p_j(k)$ we have:

$$Q = WP. \tag{2}$$

The problem of inferring both the weights, $w_{ij}$, and constituent distributions, $p_j$, from a collection of mixtures, $q_i$, is equivalent to finding the factors $W$ and $P$ given $Q$. All three matrices are constrined to be row-stochastic, meaning that all entries are nonnegative, with rows summing to unity.

The problem of finding such a factorization is known as nonnegative matrix factorization. Such factorizations are not unique, so perfect recovery of $W$ and $P$ cannot generally be achieved. On the other hand, any exact factorization of $Q$, is an equally valid mixture model for the given data. Since a variety of NMF algorithms have been proposed, this problem is partially solved. Several difficulties remain, however:

1) NMF is known to be NP-hard; thus, all efficient algorithms are heuristic, and may not yield adequate results;
2) $Q$ is not known exactly, only a finite sample for each distribution row of $Q$ is observed;
3) The samples for the rows may not have uniform size.

This list is not exhaustive, and we will address and discuss other challenges as well.

Our motivating application is mixture modeling for traces of network flows, whose distributions of packet sizes and inter-packet intervals seem to be effectively modeled as discretized mixture models, using NMF [2]. In this setting, there are several particularly challenging aspects:

1) The distribution of sample sizes is heavy-tailed, having a few very large samples, and many very small samples;
2) The constituent distributions are not uniformly represented: the most prevalent distribution has much larger average weights than the next, and so on.

We will demonstrate using simulated data why both of these properties make factor recovery particularly difficult.

To evaluate the effectiveness of NMF techniques for discrete mixture model recovery, we use synthetic data, since otherwise the true factors are unknown. To generate synthetic data, we use saw-tooth patterns as constituent distributions, densities of which are shown in Figure 1. These distributions are visually distinctive and not well-approximated by standard parametric distributions. The Pareto distribution is the classic heavy-tailed distribution, and describes the distribution of flow sizes in network trace studies [3], [4]. Accordingly, we choose sample sizes for each synthetic mixture from a Pareto distribution. Our synthetic weight matrices are also generated such that the prevalences of the component distributions — i.e. the column sums of $W$ — follow a power law. Figure 2 is a matrix plot of sample rows of randomly generated weights.

In short, we find that none of the existing NMF algorithms can accurately recover the constituent distributions used to generate synthetic mixtures. Recovered $P_*$ distributions for Lee and Seung's algorithms [5], and Kim and Park's alternating nonnegative least squares (ANLS) algorithm [6] are shown in Figure 3, dramatically illustrating their failure to

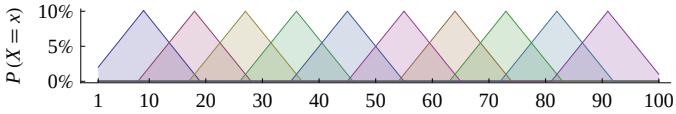


**Figure 1:** Discrete distributions used to generate synthetic mixtures.

**Figure 2:** Transposed matrix plot of 100 sample weight vectors.

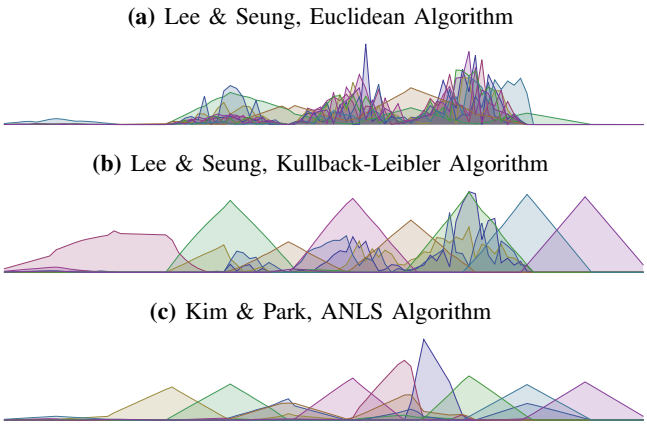


**Figure 3:** Recovered $P_*$ distributions for standard NMF algorithms, with random initialization and perfect knowledge of $Q$.

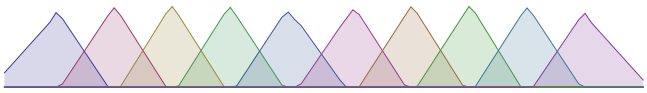

**Figure 4:** $P_0$ computed using SVD/$k$-means initialization.

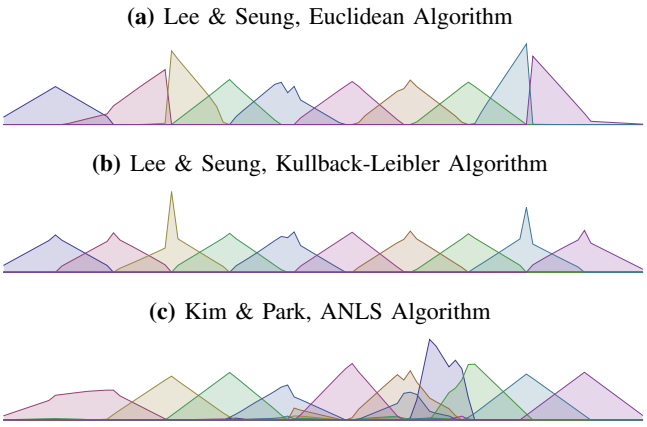


**Figure 5:** Recovered $P_*$ distributions for standard NMF algorithms, with SVD/$k$-means initialization and perfect knowledge of $Q$.

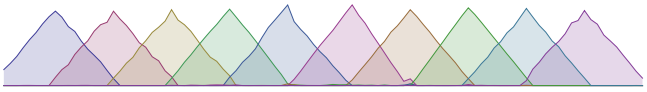

**Figure 6:** Accurate recovery of $P_*$ from sampled data.

accurately reconstruct the original distributions. It is well known that these algorithms may not converge to a globally optimum factorization. The quality of the end result is largely dependent on the matrices used to initialize the algorithms, which typically are random nonnegative matrices. To find a better factorization, we use a variation of the most promising initialization technique proposed by Langville et al. [7]:

1) Let $Q = USV'$, the singular value decomposition (SVD),
2) Use $k$-means to find $r$ clusters of columns in $V$,
3) Let $W_0$ be corresponding column centroids in $Q$,
4) Let $P_0$ be nonnegative minimizing $\|Q - W_0 P_0\|_F$.

Figure 4 shows that even without any further refinement, this initialization technique already recovers the overall shape of $P$ remarkably well. Although this recovery appears visually close to optimal, there are differences in shape which limit the quality of the approximation $W_0 P_0$. Intuitively, this initialization sits on a ridge above a deep valley, where the perfect recovery lies. If care is not taken, however, the NMF algorithms may descend the wrong side of the ridge, moving away from the optimal recovery rather than toward it. This is precisely what happens when any of these algorithms are applied naïvely but initialized with $W_0$ and $P_0$, shown in Figure 5.

Through a combination of intuition and trial and error, we have found that applying the ANLS algorithm for a few dozens of iterations, followed by the Kullback-Leibler algorithm, followed finally by the Euclidean algorithm results in perfect recovery, with error limited only by how many iterations of the Euclidean algorithm one is willing to wait for.

The above results all assume perfect knowledge of $Q$. In real situations, where discrete mixture models must be recovered from experimental data, there is highly imperfect knowledge of $Q$. Only a finite sample of values from each $q_i$ distribution are observed. These samples may be of different sizes, and in many situations, the vast majority of the samples are very small, many even consisting of only a single value. Can the basis matrix, $P$, still be recovered under such circumstances? Our meta-algorithm can accurately recover $P$ from sampled data when applied to appropriate estimates of $Q$.

The simplest estimate of $Q$ is based on the sample histogram matrix, $H$: $H_{ik}$ is the number times the value $k$ was sampled from the distribution $q_i$. We can approximate $Q$ by the row-stochasticized version of $H$, where each row is divided by its sum; call this matrix $C$. If nothing else was known about $Q$, we could not do much better than this approximation. We know, however, that $Q$ has rank $r$: we find the best rank $r$ approximation of $H$ using SVD. Why apply SVD to $H$ rather than $C$? By using SVD on $H$, we are giving each row weight proportional to its sample size in computing our approximation, thus giving more impact to more precisely known rows. The SVD approximation matrix may have negative values; since we know $Q$ has none, we truncate these to zero.[1] Let $A$ denote this nonnegative, nearly rank $r$ approximation of $H$, and let be $R$ its row-stochastization.

To successfully apply our meta-algorithm to sampled data, the following steps may be applied:

1) Compute $W_0, P_0$ from SVD/$k$-means on $R$.
2) Iterate ANLS on $R$ starting with $W_0, P_0$.
3) Let $W_*$ be nonnegative minimizing $\|H - W_* P_*\|_F$.
4) Iterate Kullback-Leibler on $H$ starting with $W_*, P_*$.
5) Iterate Euclidean on $H$ starting with $W_*, P_*$.

This meta-algorithm can recover $P$ accurately, even when $Q$ is sampled with infinite variance (heavy-tailed) and prevalences of constituent distributions follow a power law. An accurate recovery from sampled data is shown in Figure 6.

[1] Projection into the nonnegative orthant may cause the approximation to become no longer rank $r$. It will, however, remain nearly rank $r$, in the sense that it will have $r$ large singular values and the rest much smaller.

## REFERENCES


[1] G. McLachlan and D. Peel, *Finite Mixture Models*. New York NY, USA: John Wiley and Sons, 2000.

[2] S. Karpinski, E. Belding, K. Almeroth, and J. Gilbert, "Linear representation of network traffic with special application to wireless workload generation," *Mobile Networks and Applications*, November 2008.

[3] F. Hernández-Campos, M. Karaliopoulos, M. Papadopouli, and H. Shen, "Spatio-temporal modeling of traffic workload in a campus WLAN," in *2nd Annual International Workshop on Wireless Internet*, Boston MA, USA, August 2006.

[4] M. Karaliopoulos, M. Papadopouli, E. Raftopoulos, and H. Shen, "On scalable measurement-driven modeling of traffic demand in large WLANs," in *IEEE Workshop on Local and Metropolitan Area Networks*, Princeton NJ, USA, June 2007.

[5] D. Lee and H. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing*, vol. 13, 2001.

[6] H. Kim and H. Park, "Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method," *SIAM Journal in Matrix Analysis and Applications*, vol. 30, no. 2, pp. 713–730, May 2008.

[7] A. Langville, C. Meyer, R. Albright, J. Cox, and D. Duling, "Algorithms, initializations, and convergence for the nonnegative matrix factorization," in *SIGKDD International Conference on Knowledge Discovery and Data Mining*, Philedelphia PA, USA, August 2006.