Analysis of L2-loss for Probabilistically valid Factorizations under General Additive Noise

Tamir Hazan and Amnon Shashua
School of Computer Science and Engineering,
The Hebrew University,
Jerusalem 91904, Israel
e-mail: {tamir,shashua}@cs.huji.ac.il

Abstract

We address the problem of pLSA with (i) multivariate (more than two), (ii) under L_2 loss. We show that a statistical valid solution can be derived with superior qualities compared to a Maximum-Likelihood solution (relative entropy loss). We derive bounds showing that an L_2 loss over co-occurrence data (non-linear) under simplex (probability) constraints behaves well under general additive noise and derive bounds showing that an ML solution would in some (typical for applications) cases behave badly. We then derive an update rule for the variables of the decomposition which guarantees a statistically valid locally optimal solution. The update rule is based on a simple building block referred to as ”projection onto the probability simplex” whose solution can be found in a finite number of steps.

1 Introduction

We will be looking at the following class conditional (or aspect model) problem:

$$\min_{P \in \mathcal{Q}} loss(\hat{P}, P) \tag{1}$$

where $loss(\cdot,\cdot)$ is an error measure (to be defined), $\hat{P}$ is the empirical distribution represented by the co-occurrences of the data and $\mathcal{Q}$ is the model distribution family defined below:

$$\mathcal{Q} = \{\sum_{j=1}^{k} \lambda_j P^j \ : \ \boldsymbol{\lambda} \geq 0, \boldsymbol{\lambda}^\top \mathbf{1} = 1, P^j \in \mathcal{H}\} \tag{2}$$

where $\mathcal{H}$ is the set of d-way rank-1 tensors with convex constraints:

$$\mathcal{H} = \{\otimes_{i=1}^{d} \mathbf{u}_i \ : \ \mathbf{u}_i \geq 0, \ \mathbf{u}_i^\top \mathbf{1} = 1, i = 1, ..., d\}$$

where we use the notation $\otimes_{i=1}^{d}\mathbf{u}_i = \mathbf{u}_1 \otimes \cdots \otimes \mathbf{u}_d$ representing the outer-product of d vectors. The array $U = \otimes_{i=1}^{d}\mathbf{u}_i$ is indexed by $i_1, ..., i_d$ where $U_{i_1,...,i_d} = u_{1_{i_1}} \cdots u_{d_{i_d}}$.

When $d = 2$, and $loss(\cdot,\cdot)$ is the relative entropy measure $D(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log(p_i/q_i)$, the formulation above reduces to the well known pLSA (Hofmann, 1999) interpreted as follows: Let X, Y be two observable random variables generating a co-occurrence matrix $G_{ij} = \hat{P}(X = x_i, Y = y_j)$ and let Z be a hidden variable inducing conditional independence between X, Y, i.e., $X \perp Y \mid Z$. Then, the distribution model $\mathcal{Q}$ is represented by the mixture:

$$\sum_{j=1}^{k} P(Z = z_j) P(X \mid Z = z_j) P(Y \mid Z = z_j)$$

[1] Hebrew University, School of Computer Science and Engineering, Technical Report TR-13/2007, April 2007.

If we set λ_j to represent $P(Z = z_j)$ forming a diagonal matrix Σ, column vectors $\mathbf{u}_j$ forming a matrix U to represent $P(X \mid Z = z_j)$, row vectors $\mathbf{v}_j^\top$ forming a matrix V to represent $P(Y \mid Z = z_j)$, then the model distribution $\mathcal{Q}$ is the set $U\Sigma V^\top$ under convex constraints $\Sigma, U, V \geq 0$ and $U^\top \mathbf{1} = V\mathbf{1} = \mathbf{1}^\top \Sigma \mathbf{1} = 1$. The optimization problem becomes:

$$\min_{\Sigma,U,V\geq 0} loss(G, U\Sigma V^\top) \ \ s.t.\ U^\top \mathbf{1} = V\mathbf{1} = \mathbf{1}^\top \Sigma \mathbf{1} = 1. \tag{3}$$

An equivalent representation, closely related to Non-Negative Matrix (NMF) factorization (Paatero & Tapper, 1994; Lee & Seung, 1999), would be to consider a column normalized co-occurrence matrix $\bar{G}_{ij} = \hat{P}(X = x_i \mid Y = y_j)$, i.e., $\bar{G}^\top \mathbf{1} = 1$, and noting that $P(X \mid Y) = \sum_j P(X \mid Z = z_j)P(Z = z_j \mid Y)$ we have the model distribution $\mathcal{Q}$ is represented by the product UV where U is defined as above and columns of V represent $P(Z \mid Y = y_i)$. The conditions on U, V are therefore $U, V \geq 0$ and $U^\top \mathbf{1} = V^\top \mathbf{1} = 1$, and the optimization problem becomes:

$$\min_{U,V\geq 0} loss(\bar{G}, UV) \ \ s.t.\ U^\top \mathbf{1} = V^\top \mathbf{1} = 1. \tag{4}$$

The pLSA formulation was introduced in the context of the relative-entropy error measure which guarantees a (locally optimal) Maximum Likelihood (ML) solution. Generally, the problem setup of eqn. 1 under the relative entropy loss with the mixture model $\mathcal{Q}$ defined in eqn. 2 is an ML problem which falls under the general Expectation-Maximization (EM) scheme. The EM algorithm introduces auxiliary variables and breaks down the problem into pieces which could be managed without the need to explicitly enforce the non-negative constraints $\Sigma, U, V \geq 0$. As a result the EM approach not only seeks an ML solution but also is easy to manage in the context of non-negative constraints.

The NMF formulation (and its $d > 2$ extension known as NTF (Welling & Weber, 2001; Shashua & Hazan, 2005; Hazan et al., 2005)) has been introduced under both L_2 and relative-entropy loss measures but mostly without the affine constraints, i.e., factor G into UV where $U, V \geq 0$. Under the relative entropy loss, however, the affine constraints $U^\top \mathbf{1} = V^\top \mathbf{1} = 1$ in eqn. 4, i.e., with column normalized co-occurrences, can be satisfied following a simple modification to the Lee-Sueng (Lee & Seung, 1999) update rule (Gaussier & Goutte, 2005; Ding et al., 2006).

A solution for NMF with affine constraints, i.e., eqn. 4, under an L_2 loss function is more problematic and so far has not been dealt with in a satisfactorily manner. There are some discussions about "mixed" conic and affine constraints (Lee & Seung, 1997) such as: (i) $V^\top \mathbf{1} = 1, 0 \leq U \leq 1$ and $V \geq 0$, or (ii) $0 \leq U \leq 1, V \geq 0$, or (iii) $V^\top \mathbf{1} = 1, 0 \leq U \leq 1$, but those do not satisfy the probabilistic constraints. Therefore, the L_2 solution for NMF with the full set of affine constraints is still an open problem.

We will refer, from now on, to the general aspect model described in eqn. 1 under the L_2 loss with the mixture model $\mathcal{Q}$ defined in eqn. 2 as "probabilistic non-negative tensor factorization" abbreviated by pNTF. The special case of $d = 2$ would be abbreviated by pNMF. We focus on the L_2 loss because, as stated above, the relative entropy loss reduces the problem to the well known ML using the popular and well understood EM algorithm. Our contribution in this paper is two fold.

[1]: We argue that applications of interest where an aspect model is of use are typically subject to additive noise. The behavior of an L_2 loss in the context of a linear model is well understood — it provides the ML solution under Gaussian noise — however, when the data consists of co-occurrence matrices (or tensors) the rational for an L_2 loss is unclear. We will develop bounds (for general additive noise) both for L_2 and relative-entropy loss showing that an L_2 is greatly superior to a ML solution — thus paving the way to our second contribution.

[2]: Our second contribution is algorithmic. As stated above, the pNTF (or pNMF) lacks a satisfactory algorithm. The Lee-Seung update rule (for L_2 loss) would not account for the affine constraints and the NTF extension are also based on the Lee-Seung update rule and thus would not satisfy the constraints. We will show that the pNTF problem can be broken down into two simple components. The first is an iterative update rule based on the "projection onto probability simplex" sub-problem and the second component is a small size QLP which can be handled by any standard solver.

As for applications, the original application domain for the aspect model is typically found in document analysis using the bag-of-words concept where the goal is to map a document into a set of topics (Hofmann, 1999). The co-occurrence matrix holds the joint frequency of words and documents and the set of topics are represented by the rank-1 matrices $\lambda_j \mathbf{u}_j \mathbf{v}_j^\top$ of the decomposition. In computer vision there has been a recent trend of adapting the bag-of-words concept to co-occurrence of features (Fei-Fei & Perona, 2005; Quelhas et al., 2005; Sivic et al., 2005) for identifying key features and performing object classification. Non-negative tensor decompositions are also used for sparse image coding (Hazan et al., 2005) where the class of training images are stacked to form a 3D cube (tensor) whose factors correspond to basic (sparse) features.

For the sake of minimizing technical jargon we will introduce our derivations for the special case of $d = 2$, i.e., pNMF, as most readers are familiar with matrix manipulations. We will devote Section 4 to fill-in the necessary material going from pNMF to pNTF.

2 In Favor of L_2 Loss for Co-occurrence Data

To recap, using an L_2 loss for factor analysis represented in eqn. 3, or more generally in eqn. 1, would not provide us with the ML solution — which for lack of other arguments seems like the right thing to do. In other words, minimizing the Frobenius norm $\|G - \sum_j \lambda_j \mathbf{u}_j \mathbf{v}_j^\top\|_F^2$ under the constraints of Eqn. 3 would provide a probabilistically admissible solution, but would not be the ML solution — which seems to go against conventional wisdom.

The behavior of an L_2 loss is well understood when approximating a linear model, i.e., one obtains the ML solution under Gaussian additive noise. The question is what can we expect from an L_2 under a simplex model (the convex constraints) operating on co-occurrence (non-linear by nature) data and under *general* additive noise? We will provide bounds both for L_2 and relative entropy in order to measure the effect of additive noise between L_2 and ML solutions.

Let E be a bounded *perturbation* matrix with $\|E\|_\infty = \epsilon < \infty$, i.e., the maximum absolute value of an entry of E is bounded from above by some ϵ. Let $\mathcal{Q}$ be defined as in eqn. 2, i.e., the space of low rank (rank = k) probability matrices. Consider some $P_0 \in \mathcal{Q}$ additively perturbed by E, and let $P_\epsilon^* \in \mathcal{Q}$ the closest rank-k probability matrix to $P_0 + E$. The approximation is "well behaved" if the difference between P_0 and P_ϵ^* is small. To use a "neutral" measure of difference we will measure the difference between P_0 and P_ϵ^* by the infinity norm (the value of the largest entry). A low value for $\|P_0 - P_\epsilon^*\|_\infty$ is definitely desired whereas a large value indicates that something in the approximation is wrong. We introduce the following result:

Proposition 1 *Let $P_0 \in \mathcal{Q}$ be a probability matrix and let $P_\epsilon^* \in \mathcal{Q}$ be the closest admissible matrix to $P_0 + E$ in Frobenius norm. Then,*

$$\|P_0 - P_\epsilon^*\|_\infty \leq 2\sqrt{\epsilon}$$

Proof: see Appendix. □

We see that the approximation error of the Frobenius norm is governed by the *maximal* entry of the perturbation matrix E. Therefore, when $\|E\|_\infty$ is small so is the approximation error — that is, the approximation behaves well.

We will turn our attention to ML approximation, i.e., the approximation using the relative entropy measure in eqn. 3. We will focus on a reduced case where $\mathcal{Q}$ is the set of rank-1 $n \times n$ probability matrices. Define:

$$\alpha_1 = \frac{\|E\mathbf{1}\|_\infty}{(1 + \mathbf{1}^\top E\mathbf{1})} \quad \text{and} \quad \alpha_2 = \frac{\|E^\top\mathbf{1}\|_\infty}{(1 + \mathbf{1}^\top E\mathbf{1})}.$$

We state the following result:

Proposition 2 *Let $P_0 = \mathbf{p}_0\mathbf{q}_0^\top$ be a rank-1 probability matrix and let $P_\epsilon^* = \mathbf{p}_\epsilon{}^*\mathbf{q}_\epsilon^{*\top}$ be the closest rank-1 matrix to $P_0 + E$ in relative entropy. Then,*

$$\left| \|\mathbf{p}_0 - \mathbf{p}_\epsilon^*\|_\infty - \|\mathbf{p}_0\|_\infty \frac{|\mathbf{1}^\top E\mathbf{1}|}{1 + \mathbf{1}^\top E\mathbf{1}} \right| \leq \alpha_1$$

and

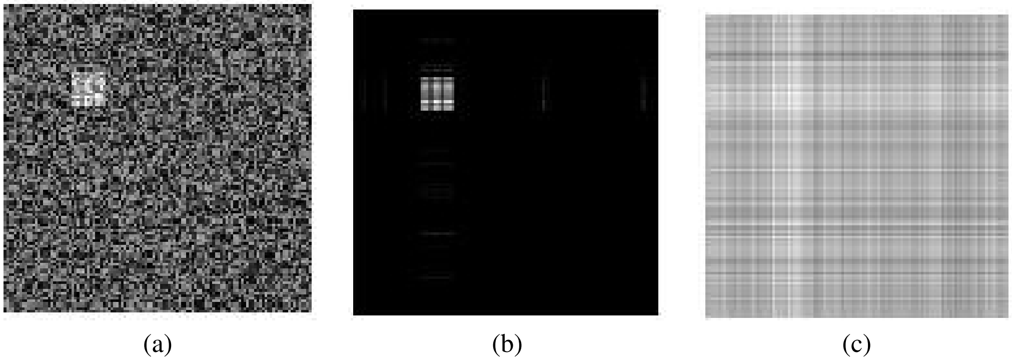
$$\left| \|\mathbf{q}_0 - \mathbf{q}_\epsilon^*\|_\infty - \|\mathbf{q}_0\|_\infty \frac{|\mathbf{1}^\top E\mathbf{1}|}{1 + \mathbf{1}^\top E\mathbf{1}} \right| \leq \alpha_2$$

and

$$\|P_0 - P_\epsilon^*\|_\infty \leq 3 \max\left\{\|\mathbf{p}_0 - \mathbf{p}_\epsilon^*\|_\infty \ , \ \|\mathbf{q}_0 - \mathbf{q}_\epsilon^*\|_\infty\right\}.$$

Proof: See Appendix.□

The bounds above would be tight when α_1, α_2 would be small, i.e., when $\|E\mathbf{1}\|_\infty \ll \mathbf{1}^\top E\mathbf{1}$. The following result bounds the value of α_1, α_2 in the case where the perturbation E satisfies $\|E\|_\infty = \epsilon$ and $P_0 + E \geq 0$ and the entries of E are sampled independently:


(a)

(b)

(c)

Figure 1: Illustration of L_2 and relative-entropy rank-1 approximation: (a) $P_0 + E$, (b) closest rank-1 (with probability constraints) in L_2, (c) ML solution.

Proposition 3 *Let β be the number of indices satisfying $P_{0ij} < \epsilon$, then,*

$$\alpha_i \leq \frac{\epsilon \cdot \min\{\beta, n\}}{\epsilon\beta + 1} \quad \textit{with high probability when} \;\; n \to \infty$$

Proof: see Appendix. □

Given the bound on α_1, α_2, then if $\beta = \omega(n)$ is super-linear, i.e., the number of entries in P_0 is larger than n, then α_i goes to zero:

Corollary 1 *If β is super-linear, abbreviated $\beta = \omega(n)$, then*

$$\|\mathbf{p}_0 - \mathbf{p}^*_\epsilon\|_\infty \overset{n\to\infty}{\longrightarrow} \|\mathbf{p}_0\|_\infty \;\; \textit{and} \;\; \|\mathbf{q}_0 - \mathbf{q}^*_\epsilon\|_\infty \overset{n\to\infty}{\longrightarrow} \|\mathbf{q}_0\|_\infty,$$

with probability one.

The results above show that when the data co-occurrence matrix has a low entropy — reflected in the requirement that (asymptotically) more than a square-root of the number of entries have a value which is smaller than $\|E\|_\infty$ (the largest entry of the perturbation matrix) then with increasing array size the ML approximation would produce an ill-behaved approximation in the sense that $\|P_0 - P^*_\epsilon\|_\infty$ is governed by $\|P_0\|_\infty$. In contrast, the L_2 loss generates well-behaved approximations which commensurate with the magnitude of the perturbation. Fig. 1 illustrates this point with a rank-1 array consisting of a block of high values with remaining entries consisting of random low value perturbation (representing $P_0 + E$). Fig. 1(b) shows the L_2 loss reconstruction (closest rank-1 array) and Fig. 1(c) shows the relative-entropy (ML via EM) reconstruction. One can clearly see that the ML solution attenuates (significantly) the gap between the high and low entries whereas the L_2 loss more or less preserves that gap.

3 pNMF via Successive Von-Neumann Projections

We turn our attention to deriving an update rule for the pNMF problem described in eqn. 3, i.e., $\min \|G - \sum_j \lambda_j \mathbf{u}_i \mathbf{v}_j^\top\|_F^2$ subject to the simplex constraints on $\boldsymbol{\lambda}, \mathbf{u}_j, \mathbf{v}_j$ and $G \geq 0$ is a $d_1 \times d_2$ matrix with unit sum.

In the context of an alternating optimization where each group of variables $\boldsymbol{\lambda}$, $\{\mathbf{u}_j\}_1^k$ and $\{\mathbf{v}_j\}_1^k$ are updated given the value of the remaining variables, each update process involves the solution to a Quadratic Linear Program (QLP). However, the QLPs associated with the update of the matrices U and V can be reduced to a simple problem, referred to as the *projection onto the probability simplex* (PPS):

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{b}\|^2 \;\; s.t. \;\; \mathbf{x} \geq 0, \;\; \|\mathbf{x}\|_1 = 1, \tag{5}$$

which has a particularly simple solution. We will begin by showing how the update rule of the vectors $\mathbf{u}_j$ and $\mathbf{v}_j$ reduce to a PPS problem.

Let $G^t = G - \sum_{j \neq t} \lambda_j \mathbf{u}_j \mathbf{v}_j^\top$ and we wish to find a probability vector $\mathbf{u}_t \geq 0$, $\|\mathbf{u}_t\|_1 = 1$ which minimizes $\|G^t - \lambda_t \mathbf{u}_t \mathbf{v}_t^\top\|_F^2$ — this can be reduced to building block (eqn. 5) as follows:

$$\begin{aligned}
&\text{argmin}_{\mathbf{u}_t} \ \|G^t - \lambda_t \mathbf{u}_t \mathbf{v}_t^\top\|_F^2 = \\
&\text{argmin}_{\mathbf{u}_t} \ \lambda_t^2 \|\mathbf{u}_t\|^2 \|\mathbf{v}_t\|^2 - 2\lambda_t \mathbf{u}_t^\top G^t \mathbf{v}_t = \\
&\text{argmin}_{\mathbf{u}_t} \ \|\frac{G^t \mathbf{v}_t}{\lambda_t \|\mathbf{v}_t\|^2} - \mathbf{u}_t\|^2
\end{aligned}$$

Thus we wish to find a probability vector $\mathbf{u}_t$ closest (in L_2 norm) to the point $\mathbf{b} = G^t \mathbf{v}_t / \lambda_t \|\mathbf{v}_t\|^2$. Likewise, the same argument holds if we wish to solve for $\mathbf{v}_t$ given the remaining vectors — we end up solving a PPS problem. We will defer the details of the solution of a PPS problem to the next section and focus next on the update rule for $\boldsymbol{\lambda}$.

The second type of QLP problem arises when we desire to solve for $\boldsymbol{\lambda}$ given the values of $\{\mathbf{u}_j\}_1^k$ and $\{\mathbf{v}_j\}_1^k$. Let $\mathbf{b} = vec(G)$ the vector representation of the matrix G (by some fixed order, say lexicographic order of the indices) and let A be the matrix whose j'th column is the probability vector $vec(\mathbf{u}_t \mathbf{v}_t^\top)$. Given $\mathbf{b}$ and A we desire to find a vector $\boldsymbol{\lambda}$ which minimizes the following constrained optimization:

$$\min_{\boldsymbol{\lambda}} \|A\boldsymbol{\lambda} - \mathbf{b}\|^2 \ \ s.t \ \ \boldsymbol{\lambda} \geq 0, \ \|\boldsymbol{\lambda}\|_1 = 1. \tag{6}$$

This quadratic program defines a projection of the vector $\mathbf{b}$ onto the convex hull spanned by the columns of A — a problem which we refer to as the "*projection onto a general simplex*". Since the number of variables k is relatively small (in applications it is typically in the order of a dozen, e.g, the number of topics contained in a document) we leave this problem for a standard QLP solver (such as the one supplied by Matlab).

3.1 Finding the Projection onto the Probability Simplex

We wish to solve the QLP described in eqn. 5. We first show how the projection can be derived by an iterative scheme and then introduce a modification which guarantees a finite number of steps (at most d).

We define below two subproblems, each with a closed-form solution:

$$P_1(\mathbf{y}) = \text{argmin}_{\mathbf{x}} \|\mathbf{x} - \mathbf{y}\|^2 \ \ s.t. \ \ \mathbf{x}^\top \mathbf{1} = 1, \tag{7}$$

$$P_2(\mathbf{y}) = \text{argmin}_{\mathbf{x}} \|\mathbf{x} - \mathbf{y}\|^2 \ \ s.t. \ \ \mathbf{x} \geq 0. \tag{8}$$

We will use the Von-Neumann (Neumann, 1950) successive projection lemma stating that $P_1 P_2 P_1 P_2 ... P_1(\mathbf{b})$ will converge to the projection of $\mathbf{b}$ onto the intersection of the affine and conic subspaces described above[1]. Therefore, what remains to show is that the projections P_1 and P_2 can be solved efficiently (and in closed form).

We begin with the solution for P_1. Setting the derivative of the Lagrangian corresponding to eqn. 7 to zero yields: $\mathbf{x} - \mathbf{y} = \lambda \mathbf{1}$ where λ is the Lagrange multiplier. Multiplying both sides by the vector $\mathbf{1}$ provides $\lambda = (1/d)(1 - \mathbf{y}^\top \mathbf{1})$ which after substitution provides the solution: $P_1(\mathbf{y}) = \mathbf{y} + \frac{1}{d}(1 - \mathbf{y}^\top \mathbf{1})\mathbf{1}$.

The projection $P_2(\mathbf{y})$ can also be described in a simple closed form manner. Let I_+ be the set of indices corresponding to positive entries of $\mathbf{y}$ and I_- the set of non-positive entries of $\mathbf{y}$.The criterion function $\|\mathbf{x} - \mathbf{y}\|^2$ becomes: $\|\mathbf{x} - \mathbf{y}\|^2 = \sum_{i \in I_+} (x_i - y_i)^2 + \sum_{i \in I_-} (x_i - y_i)^2$.

Clearly, the minimum energy over $\mathbf{x} \geq 0$ is obtained when $x_i = y_i$ for all $i \in I_+$ and zero otherwise. Let $th_{\geq 0}(\mathbf{y})$ stand for the operator that zeroes out all negative entries of $\mathbf{y}$. Then, $P_2(\mathbf{y}) = th_{\geq 0}(\mathbf{y})$.

The two projection operators can be integrated into a finite step scheme by noting that when $P_1(\mathbf{y})$ has a non-positive coordinate (thus is set to zero by P_2) it will remain non-positive in future P_1 projections (proof omitted due to limited space) therefore without loss of generality a non-positive coordinate can be removed. Therefore, the scheme can have at most d steps. Above we summarize the projection onto the probability simplex algorithm.

Algorithm 1 (Projection onto Probability Simplex) *The following algorithm projects a general point* $\mathbf{b} \in R^d$ *onto the intersection of the non-negative orthant of* R^d *and the unit hyperplane, i.e., finds* $\mathbf{x}$ *which minimizes (global optimum)* $\|\mathbf{x} - \mathbf{b}\|^2$ *under the constraints* $\mathbf{x} \geq 0$ *and* $\mathbf{x}^\top \mathbf{1} = 1$.

1. *Set the estimated projection* $\mathbf{x} = \mathbf{b}$. *Set the active face* $I = \{1, ..., d\}$

[1] actually, the Von-Neumann lemma applies only to linear subspaces. Nevertheless, it can be shown to apply also to the PPS problem.

```
2. for $t = 1, ..., d$
    (a) $\mathbf{x}_I \leftarrow \mathbf{x}_I + \frac{1}{d}\left(1 - \sum_{i \in I} x_i\right)$
    (b) **if** $I = \{i \in I : x_i > 0\}$ **then** break.
    (c) $\forall i \in \{1, ..., d\} \setminus I$   set $x_i = 0$
    (d) $I \leftarrow \{i \in I : x_i > 0\}$
```

The algorithm makes use of very simple computations and with a linear upper-bound on the number of steps. In practice the number of steps are a small fraction of the theoretical upper bound. It is important to note that a general QLP solver would make the pNMF unwieldy as the dimension d of the problem is the number of rows or columns of the input matrix G. For example, in computer vision applications each column of G represents an image thereby making $d \approx 10^5$. Our algorithm, which is specialized to the probability simplex, has at most d very simple calculations and in practice much fewer than d.

4 pNTF: Beyond Matrix Factorizations

The derivations and algorithms presented so far scale up beyond matrices. The general aspect-model presented in eqn. 1 with distribution model $\mathcal{Q}$ of eqn. 2 takes the following form:

$$\min_{\boldsymbol{\lambda} \geq 0, \mathbf{u}_i^j \geq 0} \|G - \sum_{j=1}^{k} \lambda_j \otimes_{i=1}^{n} \mathbf{u}_i^j\|_F^2 \;\; s.t. \;\; \|\boldsymbol{\lambda}\|_1 = 1, \; \|\mathbf{u}_i^j\|_1 = 1, \tag{9}$$

where $G \in R^{d_1 \times \ldots \times d_n}$ is an n-dimensional (probability) array, and $\otimes_i \mathbf{u}_i^j$ stands for the rank-1 tensor $\mathbf{u}_1^j \otimes \ldots \otimes \mathbf{u}_n^j$ described by the outer-product of n vectors.

The update computation of a vector $\mathbf{u}_r^t$ given the values of the remaining vectors and of $\boldsymbol{\lambda}$ reduces to the PPS problem of eqn. 5 and whose solution is described in Section 3.1. To derive the reduction we need to introduce a contraction notation: Let $A \odot \otimes_{m \neq r} \mathbf{u}_m$ stand for the vector resulting from contracting the tensor A with the vectors $\mathbf{u}_i$ where $i \neq r$:

$$A \odot \otimes_{m \neq r} \mathbf{u}_m = \sum_{i_1,..,i_{r-1},i_{r+1},...,i_n} A_{i_1,...,i_n} \Pi_{m \neq r} u_{m,i_m}$$

the result is a vector since the index i_r remains free. So, for example in the 2D case $(n, r = 2)$ $A \odot \mathbf{u}_1 = \sum_{i_1} a_{i_1,i_2} u_{1,i_1} = A^\top \mathbf{u}_1$ and for $n = 3, r = 2$ we have $A \odot (\mathbf{u}_1 \otimes \mathbf{u}_3) = \sum_{i_1,i_3} a_{i_1,i_2,i_3} u_{1,i_1} u_{3,i_3}$. Using this notation, first let $G^t = G - \sum_{j \neq t} \lambda_j \otimes_i \mathbf{u}_i^j$ and we wish to find $\mathbf{u}_r^t \geq 0$ which minimizes $\|G^t - \lambda_t \otimes_i \mathbf{u}_i^t\|_F^2$, then it is possible to show that:

$$\begin{aligned} &\text{argmin}_{\mathbf{u}_r^t \geq 0, \|\mathbf{u}_r^t\|_1 = 1} \|G^t - \lambda_t \otimes_i \mathbf{u}_i^t\|_F^2 = \\ &\text{argmin}_{\mathbf{u}_r^t \geq 0, \|\mathbf{u}_r^t\|_1 = 1} \|\frac{G^t \odot \otimes_{i \neq r} \mathbf{u}_i^t}{\lambda_t \Pi_{i \neq r} \|\mathbf{u}_i^t\|^2} - \mathbf{u}_r^t\|^2, \end{aligned}$$

which is achieved by projecting the point $\mathbf{b} = (G^t \odot \otimes_{i \neq r} \mathbf{u}_i^t)/(\lambda_t \Pi_{i \neq r} \|\mathbf{u}_i^t\|^2)$ onto the probability simplex (Section 3.1). To update the weights $\lambda_1, ..., \lambda_k$ we follow the same path as described in pNMF: let $\mathbf{b} = vec(G)$ the vector representation of the tensor G (by some fixed order, say lexicographic order of the indices) and let A be the matrix whose j'th column is $vec(\otimes_i \mathbf{u}_i^j)$. Given $\mathbf{b}$ and A we desire to find a vector $\boldsymbol{\lambda}$ which minimizes eqn. 6.

5 Experiments

We have performed a wide range of tests to compare the pNMF/pNTF performance against conventional NMF/NTF (with L_2 loss) for the purpose of gauging the importance of the integral weights recovered during decompositions. We also performed tests against EM and "tempered" EM (Hofmann, 1999) which also provide a weighted decomposition but under the relative-entropy error measure (this also covers NMF with relative-entropy loss).

Our first series of tests is on face images taken from the MIT CBCL database containing 2429, 19 × 19, images. We wish to recover local-part features from the image class via a non-negative decomposition and then use those features as

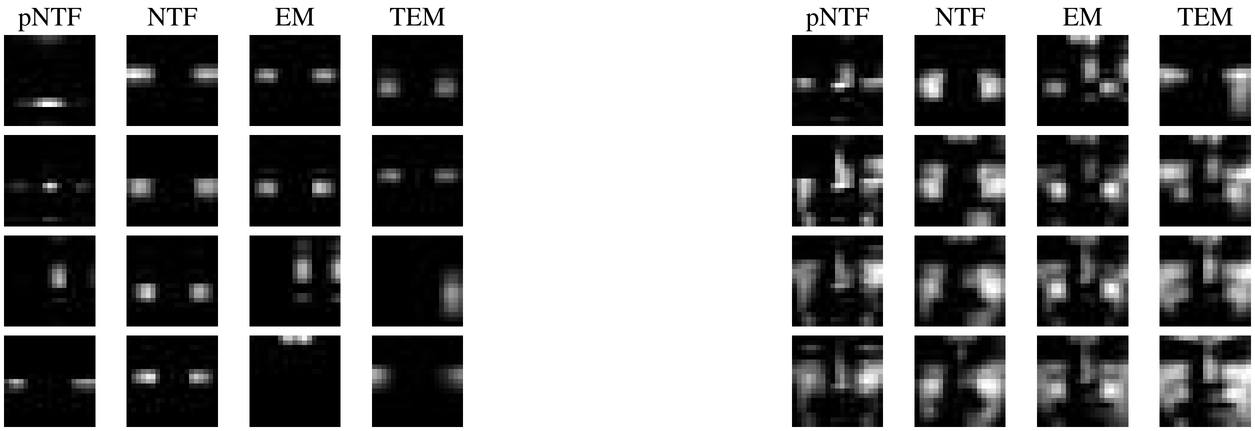
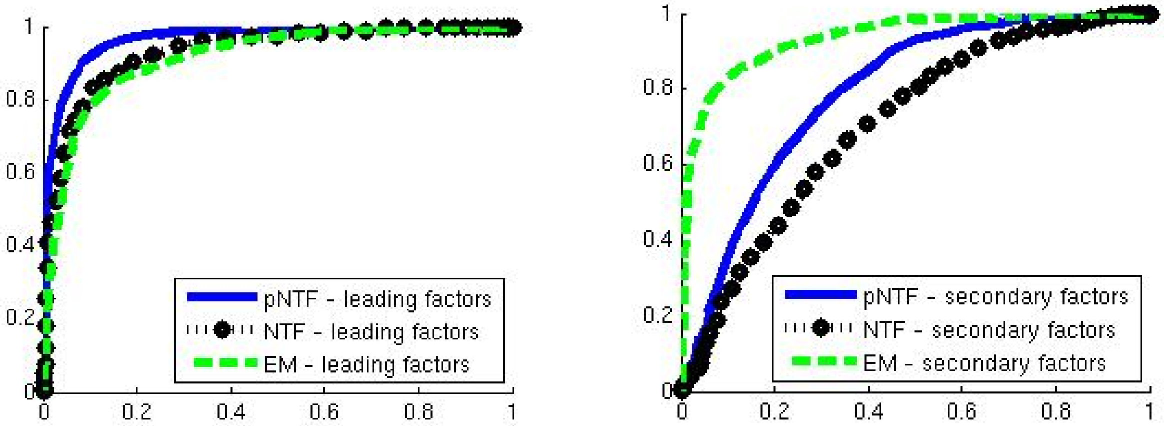


Figure 2: *Left: Comparing leading factors of pNTF versus NTF, EM and Tempered-EM factors on face images. Right: Superimposing the* $q = 4, 8, 12, 16$ *(from top to bottom) leading factors on face images.*



(a)

(b)

Figure 3: *ROC curves generated by SVM using the filter responses of (a) leading pNTF, NTF and EM factors 1–5, and (b) non-leading factors 6–10*

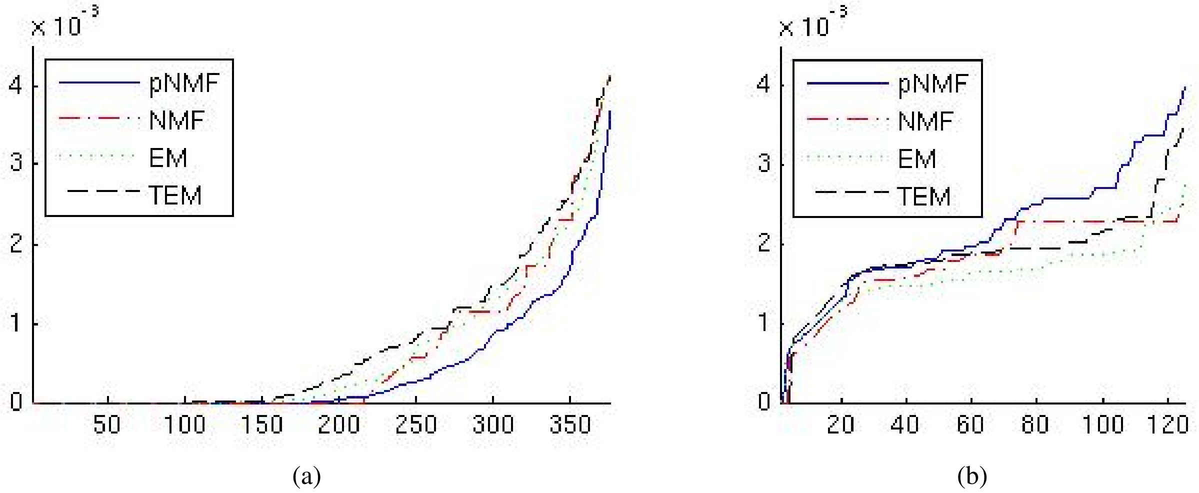


Figure 4: *The "science" factor coefficient in the reconstruction of (a) non "science" postings (graph values should be low), and (b) "science" postings (values should be high).*

measurements for a face detection classification and compare the ROC performance across pNTF, NTF and EM. It has been argued in the past (Shashua & Levin, 2001; Hazan et al., 2005) that a natural representation of a set of images is a 3D cube (rather than a matrix whose columns correspond to the vectorized images). Consequently, we construct a 3-way array from the set of face images and perform the decomposition methods directly on the array. The factors of the decomposition are rank-1 tensors each represented by an outer-product of three vectors. The slices of the rank-1 factor are scaled versions of a single rank-1 matrix. In other words, if $A_1, ..., A_m$ represent the set of m images as slices of the tensor G, then the tensor decomposition produces a set of k rank-1 matrices $\lambda_1 \tau_1, ..., \lambda_k \tau_k$ whose linear combinations reconstruct (to a least-squares approximation) the original images. We refer to τ_j as the factors of the decomposition.

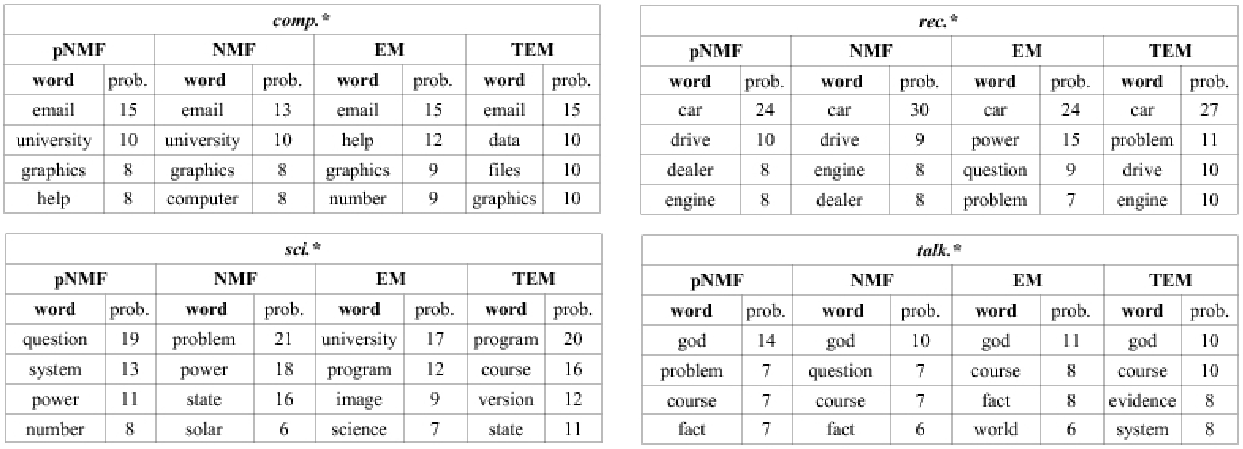
In Fig. 2 we show the four leading factors (out of a decomposition of $k = 50$ factors) of pNTF, NTF, EM, and tempered-EM. The local-part nature of the factors is evident in all the decomposition methods but one can see a sharper distinction between the factors in pNTF compared to NTF (where the factors seem to have a higher overlap with each other). This sharper distinction becomes more evident when we take a super-position of the $q = 4, 8, 12, 16$ leading factors (weighted by λ_j) as shown in right hand side panel of Fig. 2. The gradual face "build-up" as generated by local parts is much more apparent with pNTF than with any of the other decompositions.

For a quantitative measure of performance, we used the leading five factors as convolution kernels for obtaining a vector of measurements per input images. The part buildup that we see in Fig. 2 suggests that the pNTF factors potentially form a "relevant" set of features. An image was mapped into a vector of five entries by convolution with the leading factors. The measurement vectors were then fed into a SVM classifier (Boser et al., 1992) with a RBF kernel for classifying faces versus non-faces. The decomposition was performed on a subset of 429 face images chosen randomly from the dataset, the training of the SVM was performed on a subset of 1000 images of the dataset (different from the decomposition set) and the testing was done on the remaining 1000 images. Fig. 3a shows the ROC curves of the SVM using the features produced by pNTF, NTF and EM (tempered-EM produced the same ROC as EM). One can see that the pNTF generalizes the best of the three and this is consistent with the previous qualitative judgement about the sharpness (and relevance) of the factors. We then performed the experiment using the next five factors (numbered $6 - 10$). If indeed the pNTF factors are more "relevant" than we would expect the performance to degrade more rapidly with non-leading factors. Fig. 3b supports this notion as the ROC using the EM features is much better than pNTF.

Our second experiment is in the domain of text analysis. We used the reduced version of the "20newsgroups" dataset[2]. Documents are represented by 100 dimensional binary co-occurrence vectors and tagged as a member of one out of four classes "computers", "recreational activities", "science" and "talks". Documents contain approximately 4% of the words, averaged across the 16242 postings. We performed pNMF,NMF,EM and tempered-EM to recover four factors which were used as filter measurements to an SVM classifier.

To gain a better understanding to the nature of the different solutions we first show in Fig. 5 the leading (high value/probability)

[2] http://www.cs.toronto.edu/∼roweis/data.html



comp.*							
pNMF		**NMF**		**EM**		**TEM**	
word	prob.	**word**	prob.	**word**	prob.	**word**	prob.
email	15	email	13	email	15	email	15
university	10	university	10	help	12	data	10
graphics	8	graphics	8	graphics	9	files	10
help	8	computer	8	number	9	graphics	10

rec.*							
pNMF		**NMF**		**EM**		**TEM**	
word	prob.	**word**	prob.	**word**	prob.	**word**	prob.
car	24	car	30	car	24	car	27
drive	10	drive	9	power	15	problem	11
dealer	8	engine	8	question	9	drive	10
engine	8	dealer	8	problem	7	engine	10

sci.*							
pNMF		**NMF**		**EM**		**TEM**	
word	prob.	**word**	prob.	**word**	prob.	**word**	prob.
question	19	problem	21	university	17	program	20
system	13	power	18	program	12	course	16
power	11	state	16	image	9	version	12
number	8	solar	6	science	7	state	11

talk.*							
pNMF		**NMF**		**EM**		**TEM**	
word	prob.	**word**	prob.	**word**	prob.	**word**	prob.
god	14	god	10	god	11	god	10
problem	7	question	7	course	8	course	10
course	7	course	7	fact	8	evidence	8
fact	7	fact	6	world	6	system	8

Figure 5: *comparing the most probable words contained in the four factors of the newsgroup data-set.*

entries (corresponding to the most popular words) of the four factors for each of the four decomposition method. All methods produce plausible results in the sense that the most popular words indeed correspond to what a Layman would consider appropriate for the topic — hence a visual inspection of the factors does not reveal useful information for comparing the different methods. Next we took a new ”science” posting (represented as a binary vector $\mathbf{g}$ with 100 entries normalized to unit L_1 norm) and reconstructed it using the four factors. One would expect that the coefficient corresponding to the ”science” factor would have a high value if $\mathbf{g}$ is a science posting and low if $\mathbf{g}$ is some other posting. By running over all ”science” postings we obtain a graph, one for each method, of the ”science” factor coefficient value. Likewise, by running over all other postings (non ”science”) and plotting the ”science” coefficient value (a graph per method) we obtain four graphs whose values should be low. The ”winning” method should have a consistent high value in the first graph and a consistent low value in the second graph. As for the details of reconstruction, in pNMF we solve the problem $\min \|A\mathbf{x} - \mathbf{g}\|^2$ under the simplex conditions $\mathbf{x} \geq 0$ and $\|\mathbf{x}\|_1 = 1$. For NMF we solve the same criterion function but under $\mathbf{x} \geq 0$ only. For EM we minimize under the relative entropy loss with the simplex constraint.

Fig. 4 shows the graphs corresponding to the value of the ”science” factor coefficient in the reconstruction. One can see that over all ”science” postings the pNMF coefficient is consistently the highest, and over all non ”science” postings the pNMF coefficient is consistently the lowest. This result agrees with the face image dataset results we obtained above where it appears that the pNTF factors are ”sharper” and more ”relevant” than the factors obtained by the other decomposition methods.

6 Summary

There are two main points in our work. First is the justification of an L_2 loss for probabilistically valid decompositions of co-occurrence data. The L_2 loss for a linear model is well understood (ML under Gaussian additive noise) but for co-occurrence data, which is by nature non-linear, we have shown that under general additive noise (not necessarily Gaussian) the approximations commensurate well with the magnitude of the noise — unlike a ML solution. It is worthwhile noting that our framework is constrained as the solutions must lie in the probability simplex thus one should expect a different behavior compared to standard unconstrained L_2 approximations (where L_2 has a ”bad reputation” with respect to outliers for example). The length of the paper does not allow us to explore this avenue in more details — such as the role of an ”Euclidean” entropy for driving sparse solutions and so forth. The second point of our work is deriving an update rule, replacing the Lee-Seung update rule for NMF/NTF, which would guarantee a probabilistically valid (locally optimal) solution.

References

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifers. *Proc. of the 5th ACM Workshop on Computational Learning Theory* (pp. 144–152). ACM Press.

Ding, C., Li, T., & Peng, W. (2006). Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence, chi-square statistic, and a hybrid method. *Proc. of AAAI National Conf. on Artificial Intelligence (AAAI).*

Dykstra, R. (1983). An algorithm for restricted least squares regression. *J. of the Amer. Stat. Assoc.*, *78*, 837–842.

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. San Diego, USA.

Gaussier, E., & Goutte, C. (2005). Relation between plsa and nmf and implications. *SIGIR '05: Proceedings of the ACM SIGIR conference on Research and development in information retrieval* (pp. 601–602). New York, NY, USA.

Hazan, T., Polak, S., & Shashua, A. (2005). Sparse image coding using a 3d non-negative tensor factorization. *Proceedings of the International Conference on Computer Vision*. Beijing, China.

Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proc. of Uncertainty in Artificial Intelligence, UAI'99*. Stockholm.

Lee, D. D., & Seung, H. S. (1997). Unsupervised learning by convex and conic coding. *Proceedings of the conference on Neural Information Processing Systems (NIPS).*

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Neumann, J. V. (1950). *Functional operators vol. ii*. Princeton University Press.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Envirometrics*, *5*, 111–126.

Quelhas, P., Monay, F., Gatica-Perez, D., Tuytelaars, T., & Gool, L. V. (2005). Modelling scenes with local descriptors and latent aspects. *Proceedings of the International Conference on Computer Vision*. Beijing, China.

Shashua, A., & Hazan, T. (2005). Non-negative tensor factorization with applications to statistics and computer vision. *Proceedings of the International Conference on Machine Learning (ICML).*

Shashua, A., & Levin, A. (2001). Linear image coding for regression and classification using the tensor-rank principle. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Hawaii.

Sivic, J., Russel, B., Efros, A., Zisserman, A., & Freeman, W. (2005). Discovering objects and their location in images. *Proceedings of the International Conference on Computer Vision*. Beijing, China.

Welling, M., & Weber, M. (2001). Positive tensor factorization. *Pattern Recognition Letters*, *22*, 1255–1261.

A Proofs of Propositions

Proof of Proposition 1:

$$\begin{aligned} P^*_\epsilon &= \text{argmin}_{P\in\mathcal{Q}} \|P_0 + E - P\|_F^2 \\ &= \text{argmin}_{P\in\mathcal{Q}} \|P_0 - P\|_F^2 - 2\langle P, E\rangle, \end{aligned}$$

where $\langle A, B\rangle = \sum_{i,j} A_{i,j}B_{i,j}$ denote the inner product operation. As $P_0 \in \mathcal{Q}$ and P^*_ϵ attains the minimal value we infer $\|P_0 - P^*_\epsilon\|_F^2 - 2\langle P^*_\epsilon, E\rangle \leq \|P_0 - P_0\|_F^2 - 2\langle P_0, E\rangle$ which implies by the Hölder inequality $|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|_1\|\mathbf{y}\|_\infty$ that

$$\|P_0 - P^*_\epsilon\|_F^2 \leq 2\langle P^*_\epsilon, E\rangle - 2\langle P_0, E\rangle \leq 4\|E\|_\infty$$

and the norm inequality $\|P_0 - P^*_\epsilon\|_\infty \leq \|P_0 - P^*_e\|_F$ establishes the proof. □

Proof of Proposition 2: Set $\gamma = \mathbf{1}^\top E\mathbf{1}$ and $\mathbf{e} = E\mathbf{1}$, then $\mathbf{p}^*_e = (\mathbf{p}_0 + \mathbf{e})/(1+\gamma)$ and

$$|p_{0i} - p^*_{e\,i}| = \left| p_{0i}\frac{\gamma}{1+\gamma} - \frac{\mathbf{e}_i}{\gamma}\right|$$

We set $\alpha_1 = \|\mathbf{e}\|_\infty/(1+\gamma)$ and bound $\|\mathbf{p}_0 - \mathbf{p}^*_\epsilon\|_\infty$ using the relation $|x| - |y| \leq |x - y| \leq |x| + |y|$. In a similar manner we derive the bound on $\|\mathbf{q}_0 - \mathbf{q}^*_\epsilon\|_\infty$.
Set $\delta = \mathbf{p}^*_e - \mathbf{p}_0$ and $\mu = \mathbf{q}^*_e - \mathbf{q}_0$ then for every i, j hold $|(\mathbf{p}_0\mathbf{q}_0^\top - \mathbf{p}^*_\epsilon {\mathbf{q}^*_\epsilon}^\top)_{i,j}| = |p_{0i}q_{0j} - p^*_{e\,i}q^*_{e\,j}| = |p_{0i}\mu_j + q_{0j}\delta_i + \delta_i\mu_j| \leq 3\max\{|\delta_i|, |\mu_j|\}$ since $p_{0i}, q_{0j}, \delta_i, \mu_j \leq 1$. □

Proof of Proposition 3: The expectation of uniformly chosen perturbation elements $\epsilon' \leq E_{ij} \leq \epsilon$ is $(\epsilon - \epsilon')/2$. To compute the expectation of $E\mathbf{1}$ and $\mathbf{1}^\top E\mathbf{1}$ we introduce the collection of indexes $\mathcal{F} = \{(i,j) :\ {p_0}_i {q_0}_j < \epsilon\}$ and its power $\beta = |\mathcal{F}|$.

$$\begin{aligned} Exp[\sum_{ij} E_{ij}] &= \sum_{ij \in \mathcal{F}} \frac{\epsilon - {p_0}_i {q_0}_j}{2} \geq \frac{\epsilon\beta - 1}{2} \\ Exp[\sum_{j} E_{i_0 j}] &= \sum_{j:\ (i_0, j) \in \mathcal{F}} \frac{\epsilon - {p_0}_{i_0} {q_0}_j}{2} \leq \frac{\epsilon \cdot \min\{\beta, n\}}{2} \end{aligned}$$

Since independent trials are concentrated around their mean, we deduce that with high probability

$$\frac{E\mathbf{1}}{1 + \mathbf{1}^\top E\mathbf{1}} \leq \frac{\epsilon \cdot \min\{\beta, n\}}{\epsilon\beta + 1} \qquad \text{whenever} \quad n \to \infty$$

□