pLSA for Sparse Arrays With Tsallis Pseudo-Additive Divergence: Noise Robustness and Algorithm

Tamir Hazan Roee Hardoon Amnon Shashua
School of Computer Science and Engineering
Hebrew University of Jerusalem

Abstract

We introduce the Tsallis divergence error measure in the context of pLSA matrix and tensor decompositions showing much improved performance in the presence of noise. The focus of our approach is on one hand to provide an optimization framework which extends (in the sense of a one parameter family) the Maximum Likelihood framework and on the other hand is theoretically guaranteed to provide robustness under clutter, noise and outliers in the measurement matrix under certain conditions. Specifically, the conditions under which our approach excels is when the measurement array (co-occurrences) is sparse — which happens in the application domain of "bag of visual words".

1. Introduction

We introduce a robust version of pLSA which is the problem of factorizing a multi-way array (matrix or tensor) into a linear combination of rank-1 factors subject to probability (simplex) constraints. We are mostly interested in applications where the input array (representing co-occurrances between features and images, or between words and documents) is *sparse* and in this context would like to obtain decompositions into factors in a way which is insensitive to additive noise, clutter and outliers. In the context of feature/image or word/document associations noise takes the form of clutter images (outside of the object classes of interest) and irrelevant feature fragments. We also wish to remain with the Maximum Likelihood (ML) framework which the original pLSA provides. The line between obtaining more robust solutions than original pLSA while not deviating much from ML solutions can be treaded carefully by working with an extended divergence measure, known as the Tsallis divergence [8], which is a one parameter extension of relative entropy. We will show that optimization with Tsallis divergence can be done within an Expectation-Maximization (EM) framework, thus generalizing the original pLSA algorithm, and prove robustness claims applicable to sparse input arrays.

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., $X \perp Y \mid Z$. The standard pLSA [3] is described as the maximum-likelihood factorization of the co-occurrence matrix G into the product $U\Sigma V^\top$:

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

where $D(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log(p_i/q_i)$ is the relative entropy measure (a.k.a KL-divergence). The statistical interpretation of the decomposition is based on the mixture:

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

where $P(Z = z_j)$ form the diagonal of Σ, $P(X \mid Z = z_j)$ are the column vectors of U, and $P(Y \mid Z = z_j)$ are the row vectors of $V^\top$. Optimization under KL-div error guarantees a maximum-likelihood (ML) estimation and can be obtained via the celebrated Expectation-Maximization (EM) algorithm [1].

The pLSA algorithm is closely related to non-negative matrix and tensor factorizations but with two distinctions (i) ML solution is sought after, and (ii) the decomposition is governed by simplex constraints required for obtaining probabilistically valid solutions. The use of pLSA in visual recognition has been gaining attention and in particular where visual analogue of the "bag of words" representations of image features across multiple object classes is used. In this context a factor corresponds to an object class and the factorization of the data matrix, representing co-occurrences of image features and images, decomposes a collection of images over multiple object classes into separate classes with their associated image features [5] — more details on this process can be found in Section 4. The bag of visual words application domain is promising but the level of success relies on the scalability of the process and in turn in the performance in the presence of clutter, noise and out-

liers. The focus of this paper is to show how the pLSA framework can be extended in order to handle noise in a satisfactory manner in the domain of sparse co-occurrence arrays (which is the domain of bag-of-visual-words).

We introduce a matrix/tensor pLSA algorithm governed by the Tsallis one parameter divergence family $D_q(\mathbf{x} \,||\, \mathbf{y}) = (1 - \sum_i x_i^q y_i^{1-q})/(1-q)$. A convenient property of Tsallis divergence is that $D_{q\to 1}(\mathbf{x}||\mathbf{y}) \to D(\mathbf{x}||\mathbf{y})$, i.e., it is one parameter extension of the KL-div. Our contribution is two-fold: we show that a factorization minimizing the D_q energy between a *sparse* measurement tensor and a low-rank statistically admissible model is largely insensitive to measurement outliers when $q \to 0$. Our second contribution is to derive an EM extension for D_q optimization thereby introducing a simple locally-optimum iterative scheme using auxiliary variables.

2. On Tsallis (Non-extensive) Divergence Measure

Tsallis entropy [7] defined below,

$$S_q(\mathbf{x}) = \frac{1 - \sum_i x_i^q}{q-1} \quad (\mathbf{x} \geq 0,\ \sum_{i=1}^{n} x_i = 1)$$

where q is a real parameter, is a generalization of the standard Boltzmann-Gibbs (and Shannon) entropy. In the limit as $q \to 1$, we have that $x_i^{q-1} = e^{(q-1)\ln x_i} \approx 1 + (q-1)\ln x_i$, hence $S_1 = -\sum_i x_i \ln x_i$, which is the normal Shannon entropy.

Tsallis entropy can also be thought of a q-deformation of Shannon's entropy by noting that $S_q(\mathbf{x}) = -\sum_i x_i \ln_q x_i$ where $\ln_q(x) = (x^{1-q} - 1)/(1-q)$ is the q-logarithm with the property $\ln_q(x) \to \ln x$ when $q \to 1$.

As for properties, $S_q \geq 0$, for $q > 0$, and equals to zero when all probabilities but one vanishes; like Shannon's entropy, S_q attains its maximum (for $q > 0$) for uniform distribution ($x_i = 1/n$), thus becoming $S_q = ln_q$.

Finally, S_q is *not extensive* in the sense that given two independent random variables $X \perp Y$, i.e., $P(X,Y) = P(X)P(Y)$, then

$$S_q(X,Y) = S_q(X) + S_q(B) + (1-q)S_q(X)S_q(Y).$$

From this result it is evident that q is a measure of the departure from extensivity. Tsallis relative entropy [8] $D_q(\mathbf{x} \,||\, \mathbf{y})$ can be described as a q-deformation of relative entropy $D(\mathbf{x} \,||\, \mathbf{y})$:

$$D_q(\mathbf{x} \,||\, \mathbf{y}) = -\sum_i x_i \ln_q \frac{y_i}{x_i} = \frac{\sum_i x_i^q y_i^{1-q} - 1}{q-1} \qquad (2)$$

Like the entropy function, $D_q \to D$ in the limit when $q \to 1$. It can also be shown that $D_q(\mathbf{x} \,||\, \mathbf{y}) \geq 0$, for $q > 0$, and vanishes if and only if $\mathbf{x} = \mathbf{y}$. Further, for $q > 0$, D_q is a convex function of $\mathbf{x}$ and $\mathbf{y}$. Like Tsallis entropy, D_q satisfies the pseudo-additivity of the form:

$$D_q(X_1, X_2 \,||\, Y_1, Y_2) = D_q(X_1 \,||\, Y_1) + D_q(X_1 \,||\, Y_1) + (q-1)D_q(X_1 \,||\, Y_1)D_q(X_1 \,||\, Y_1)$$

where X_1, X_2 and Y_1, Y_2 are independent pairs. It is worthwhile noting that the non-extensive nature of S_q and the pseudo-additivity of D_q is a hindrance to using S_q and D_q for statistical inference because it does not allow one to take advantage of the i.i.d. property of observations and thereby one must work with the distribution over the entire training set.

To make this point in some detail, the standard EM algorithm over observations X, model parameters θ and latent variables Z consists of iterating the two steps:

E-step: $Q(\theta, \theta^{(t)}) = \sum_Z P(Z \mid X, \theta^{(t)}) \ln P(X, Z \mid \theta)$
M-step: $\theta^{(t+1)} = \text{argmax}_\theta Q(\theta, \theta^{(t)})$.

If the observation sample X are i.i.d. the conditional expectation can be simplified:

$$Q(\theta, \theta^{(t)}) = \sum_{j=1}^{k} \sum_{i=1}^{m} D(P(z^i = j \mid x^i, \theta^{(t)}) \,||\, P(x^i, z^i = j \mid \theta))$$

where x^i is the i'th observation and $z^i \in \{1, ..., k\}$ is the value of the hidden variable of the i'th observation. The M-step becomes a minimization over θ. The posteriors $w_{ij}^{(t)} = P(z^i = j \mid x^i, \theta^{(t)})$ are updated using the Bayes rule. The point about the non-extensive nature of Tsallis entropy and divergence is that although it is completely valid to replace $\ln$ with $\ln_q$ in the E-step, that does *not carry over* in the simplified i.i.d. version of the conditional expectation, i.e., one cannot replace the relative entropy D with D_q in the simplified form (can be done only in the limit $q \to 1$). Therefore, the use of Tsallis entropy in statistical inference has been limited to the update of the posteriors $w_{ij}^{(t)}$ in Deterministic Annealing EM where the Bayes update rule is replaced with a MaxEnt principle [9] but where Tsallis entropy is used instead of Shannon's [6].

In the section below we will derive an EM version of D_q minimization between a multi-way array (representing an empirical distribution under i.i.d. data samples or co-occurrence array) and a low-rank factorization model. We will argue and prove that the optimization is robust under sparse co-occurrence arrays subject to sparse random noise (outliers) in the limit $q \to 0$ — which makes it advantageous for bag-of-words or visterm representations in computer vision.

3. pLSA under Tsallis Divergence

For clarity of presentation we will present first the derivations for matrix factorizations and later summarize the main steps for higher-valence arrays. The pLSA problem written as an algebraic optimization takes the following form: Given a $d_1 \times d_2$ matrix G representing a co-occurrence array ($G \geq 0$, $\mathbf{1}^\top G \mathbf{1} = 1$) we wish to find a low-rank model: $\sum_{r=1}^{k} \lambda_r \mathbf{u}_r \mathbf{v}_r^\top$ under the probabilistic (simplex) constraints: (i) all parameters are non-negative, and (ii) $\|\boldsymbol{\lambda}\|_1 = 1$, and $\|\mathbf{u}_r\|_1 = \|\mathbf{v}_r\|_1 = 1$ for $r = 1, ..., k$. The low-rank model is then found by minimizing the relative entropy which guarantees a maximum likelihood solution. Instead of relative entropy we will employ Tsallis relative entropy D_q:

$$\min_{\boldsymbol{\lambda}, \mathbf{u}_r, \mathbf{v}_r \geq 0} D_q(G \,||\, \sum_{r=1}^{k} \lambda_r \mathbf{u}_r \mathbf{v}_r^\top) \;\; s.t. \; \|\boldsymbol{\lambda}\|_1 = \|\mathbf{u}_r\|_1 = \|\mathbf{v}_r\|_1 = 1 \tag{3}$$

where $r = 1, ..., k$ and $0 < q < 1$ is a fixed real parameter. We show below that if G is sparse and subject to sparse random additive noise, i.e., $G + E$ for some perturbation matrix E, then as $q \to 0$ the influence of the perturbation E diminishes, i.e., we obtain a robust estimation.

3.1. Performance Bounds in the Presence of Noise

We wish to investigate the sensitivity of D_q minimization in the presence of additive noise as $q \to 0$. The results below show that if G is *sparse* and the perturbation matrix E has its non-vanishing entry locations selected randomly, then the closest admissible solution to $G + E$ under D_q is unique and consists of G itself in the limit when $q \to 0$. Conversely, if G is not sparse then minimization of D_q would lead to multiple global solutions and therefore is not the right energy error (unless $q \to 1$).

Let $\mathcal{P}$ define the set of admissible models (rank-k probabilistic matrices):

$$\mathcal{P} = \left\{ \sum_{r=1}^{k} \lambda_r \mathbf{u}_r \mathbf{v}_r \;\; : \;\; \begin{array}{l} \boldsymbol{\lambda}, \mathbf{u}_r, \mathbf{v}_r \geq 0 \\ \|\boldsymbol{\lambda}\|_1 = \|\mathbf{u}_r\|_1 = \|\mathbf{v}_r\|_1 = 1 \end{array} \right\}$$

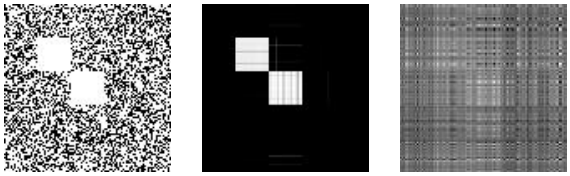
Let $G \in \mathcal{P}$ be an admissible model and let $E \geq 0$ be a perturbation matrix. Note that in case G is sparse, we are still guaranteed that $G + E \geq 0$. Let $\alpha = 1/(1 + \sum_{ij} E_{ij})$ be a normalizing factor and we wish to find $P \in \mathcal{P}$ that minimizes $D_q(\alpha(G + E) \,||\, P)$.

Claim 1 *Let $P \in \mathcal{P}$, then in the limit $q \to 0$ we have:*

$$\lim_{q \to 0} D_q(\alpha(G + E)||P) = 1 - \sum_{i \in supp(G+E)} P_i,$$

where we define the support of a non-negative array $\mathbf{x}$ *as*

$$supp(\mathbf{x}) \stackrel{def}{=} \{i \;:\; x_i > 0\}.$$



(a) (b) (c)

Figure 1. Illustration of Tsallis-divergence and relative-entropy rank-k approximation with 60% random noise: (a) $G + E$, (b) closest rank-k for Tsallis divergence with $q = 0.05$, (c) ML solution.

Proof: follows from the definition of D_q in eqn. 2.□
The immediate implication is that if $G + E$ is *not sparse* then $\lim_{q \to 0} D_q(\alpha(G + E)||P) = 0$ for *all* $P \in \mathcal{P}$. To get a better glimpse into what really happens when G is sparse, let $\mathcal{H}_q$ denote the solution set for a fixed $q > 0$:

$$\mathcal{H}_q = \{P^* : \; P^* = \text{argmin}_{P \in \mathcal{P}} D_q(\alpha(G + E)||P)\}$$

and let $\mathcal{H}_0$ stand for $\mathcal{H}_{q \to 0}$. We can state the following corollary:

Corollary 1

$$\lim_{q \to 0} D_q(\alpha(G + E)||P) = 0$$

iff $supp(P) \subseteq supp(G + E)$ from which it follows that

$$\mathcal{H}_0 = \{P \in \mathcal{P} : \; supp(P) \subseteq supp(G + E)\}$$

We conclude that in general, even if G is sparse, the solution space $\mathcal{H}_0$ is not unique. In fact any matrix whose support is equal or contained in the support of $G + E$ would in the limit $q \to 0$ generate vanishing error to $\alpha(G + E)$:

Corollary 2 *$\forall P \in \mathcal{P}$, if $supp(P) \subseteq supp(G + E)$, then $P \in \mathcal{H}_0$.*

We consider next the situation which guarantees for each q that $\mathcal{H}_q$ to consist of a single member (G itself). Let $\mathcal{H}^g \subseteq \mathcal{H}_0$ be the subset containing all $P \in \mathcal{P}$ whose support is contained in the support of G:

$$\mathcal{H}^g = \{P \in \mathcal{P} : \; supp(P) \subseteq supp(G)\}.$$

Consider the subset $\mathcal{H}_0 \setminus \mathcal{H}^g$ which consists of matrices $P \in \mathcal{P}$ whose support are contained in $supp(G + E)$ but *not in* the support of G. The following claim asserts that G is the only member of $\mathcal{H}_q$ if the set $\mathcal{H}_0 \setminus \mathcal{H}^g$ is empty:

Claim 2 *If $E \perp G$, i.e., the two matrices are disjoint in the locations of vanishing entries, and if $\mathcal{H}_0 \setminus \mathcal{H}^g = \emptyset$, then $\forall P \in \mathcal{H}_0$ we have*

$$D_q\left(\alpha(G + E)||G\right) < D_q\left(\alpha(G + E)||P\right)$$

for all $q > 0$. In particular, in the limit $q \to 0$, G becomes the unique global minimizer: $G = argmin_{P \in \mathcal{P}} D_q(\alpha(G + E) \,||\, P)$.

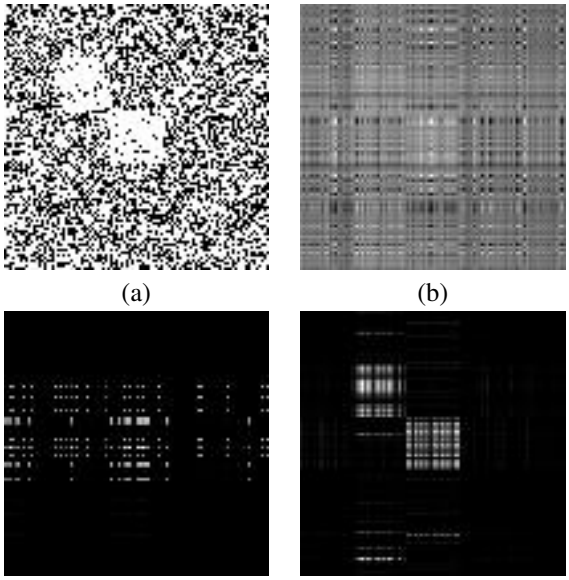


Figure 2. Illustration of Tsallis-divergence and relative-entropy rank-k approximation with 60% random noise where the data was punctured randomly at 10% of its support: (a) $G + E$, (b) ML solution, (c) closest rank-k for Tsallis divergence with $q = 0.05$, (d) closest rank-k for Tsallis divergence with $q = 0.1$. We see how too small q focus on too small support since the data squares are punctured randomly

Proof: see Appendix. □

The material ingredient in Claim 2 is the condition that $\mathcal{H}_0 \setminus \mathcal{H}^g = \emptyset$. This indirectly implies that E is "random" in the sense that "active", i.e., non-vanishing, entries are randomly placed (their actual value is not important). The requirement that E and G are disjoint, i.e., that the noise affects only vanishing entries of the signal G, is important only to make the proof easy but does not seem material in actual experiments.

To conclude, for sparse co-occurrence input matrices with "unstructured" additive noise the global minimizer, under D_q error, reconstructs the original signal in the limit $q \to 0$.

3.2. Iterative Algorithm with Auxiliary Variables

We derive an algorithm, along the lines of the EM machinery, for the optimization problem described in eqn. 3. We denote the optimization function by $L(\theta)$ where θ stands for the unknown parameters $\theta = \{\boldsymbol{\lambda}, \mathbf{u}_r, \mathbf{v}_r\}$, $r = 1, ..., k$. We introduce auxiliary variables in the form of matrices $W^1, ..., W^k$ of the same dimensions as G and define $A \odot B$ as a product array $(A \odot B)_{i,j} = A_{i,j}B_{i,j}$. The auxiliary variables are probability vectors in the sense that $W^r \geq 0$ and $\sum_{r=1}^k W^r_{i,j} = 1$ — we denote this by the shorthand statement $\sum_r W^r = \mathbf{1}$. Define an auxiliary optimization problem:

$$\min_{W,\boldsymbol{\lambda},\mathbf{u}_r,\mathbf{v}_r \geq 0} \; Q(W,\theta) \;\; s.t. \\ \|\boldsymbol{\lambda}\|_1 = \|\mathbf{u}_r\|_1 = \|\mathbf{v}_r\|_1 = 1, \;\; \sum_r W^r = \mathbf{1} \quad (4)$$

where

$$Q(W,\theta) \stackrel{def}{=} \sum_{r=1}^{k} D_q(W^r \odot G \;||\; \lambda_r \mathbf{u}_r \mathbf{v}_r^\top).$$

The relationship between minimizing $Q(W,\theta)$ versus minimizing our target criterion $L(\theta)$ is captured by the following claims.

Claim 3 *$L(\theta) \leq Q(W,\theta)$ for any choice of non-negative $W^1, ..., W^k$ which satisfy $\sum_m W^m = \mathbf{1}$. In particular $L(\theta) = Q(W,\theta) - f(W,\theta)$ for the non-negative function $f(W,\theta)$*

Proof:

$$\begin{aligned} L(\theta) &= -\sum_{i,j} G_{i,j} \ln_q \frac{\sum_m \lambda_m u_{m,i} v_{m,j}}{G_{i,j}} \\ &= -\sum_{i,j} G_{i,j} \ln_q \sum_m W^m_{i,j} \frac{\lambda_m u_{m,i} v_{m,j}}{W^m_{i,j} G_{i,j}} \\ &\leq -\sum_{i,j} G_{i,j} \sum_m W^m_{i,j} \ln_q \frac{\lambda_m u_{m,i} v_{m,j}}{G_{i.j} W^m_{i,j}} = Q(W,\theta) \end{aligned}$$

The last inequality is a direct result of the convexity of $\ln_q(x)$ derived by Jensen's inequality: $-\ln_q \sum_j p_j x_j \leq -\sum_j p_j \ln_q x_j$ when $\mathbf{p} \geq 0$ and $\sum_j p_j = 1$ implying we can turn the log-over-sum into sum-over-log. □

The strategy of EM is to minimize the upper-bound auxiliary function $Q(W,\theta)$ with the hope that if we descend on the upper-bound function we will also descend on $L(\theta)$. To see that this strategy really holds we show that the inequality above becomes an equality $L(\theta) = Q(W^*,\theta)$ for the optimal W:

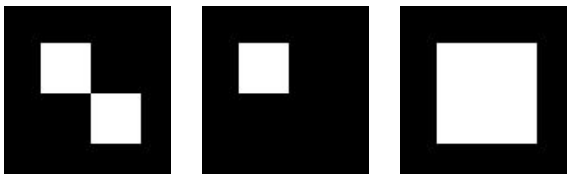
Claim 4 *Let $W^* = argmin_W Q(W,\theta)$. Then*

$$W^{*r}_{i,j} = \frac{\lambda_r u_{r,i} v_{r,j}}{\sum_{s=1}^k \lambda_s u_{s,i} v_{s,j}} \quad (5)$$

and $L(\theta) = Q(W^*,\theta)$.

Proof:

$$\begin{aligned} Q(W^*,\theta) &= \sum_{i,j} G_{i,j} \sum_m W^{*m}_{i,j} \ln_q \frac{\lambda_j u_{m,i} v_{m,j}}{G_{i.j} W^{*m}_{i,j}} \\ &= \sum_{i,j} G_{i,j} \sum_m W^{*m}_{i,j} \ln_q \frac{\lambda_j u_{m,i} v_{m,j}}{G_{i,j} \frac{\lambda_m u_{m,i} v_{m,j}}{\sum_r \lambda_r u_{r,i} v_{r,j}}} \\ &= \sum_{i,j} G_{i,j} \ln_q \frac{\sum_r \lambda_r u_{r,i} v_{r,j}}{G_{i,j}} = L(\theta) \end{aligned}$$



(a) (b) (c)

Figure 3. Illustration of Tsallis-divergence and relative-entropy rank-1 approximation: (a) original image, (b) closest rank-1 for Tsallis divergence with $q = 0.05$, (c) ML solution.

□

It is worthwhile noting that W^* is defined by the Bayes rule, i.e., the value of q does not enter into the formula for updating the auxiliary variables. Therefore, the auxiliary variables play the role of posteriors just as in the ML ($q \rightarrow 1$) scenario.

The point of convergence of an alternating scheme on the unknowns W, θ of $Q(W, \theta)$ is shown below to be a stationary point of $L(\theta)$.

Claim 5 *In an iterative update scheme, If $\theta^{(t)}$ is a stationary point for $Q(W^{(t)}, \theta)$ then it is a stationary point of $L(\theta)$.*

Proof: Consider the function $L(\theta) = Q(W^{(t+1)}, \theta) - f(W^{(t+1)}, \theta)$ while taking into account that $f(W^{(t+1)}, \theta)$ attains its global minima at $\theta^{(t)}$, therefore $\frac{\partial}{\partial \theta} f(W^{(t+1)}, \theta^{(t)}) = 0$. To estimate $\theta^{(t)}$ as a stationary point of $L(\theta)$ we differentiate:

$$\frac{\partial L}{\partial \theta}\left(\theta^{(t)}\right) = \frac{\partial}{\partial \theta} Q(W^{(t+1)}, \theta^{(t)}) - \frac{\partial}{\partial \theta} f(W^{(t+1)}, \theta^{(t)})$$

and vanishing derivative of $Q(W^{(t+1)}, \theta)$ near $\theta^{(t)}$ implies a vanishing derivative of $L(\theta)$ near $\theta^{(t)}$. □

The alternating optimization over the parameters W, θ of $Q(W, \theta)$ involve updating W via the Bayes rule (eqn. 5) and θ via the partial derivatives of $Q(W, \theta)$ with respect to the unknowns $\boldsymbol{\lambda}, \mathbf{u}_r$ and $\mathbf{v}_r$. The update of the parameters θ benefit from the same feature contained in KL-div (the case $q \rightarrow 1$) whereby the non-negativity constraint comes "for free". This is rooted in the result that the minimizer of $\min_{\mathbf{x}} D_q(\mathbf{b} \,||\, \mathbf{x})$ under the simplex constraints $\mathbf{x} \geq 0$ and $\sum_i x_i = 1$ is $\mathbf{x}^* = (1/\sum_i b_i)\mathbf{b}$.

Taken together, the algorithm (we refer to as q-EM) for updating the unknowns until convergence is described below (we omit the derivations):

q-EM Algorithm:

1. *Start with an initial guess for the parameters* $\boldsymbol{\lambda}^{(1)}, \mathbf{u}_r^{(1)}, \mathbf{v}_r^{(1)}$
2. *for* $t = 1, 2, ...$
   (a) $W_{ij}^{r\,(t+1)} \leftarrow \frac{\lambda_r^{(t)} u_{r,i}^{(t)} v_{r,j}^{(t)}}{\sum_{s=1}^{k} \lambda_s^{(t)} u_{s,i}^{(t)} v_{s,j}^{(t)}}$
   (b) $\lambda_r^{(t+1)} \leftarrow \frac{1}{z} \sqrt[q]{\sum_{i,j} (W_{i,j}^{r\,(t+1)} G_{i,j})^q (u_{r,i}^{(t)} v_{r,j}^{(t)})^{1-q}}$ *where z is a normalizing factor to make* $\sum_i \lambda_i^{(t+1)} = 1$.
   (c) *for* $r = 1, ..., k$, *set* $H = W^{r\,(t+1)} \odot G / \lambda_r^{(t+1)}$
      i. $u_{r,i}^{(t+1)} \leftarrow \frac{1}{z} \sqrt[q]{\sum_j H_{i,j}^q v_{r,j}^{(t)\,1-q}}$ *(where z is a normalizing factor).*
      ii. $v_{r,j}^{(t+1)} \leftarrow \frac{1}{z} \sqrt[q]{\sum_i H_{i,j}^q u_{r,i}^{(t+1)\,1-q}}$ *(where z is a normalizing factor).*

3.3. Tensor Factorization Update Formulas

The q-EM algorithm for multi-way arrays (tensors) proceeds as follows. The constrained optimization problem becomes:

$$\min_{\boldsymbol{\lambda}, \mathbf{u}_i^r \geq 0} D_q(G \,||\, \sum_{r=1}^{k} \lambda_r \otimes_i \mathbf{u}_i^r) \;\; s.t. \; \|\boldsymbol{\lambda}\|_1 = \|\mathbf{u}_i^r\|_1 = 1 \quad (6)$$

where $G \in R^{d_1 \times ... \times d_n}$ is an n-way array indexed by $G_{i_1,...,i_n}$ where $1 \leq i_j \leq d_j$ and $\otimes_i \mathbf{u}_i^r$ is a short-hand notation for the outer-product $\mathbf{u}_1^r \otimes \cdots \otimes \mathbf{u}_n^r$. Hence, G is described as a linear combination of k rank-1 tensors. The update formulas of the q-EM algorithm become:

1. *Start with an initial guess for the parameters* $\boldsymbol{\lambda}^{(1)}, \mathbf{u}_i^{r\,(1)}$, $r = 1, ..., k$ *and* $i = 1, ..., n$.
2. *for* $t = 1, 2, ...$
   (a) $W_{i_1,...,i_n}^{r\;(t+1)} \leftarrow \frac{1}{z} \lambda_r^{(t)} \prod_{j=1}^{n} u_{i,i_j}^{r\;(t)}$ *where z is a normalization factor to make* $\sum_r W_{i_1,...,i_n}^{r\;(t+1)} = 1$.
   (b) $\lambda_r^{(t+1)} \leftarrow \frac{1}{z} \sqrt[q]{\sum (W^{r\,(t+1)} \odot G)_s^q (\prod_{j=1}^{n} u_{i,i_j}^{r\;(t)})^{1-q}}$ *where z is a normalizing factor and the summation is over the indexes* $s = (i_1, ..., i_n)$.
   (c) *for* $r = 1, ..., k$, *set* $H = W^{r\,(t+1)} \odot G / \lambda_r^{(t+1)}$
      for $j = 1, ..., n$
      $u_{r,i} \leftarrow \frac{1}{z} \sqrt[q]{\sum H_s^q (\prod_{j=1}^{n} u_{i,i_j}^{r})^{1-q}}$ *where z is a normalizing factor and the summation is over the indexes* $s = (i_1, ..., i_n)$.

4. Experiments

We will begin with a number of experiments whose purpose is didactic in the sense of highlighting the advantages of D_q versus ML optimization ($q \rightarrow 1$). We will then move our attention to a specific real world application using the bag of visual words representation and compare the performance of q-EM and EM.

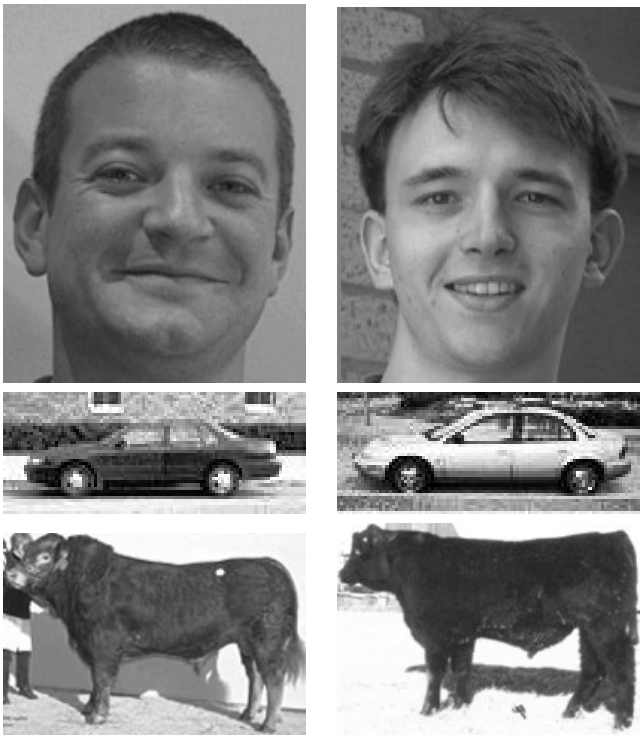


Figure 4. For a real image experiment we constructed image-fragment/images co-occurrences matrix consisting 3000 image-fragments and 20 images from three image classes — faces, cars and cows. Here we present image samples from the database.

To illustrate the theoretical results of Section 3.1, consider a matrix with two uniform blocks (two White blocks in Fig. 1a) with added sparse noise where the amount of noise is significant and stands around 60% of the number of entries. The closest rank-2 matrix under KL-div, illustrated in Fig. 1c, fails to reconstruct the original signal (the two blocks), whereas the D_q with $q = 0.05$ succeeds in reconstructing the original signal (Fig. 1b).

Similar results are obtained when the signal (the two blocks) is not uniform, i.e., each of the blocks is sparse as well. Fig. 2 illustrates this experiment where it is shown that when q is too small the solution focuses only partially on the signal (Fig. 2c) but with a higher q the original signal is largely recovered (Fig. 2d). This experiment also makes the point that the value of q needs to be tuned to the particular characteristics of the signal. Therefore in practice one needs to go through some trial-and-error until the right value of q is found. Note that by setting $q \to 1$ the system falls back to the ML solution.

As a final didactic illustration, Fig. 3 shows the effect of recovering the wrong number of factors. The signal consists of two blocks (without noise) whereas the system is recovering a rank-1 matrix. The ML solution takes the union of the two blocks (Fig. 3c) whereas D_q recovers one of the blocks.

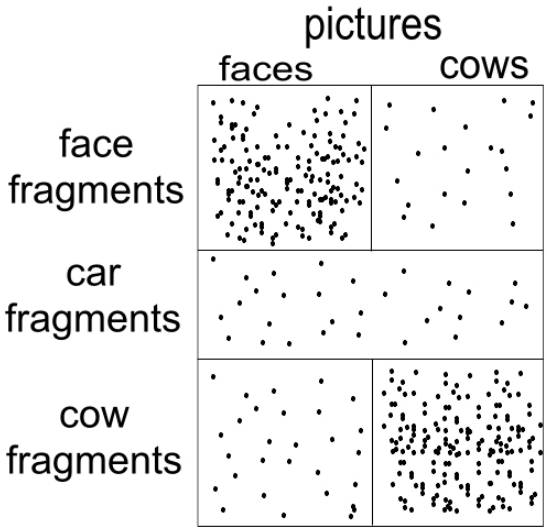
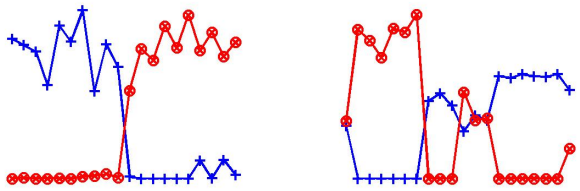


Figure 5. An illustration of the 3000×20 co-occurrence matrix. Two sub-matrices correspond to face-fragments/face-images and cow-fragments/cow-images are more dense than the other parts of the matrix. The cars fragments add sparse irrelevant cluttered data to the occurrence matrix.

For real image experiments we constructed a co-occurrence image-fragments/images matrix (in line with the work of [2, 4, 5]) where the image fragments consisted of a random selection of 3000 rectangular shaped regions with varying size and aspect ratio from three object classes: frontal faces, side-view cows and side-view cars (see Fig. 4). An image-fragment is said to match an image if the cross-correlation between the fragment and the image at the prescribed location of where the fragment was extracted is above threshold.

The co-occurrences were computed between each of the image fragments and 20 images from only two of the object classes: Faces and Cows. In other words, the co-occurrence matrix consisted of a number of *irrelevant* fragments (corresponding to spurious rows) which occasionally have matches with Faces and Cows. Fig. 5 illustrates the structure of the input matrix: rows correspond to fragments and columns correspond to images of Faces and Cows which together form a 3000×20 frequency of occurrence matrix G. Following factorization into a rank-2 model: $\lambda_1 \mathbf{u}_1 \mathbf{v}_1^\top + \lambda_2 \mathbf{u}_2 \mathbf{v}_2^\top$ the vectors $\mathbf{u}_i, \mathbf{v}_i$ should contain information about object classes and relevance of image fragments to the classes.

In particular, $\mathbf{v}_1, \mathbf{v}_2$ should contain the distribution of the 20 images to object classes. If all goes well one should observe a concentration of energy (high values) along the entries associated with images of a single object class (Faces or Cows). Fig. 6a shows $\mathbf{v}_1, \mathbf{v}_2$ as recovered by q-EM where one can clearly see the sharp split between the two object classes compared to EM reconstruction shown in Fig. 6b.



(a) (b)

Figure 6. The vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ of the decomposition $G \approx \lambda_1 \mathbf{u}_1 \mathbf{v}_1^\top + \lambda_2 \mathbf{u}_2 \mathbf{v}_2^\top$ where G is the co-occurrence matrix between fragments (rows) and images (columns). The values of $\mathbf{v}_1$ are marked by '+' and those of $\mathbf{v}_2$ are marked by 'o' in the display. (a) Reconstruction under D_q with $q = 0.1$ showing a sharp split between images of Faces and of Cows. (b) Reconstruction under KL-div is much more sensitive to the clutter introduced by the non-relevant Car fragments.

Figure 7. 20 additional random images of nature and urban scenes were added, and we assembled a 3000×40 fragments/images matrix by adding the 20 random images to the 3000×20 original co-occurrence matrix.

This result illustrates the robustness of D_q optimization compared to the sensitivity of the ML solution to the outliers introduced by the Car fragments.

We next added spurious images as additional 20 columns to G creating an extended 3000×40 matrix G'. Those images were taken from various Nature and Urban scenes (see Fig. 7). The fragments have occasional hits with those images thus creating an additional disruption to frequency measurements of our original two classes of Faces and Cows. The structure of G' is illustrated in Fig. 8. Fig. 9 shows the recovered vectors $\mathbf{v}'_1$ and $\mathbf{v}'_2$ corresponding to the distribution of the 40 images across two factors (object classes). If all goes well we expect $\mathbf{v}'_i = (\mathbf{v}_i, 0)$ where the vanishing entries correspond to the additional 20 spurious images. One can see that the factors recovered by the D_q optimization largely is invariant to the spurious images (Fig. 9 top row) compared to the ML solution where $\mathbf{v}'_i$ is very different from $\mathbf{v}_i$ as a result of the disruption introduced by the spurious images.

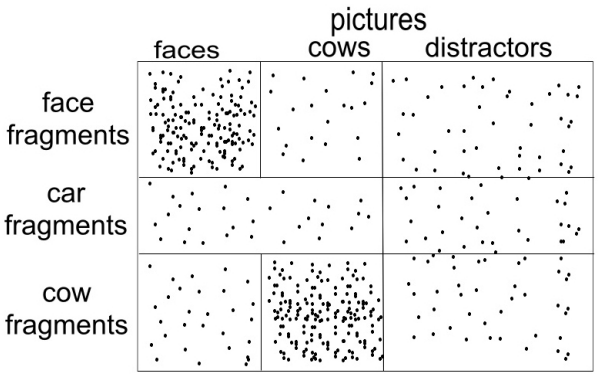


Figure 8. An illustration of the 3000×40 co-occurrence matrix. The 20 random matrices act as distractors and add 20 image columns of random sparse noise into the original 3000×20 co-occerrence matrix.

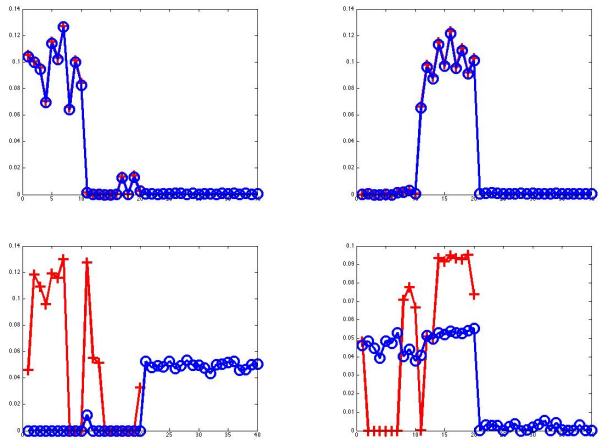


Figure 9. Top row are the Dq factor overlap the small original factors (marked with +) and large factors after adding noised rows and columns(o).. Bottom row are the KL factors overlap the small original factors (marked with +) and large factors after adding noised rows and columns(o).

Next we used the factors (recovered from G) for classification comparing the classification performance of q-EM to EM. We followed the classification scheme of [5] as follows. We used the 20 training images and recovered from the 3000×20 fragment/image frequency array G the leading two factors $\lambda_j \mathbf{u}_j \mathbf{v}_j^\top$, $j = 1, 2$. The vectors $\mathbf{u}_j$ (representing the fragment axis) form the columns of a matrix A and each test image forms a vector $\mathbf{b}$ where $\mathbf{b}$ contains the fragment frequencies matched to the test image. Solving $\min_{\mathbf{x} \geq 0} D(\mathbf{b} || A\mathbf{x})$ subject to $\sum_i x_i = 1$ provides a 2D weight vector associating the test image to each of the object categories. Statistically $\mathbf{b}$ represents $P(fragment \mid test\ image)$, the matrix A constructed from $\mathbf{u}_j$ are the learned fragment distribution for the latent topics $P(fragment \mid topic = j)$ and x_j are the posterior of the topic given the test image $P(z_j \mid test\ image)$. The classi-

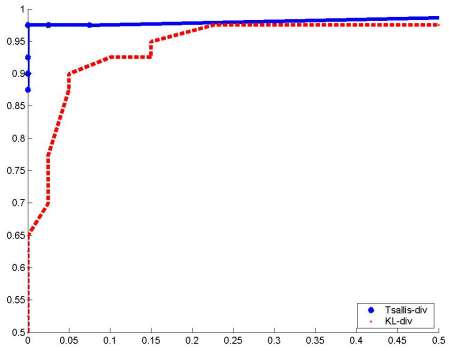


Figure 10. *ROC curves of classification over 80 test images. The success rate of q-EM is 98.75% vs 91.25% the success rate of EM.*

fication decision is based on the object associated with the recovered posteriors. Fig. 10 shows the ROC curve over 80 test images. One can clearly observe that the factors recovered from D_q optimization with $q = 0.1$ provide a much better classification performance of 98.75% compared to 91.25% from the factors recovered using ML.

5. Summary

We have introduced an extended pLSA algorithm called q-EM based on optimization over Tsallis divergence D_q providing a one-parameter extension of KL-div. We have shown that application domains generating sparse co-occurrence matrices, such as when constructing frequency arrays matching features to images or words to documents, there is a benefit to D_q for $q \to 0$ in the sense of robustness against additive noise. We have illustrated the theoretical analysis with both synthetic and real image experiments showing that factors recovered under D_q error provide much more meaningful information compared to the ML solution (when $q \to 1$). The difference is striking in the presence of clutter generated by spurious images and spurious image features — a situation which is likely to occur in real applications. The advantage of the q-EM scheme is that it is an *extension* to the existing approaches in the sense of having a one-parameter tunable dimension to allow the solution to tune into the specific characteristics of the signal. At the current stage of this work the value of q needs to be set by trial and error but we believe that with future work more insight to the relationship between signal characteristics and the value of q can be achieved.

References

[1] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Stat. Soc., series B*, 39(1):1–38, 1977. 2

[2] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, USA, June 2005. 7

[3] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999. 2

[4] P. Quelhas, F. Monay, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modelling scenes with local descriptors and latent aspects. In *Proceedings of the International Conference on Computer Vision*, Beijing, China, Oct. 2005. 7

[5] J. Sivic, B. Russel, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *Proceedings of the International Conference on Computer Vision*, Beijing, China, Oct. 2005. 2, 7, 8

[6] K. Tabushi and J. Inoue. Improvement of em algorithm by means of non-extensive statistical mechanics. In *Neural Networks for Signal Processing XI*, pages 133–142, 2001. 3

[7] C. Tsallis. Possible generalization of bolzmann-gibbs statistics. *J. of Math. Phy.*, 52:479–487, 1988. 3

[8] C. Tsallis. Generalized entropy-based criterion for consistent testing. *Phys. Rev. E.*, 58:1442–1445, 1998. 2, 3

[9] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks.*, 11:271–282, 1998. 3

A. Proof of Propositions

Claim 2 If $E \perp G$, i.e., the two matrices are disjoint in the locations of vanishing entries, and if $\mathcal{H}_0 \setminus \mathcal{H}^g = \emptyset$, then $\forall P \in \mathcal{H}_0$ we have

$$D_q\left(\alpha(G+E)||G\right) < D_q\left(\alpha(G+E)||P\right)$$

for all $q > 0$. In particular, in the limit $q \to 0$, G becomes the unique global minimizer: $G = \text{argmin}_{P\in\mathcal{P}} D_q(\alpha(G+E) \,||\, P)$.

Proof:
Orthogonality $E \perp G$ implies decomposability $D_q(\alpha(G+E)||P) = D_q(\alpha G||P) + D_q(\alpha E||P)$. In addition, for every $P \in \mathcal{H}^g$ holds $D_q(\alpha E||P) = 0$, and by the assumption $\mathcal{H}_0 \setminus \mathcal{H}^q = \emptyset$ we derive $D_q(\alpha E||P) = 0$ for every $P \in \mathcal{H}_0$ and reduce

$$\min_{P\in\mathcal{H}_0} D_q(\alpha(G+E)||P) = \min_{P\in\mathcal{H}_0} D_q(\alpha G||P)$$

The proof is concluded as the solution of $\text{argmin}_{\mathbf{x}} D_q(\mathbf{b}||\mathbf{x})$ under the convex constraints $\mathbf{x} \geq 0$ and $\sum_i x_i = 1$ is $(1/\sum_i b_i)\mathbf{b}$. □