Manifold Preserving Hierarchical Topic Models for Quantization and Approximation

Minje Kim MINJE@ILLINOIS.EDU
Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA

Paris Smaragdis PARIS@ILLINOIS.EDU
University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA
Adobe Research, Adobe Systems Inc., San Francisco, CA 94103, USA

Abstract

We present two complementary topic models to address the analysis of mixture data lying on manifolds. First, we propose a quantization method with an additional mid-layer latent variable, which selects only data points that best preserve the manifold structure of the input data. In order to address the case of modeling all the in-between parts of that manifold using this reduced representation of the input, we introduce a new model that provides a manifold-aware interpolation method. We demonstrate the advantages of these models with experiments on the hand-written digit recognition and the speech source separation tasks.

1. Introduction

Probabilistic topic models have been widely used for various applications, such as text analysis (Hofmann, 1999b;a; Blei et al., 2003), recommendation systems (Popescul et al., 2001), visual scene analysis (Cao & Fei-Fei, 2007), and music transcription (Smaragdis et al., 2006; Févotte et al., 2009). A common intuition behind such models is that they seek a convex hull that wraps the input M-dimensional data points in the $M-1$ dimensional simplex. The hull is defined by the positions of its corners, also known as basis vectors, whose linear combinations reconstruct the inputs inside the hull.

Although these linear decomposition models provide compact representations of the input by using the learned convex hull, an ambiguity exists: the hull loses the data manifold structure as it redundantly includes areas where no training data exist. This is problematic especially when the input is a mixture of distinctive data sets with heterogeneous manifolds. In this case, the desirable outcome of this analysis is not only to approximate the input, but to separate it into its constituent parts, which we will refer to as sources. In text these could be sets of topics, in signal processing they could be independent source signals, etc.

Without knowing the nature of each source, the separation task is ill-defined. Hence, it is advantageous to start with learned sets of basis vectors. Each set approximates the training data of a particular source. Figure 1 depicts a separation result using Probabilistic Latent Semantic Indexing (PLSI) (Hofmann, 1999b;a). In this example two data sets are modeled using their four-cornered convex hulls (red and blue dashed polytopes) as computed by PLSI, respectively. Once confronted with a new data point that is a linear mixture of these two classes (black square) we can decompose it using the already-known models. As seen

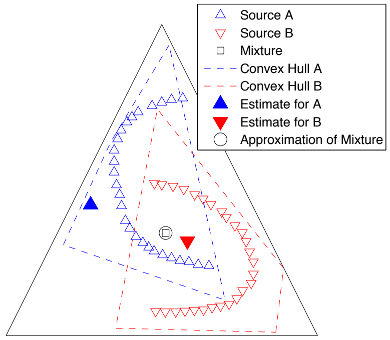


Figure 1. Separation using convex hulls of sources that are learned from PLSI.

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

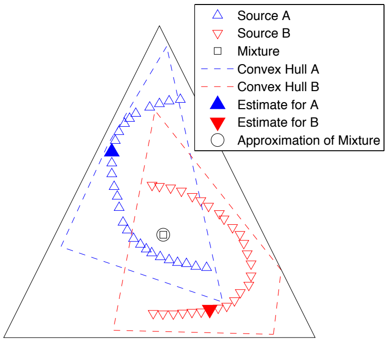


Figure 2. Separation by sparse coding on all the training data points.

in the simulation in the figure, the combination of the learned convex hulls can jointly approximate that mixture point very well (black circle), but estimates for the two source points that constitute the mixture (blue and red filled triangles) lie outside of the original two manifolds, thereby providing poor separation of sources. For instance, in the speech separation scenario the separated speeches do not reflect the characteristics of the sources while their mixture sounds a lot like the mixed signal.

If we use all the overcomplete training samples as candidate topics and force them to be activated in a very sparse way, it can be an alternative to the convex hull representation (Smaragdis et al., 2009). Doing so we can prevent reconstructions from being placed in areas away from the data manifold. For instance, in Figure 2, only one training sample from each class participated as a topic in estimating the constituent sources. Thus, the sparsity constraint can confine the source-specific reconstructions to lie on the data manifolds.

In this paper, we propose two hierarchical topic models. First, a quantization method is used to reduce the size of the overcomplete training data. Doing so is important since the overcomplete representation can often necessitate additional memory and computational resources to process larger number of parameters. Quantizing the data helps minimize redundancy in the data while retaining their expressive power. To this end, we introduce an additional latent variable that selects overcomplete candidate topics and use them to replace the entire overcomplete bases.

Second, the sparse topic model cannot always produce good source estimates especially when the training data is not dense enough, or some important part of the data is lost during the sampling procedure. To handle this issue, we propose another middle-layer latent variable, which is also dedicated to activate only selected data points: groups of neighboring data points of the current estimation of sources. That will result in more manifold-preserving reconstructions.

2. Background: Topic Models and Separation with Sparse Coding

2.1. Probabilistic Latent Semantic Indexing

Ordinary topic models, such as PLSI, take a matrix as input, whose column vectors can be seen as observations with multiple entries, e.g. news articles with finite set of words, sound spectra with frequency bin energies, vectorized images with pixel positions, etc. The goal of the analysis is to find out topics $P(f|z)$ and their mixing weights $P_t(z)$ that best describe the observations $X_{f,t}$ as follows:

$$X_{f,t} \sim \sum_z P(f|z)P_t(z), \tag{1}$$

where t, f, and z are indices for observation vectors, elements of a topic, and the latent variables, respectively. The EM algorithm is common to estimate the model parameters, and in this case this works by minimizing the sum of cross entropy between $X_{f,t}$ and $\sum_z P(f|z)P_t(z)$ for all t:

E-step:

$$P_t(z|f) = \frac{P(f|z)P_t(z)}{\sum_z P(f|z)P_t(z)}$$

M-step:

$$P(f|z) = \frac{\sum_t X_{f,t}P_t(z|f)}{\sum_{f,t} X_{f,t}P_t(z|f)}, \quad P_t(z) = \frac{\sum_f X_{f,t}P_t(z|f)}{\sum_{f,z} X_{f,t}P_t(z|f)}.$$

For example, in Figure 1, we can construct the convex hull of source A by taking source A's training data as input $X^A_{f,t}$ and getting $P_A(f|z)$ as four corners of the hull, which are designated by z. $P_t(z)$ is the mixing weight of z-th corner to reconstruct t-th input.

2.2. Sparse PLSI

The t-th data point of the mixture input $X^M_{f,t}$ is an observation drawn from a multinomial distribution, which is a convex sum of multiple sources s:

$$X^M_{f,t} \sim \sum_s P_t(f|s)P_t(s), \tag{2}$$

where t-th source multinomial $P_t(f|s)$, which corresponds to the filled triangles in Figure 1 and 2, can be further decomposed into combination of topics (corners) as in (1) by seeing $P_t(f|s)$ as input:

$$P_t(f|s) \sim \sum_z P_s(f|z)P_t(z|s). \tag{3}$$

As discussed in the previous section, it is convenient to pre-learn the source-specific topics, $P_s(f|z)$. For instance, if we learned several political topics as $P_{s_1}(f|z)$ and medical topics as $P_{s_2}(f|z)$, respectively, we can reconstruct a news article about a medical bill in the council. For a spectrum representing a mixture of speech signals of two different people, we can reconstruct it as a weighted sum of speaker-wise estimates by using each individual's sets of "topic" spectra. In other words, a mixture input $X^M_{f,t}$ can require more than one set of similar topics, as opposed to the traditional use of PLSI where the input is not a mixture of multiple sources.

With the learned and fixed topics per each source $P_s(f|z)$, the rest of the separation analysis consists of inferring global source weights $P_t(s)$ and source-wise reconstruction weights $P_t(z|s)$ using EM:

E-step:

$$P_t(s,z|f) = \frac{P_t(s)P_t(z|s)P_s(f|z)}{\sum_s P_t(s)\sum_{z\in\boldsymbol{z}^{(s)}} P_t(z|s)P_s(f|z)},$$

M-step:

$$P_t(z|s) = \frac{\sum_f X^M_{f,t}P_t(s,z|f)}{\sum_{f,z} X^M_{f,t}P_t(s,z|f)},$$

$$P_t(s) = \frac{\sum_f X^M_{f,t}\sum_{z\in\boldsymbol{z}^{(s)}} P_t(s,z|f)}{\sum_f X^M_{f,t}\sum_s\sum_{z\in\boldsymbol{z}^{(s)}} P_t(s,z|f)}, \quad (4)$$

where $\boldsymbol{z}^{(s)}$ is a set of topic indices for source s.

The sparse PLSI model additionally assumes that the weights $P_t(z|s)$ and $P_t(s)$ are sparse, so that the mixture and source estimation in (2) and (3) try to use less number of sources $P_t(f|s)$ and topics $P_s(f|z)$, respectively. Furthermore, instead of using the corners of the learned convex hull as topics, the sparse PLSI requires the topics to be the source specific training data itself. Consequently, $P_t(z|s)$ has weights on only a very small portion of the training points as active topics. These two properties result in a manifold-preserving source estimate during the separation procedure. Obviously this is a demanding operation as the training data can be a large data set resulting in an unusually high number of topics.

We will discuss about the way of employing sparsity constraints in the EM algorithm more specifically in Section 3.1.

3. The Proposed Hierarchical Topic Models

Although the proposed extensions of PLSI have different applications, both the manifold preserving quantization and interpolation share some structural similarity: an additional latent variable that weeds out unneeded topics during the analysis.

Suppose that we have some observations $X_{f,t}$. We might need to learn both $P(f|z)$ and $P_t(z)$ for training, or can fix the provided $P_s(f|z)$ and learn the encodings only. In the proposed hierarchical models, we first seek a more compact representation of $P(f|z)$ by additionally decomposing them with a new latent variable y as follows:

$$X_{f,t} \sim \sum_y \sum_z P(f|z)P(z|y)P_t(y). \quad (5)$$

Hence, we can say that the linear transformation of topics, $\sum_z P(f|z)P(z|y)$, is a selection process once the selection parameters $P(z|y)$ meet certain criteria.

3.1. Manifold Preserving Quantization

The goal of manifold preserving quantization is to represent the input data with smaller number of samples, each of which can play as a representative topic that well respects locality of the data. Usually, this quantization is to replace the overcomplete training data or their convex hull with smaller number of representatives on the manifold.

First of all, we use the input observation vectors $X_{f,t}$ as our topic multinomials $P(f|z)$ in (5) as they are as our topic multinomials as they are:

$$X_{f,t} \sim \sum_y \sum_{t'} X_{f,t'}P(t'|y)P_t(y), \quad (6)$$

where the new index t' is surely for column vectors of X as well, but introduced to distinguish from the observation indexing as they have a new usage, i.e. fixed topics. And then, we assume the selection parameter $P(z|y)$ has a smaller number of values of y than that of z, so that the selection procedure $\sum_{t'} X_{f,t'}P(t'|y)$ produces less samples than the inputs[1]. Furthermore, assumptions about sparsity of $P(t'|y)$ and $P_t(y)$ along t' and y axes, respectively, can let the learning results respect the manifold structure. For the sparsest case, assume that only t'-th element of $P(t'|y)$ is one while the others are zero. The only activation chooses an input vector as y-th representative sample. After getting the reduced number of topics $P(f|y)$ like this another sparsity constraint on y also forces each topic to represent as many surrounding inputs as possible by itself.

For the inference of the hierarchical latent variable model, we follow the conventional EM approach, but

[1] Note that we get a trivial solution when we set the same number of y as t', i.e. $P(t'|y)$ and $P_t(y)$ being identity matrices.

for each layer sequentially. However, for this particular quantization model, we can skip the first layer EM as the topical parameter $P(f|z)$ is substituted and fixed with the input vectors $X_{f,t'}$, and the other parameter $P_t(t')$ can be trivially reconstructed with the second layer parameters, $P_t(t') = \sum_y P(t'|y)P_t(y)$.

The second layer expected complete data log-likelihood $\langle\mathcal{L}\rangle$ for y is:

$$\begin{aligned}\langle\mathcal{L}\rangle = &\sum_{f,t,t',y} X_{f,t}P_t(t',y|f)\Big\{\ln P(t'|y) + \ln P_t(y)\Big\}\\ &+ \gamma_1 \sum_{t'} \phi_{t'} \ln P(t'|y) + \gamma_2 \sum_{y} \theta_y \ln P_t(y)\\ &+ \lambda_1\{1 - \sum_{t'} P(t'|y)\} + \lambda_2\{1 - \sum_{y} P_t(y)\}\\ &+ \text{Constants},\end{aligned} \tag{7}$$

where Lagrange multipliers λ_1 and λ_2 for ensuring parameters to sum to one are straightforward. The sparsity constraint terms $\gamma_1 \sum_{t'} \phi_{t'} \ln P(t'|y)$ and $\gamma_2 \sum_y \theta_y \ln P_t(y)$ are from Dirichlet priors with sparse parameters ϕ and θ, and γ_1 and γ_2 control the contribution of the sparsity terms in the objective function. One way to introduce sparsity to the Dirichlet hyper parameters is to substitute them with raised parameters,

$$\phi_{t'} = P(t'|y)^{\alpha}, \qquad \theta_y = P_t(y)^{\beta}$$

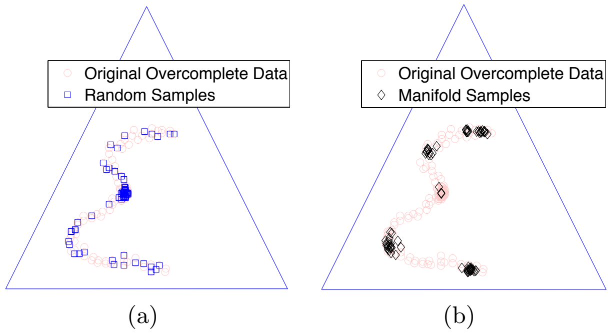
where α and β are some values bigger than 1, and parameters are the estimations from the previous EM iterations. This sparse priors can be intuitively understood, because, for instance, the L_2 norm (when $\alpha = \beta = 2$) of a p.d.f. is maximized when only one value of the variable has probability 1 while the others are 0. Generally, basing on the fact that the parameters have the same L_1 norm, α and β bigger than 1 can make the parameters sparser.

The second layer E-step for y is:

$$P_t(y, t'|f) = \frac{X_{f,t'}P(t'|y)P_t(y)}{\sum_{t'} X_{f,t'} \sum_y P(t'|y)P_t(y)}.$$

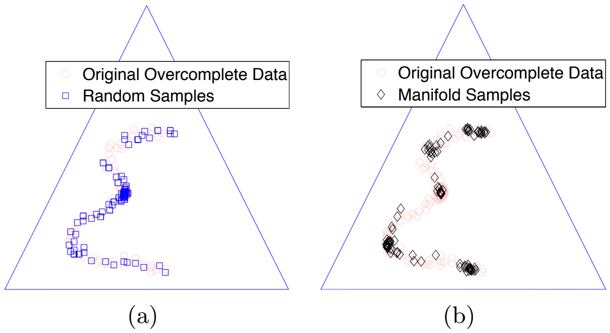
In the second layer M-step we find the solutions that make the partial derivatives of $\langle\mathcal{L}\rangle$ zero, which in turn become update rules as follow:

$$\begin{aligned}P(t'|y) &= \frac{\sum_{f,t} X_{f,t}P_t(t',y|f) + \gamma_1 P(t'|y)^{\alpha}}{\sum_{f,t,t'} X_{f,t}P_t(t',y|f) + \gamma_1 P(t'|y)^{\alpha}},\\ P_t(y) &= \frac{\sum_{f,t'} X_{f,t}P_t(t',y|f) + \gamma_2 P_t(y)^{\beta}}{\sum_{f,t',y} X_{f,t}P_t(t',y|f) + \gamma_2 P_t(y)^{\beta}}.\end{aligned} \tag{8}$$



(a) (b)

Figure 3. The repeatedly (20 times) sampled 4 bases $P(f|y)$ on an ε shaped manifold with (a) random sampling (b) proposed sampling.



(a) (b)

Figure 4. The repeatedly (20 times) sampled 5 bases $P(f|y)$ on a ε shaped manifold with (a) random sampling (b) proposed sampling.

Note that the hyper parameters are replaced at every iteration with the raised previous estimations.

Figure 3 shows the sampling results on ε-shaped manifold. The input exhibits a lower number of data points on the wings as opposed to a higher concentration in the middle. Figure 3 (a) shows the 80 random samples, which consist of four samples that are repetitively drawn 20 times. From the random sampling result (blue squares), we observe that it is very possible to have the all four samples from the center, where the population is highest, but rarely from the wings.

On the other hand, in Figure 3 (b) with the proposed quantization, the four representatives (black diamonds) tend to lie on the two elbows and the two tips, which crucially explain the manifold. They can at least form a c-shape by ignoring the kink while naïve four samples all from the populous central kink give no shape information. Note that only three out of 80 are sampled from the center using the proposed quantization.

However, it is obvious that the fifth sample should be from the kink to complete the full ε-manifold. Figure 4 (b) gives the desired result where the fifth sample successfully represents the central kink, while in Figure 4 (a) the randomly sampled five do not provide such a well-structured quantization results.

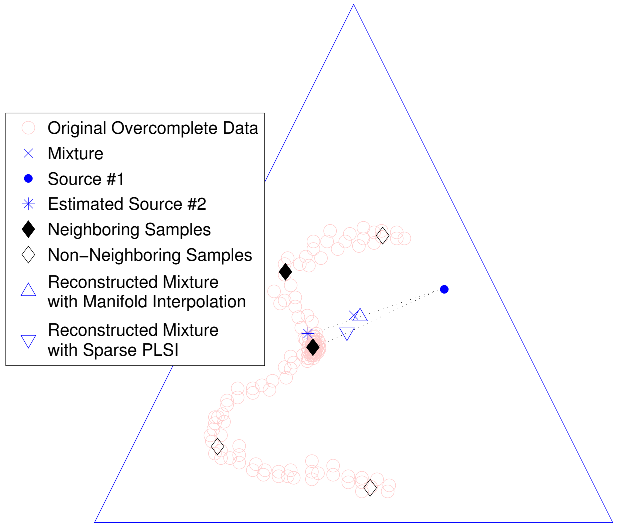


Figure 5. An illustration about the drawback of coupling manifold quantization and the sparse PLSI method. The proposed interpolation method resolves the issue by a local linear combination of samples.

3.2. Manifold Preserving Interpolation

Although the proposed manifold preserving quantization provides a compact representation, the original data points that were in-between those samples might not be modeled as accurately as with the overcomplete data plus sparse PLSI case. Moreover, it is also possible that the original training data might not be as dense in the first place.

Figure 5 describes this situation. Let us assume that we discarded all data points (pink circles) after quantizing them with only five samples (filled or empty diamonds). Given a mixture point (blue cross) and an already estimated source #1 for simplicity, the goal of the sparse PLSI is to select the best one out of the five samples, which eventually reconstructs the mixture with the reverse triangle. On the other hand, the proposed manifold preserving interpolation seeks a linear combination of neighboring samples (filled diamonds), which provides an interpolation for the missing data between the samples (blue star). Note that the estimation of mixture with this approach (triangle) is closer to the input mixture than that from the quantization only.

We start from the same hierarchical topic model introduced in (5). For the t-th mixture vector $X_{f,t}$ the goal is to reconstruct it by combining a few neighbors:

$$X_{f,t} \sim \sum_{s} \sum_{z\in\mathcal{N}_t^s} P_s(f|z)P_t(z|s)P_t(s),$$

where s indicates the sources and $\mathcal{N}_t^s$ is the set of neighboring samples of s-th source estimation for t-th input (the filled diamonds in Figure 5). The selection parameter $P_t(z|s)$ now has the index t to provide weights for each set of neighbors per an input as in (Roweis & Saul, 2000). On the contrary to the previous quantization method, $P_s(f|z)$ is fixed to hold either the overcomplete training data or quantized samples of source s as a set of topics.

Similarly to the previous derivation, we also skip the first layer EM, since $P_s(f|z)$ is fixed, and marginalization of the second layer variable s is trivial.

The second layer complete data log-likelihood for t-th input $\langle\mathcal{L}_t\rangle$ is defined as follows:

$$\begin{aligned}\langle\mathcal{L}_t\rangle = &\sum_{f,z,s} X_{f,t}P_t(z,s|f)\Big\{\ln P_t(z|s) + \ln P_t(s)\Big\}\\ &+\lambda_1\{1-\sum_z P_t(z|s)\} + \lambda_2\{1-\sum_s P_t(s)\}\\ &+\text{Constants},\end{aligned} \tag{9}$$

where λ_1 and λ_2 are Lagrange multipliers for the sum to one constraint as usual. We get the posterior probabilities $P_t(s,z|f)$ from the second layer E-step:

$$P_t(s,z|f) = \frac{P_s(f|z)P_t(z|s)P_t(s)}{\sum_s P_t(s)\sum_{z\in\mathcal{N}_t^s} P_s(f|z)P_t(z|s)}. \tag{10}$$

In the M-step, we find the parameters that maximize $\langle\mathcal{L}_t\rangle$ as follows:

$$\begin{aligned}P_t(z|s) &= \frac{\sum_f X_{f,t}P_t(s,z|f)}{\sum_{f,z} X_{f,t}P_t(s,z|f)},\\ P_t(s) &= \frac{\sum_f X_{f,t}\sum_{z\in\mathcal{N}_t^s} P_t(s,z|f)}{\sum_f X_{f,t}\sum_s\sum_{z\in\mathcal{N}_t^s} P_t(s,z|f)}.\end{aligned} \tag{11}$$

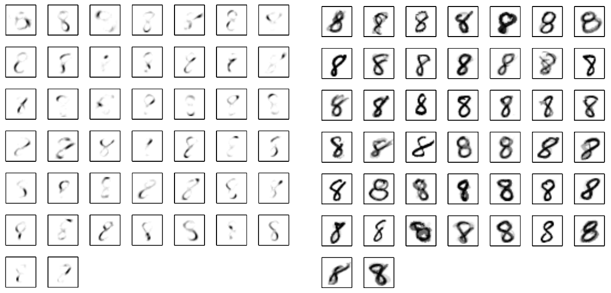
It is obvious that the update rules of the proposed interpolation method eventually become analogous to those of sparse PLSI in (4), but the difference of defining the selection parameter $P_t(z|s)$ makes the proposed method behave uniquely. Instead of imposing sparsity on the completely defined parameter $P_t(z|s)$ for all z indices, the neighbor set $\mathcal{N}_t^s$ lets the procedure focus only on the current neighbors. In other words, $P_t(z|s)$ is not smooth as it is zero for $z \notin \mathcal{N}_t^s$, and so is

$P_t(s, z|f)$ for $z \notin \mathcal{N}_t^s$ in the M-step, consequently. The smaller the number of neighbors is, the more local the reconstruction is. Likewise, in the proposed interpolation model sparse coding is achieved by finding running neighbors at every iteration, which are K-nearest samples from the current estimation of each source for t-th input, $P_t(f|s) = \sum_{z \in \mathcal{N}_t^s} P_s(f|z)P_t(z|s)$:

$$\begin{aligned}\mathcal{N}_t^s = \Big\{ z_k : \mathcal{E}\Big[P_s(f|z_k) \big\| P_t(f|s)\Big] \\ < \mathcal{E}\Big[P_s(f|z' \notin \mathcal{N}_t^s) \big\| P_t(f|s)\Big]\Big\}, \quad (12)\end{aligned}$$

where the integer index $1 \leq k \leq K$, and z' indicates all the possible topics. $\mathcal{E}[A\|B]$ can be any divergence measure, but we use cross entropy, which is a natural choice in the simplex domain,

$$\mathcal{E}[A\|B] = -\sum_i A_i \log B_i. \quad (13)$$


(a)

(b)

Figure 6. Comparison of probabilistic topics and manifold preserving samples. (a) 44 basis topic multinomials learned from ordinary PLSI. (b) 44 manifold samples drawn from the proposed quantization.

3.3. Computational Complexities

Each EM iteration for a mixture vector X_t^M of the sparse PLSI model (4) runs in time $\mathcal{O}(SFZ)$, where S, F, and Z, stand for the number of sources, features, and topics. Therefore, reducing Z to a small set of manifold samples rZ with sampling rate $r < 1$ can mitigate the computational cost of the separation procedure. The quantization can be done beforehand, so its complexity is negligible.

The interpolation method can further reduce rZ to the size of neighbors K. Finding neighbors using (12), which is a sorting operation with $\mathcal{O}(rZ \log rZ)$, is usually a lot less complex than $\mathcal{O}(SFrZ)$ with small sampling rate r. However, since calculating the error function (13) requires additional $\mathcal{O}(SFrZ)$, the complexity of the manifold interpolation is $\mathcal{O}(SFrZ)$, not $\mathcal{O}(SFK)$.

4. Empirical Results

4.1. Quantization of Hand-written Digits

The first experiment is to show the behavior of the quantization model at a sampling rate is 5% – 44 samples out of 876 images of hand-written digit, “8”, from MNIST dataset (LeCun et al., 1998). For the comparison, we learned the same number of ordinary PLSI topics, which are the corners of the convex hull that wraps the images in the digit “8” class.

Figure 6 presents the results. As we can expect, and reported in (Lee & Seung, 1999), ordinary topic models without the concept of sparsity give a parts-based representations of the data, which can be seen as building blocks to be additively combined to reconstruct the input data. It is intuitive as the corners of the convex hull that surround the data points would be more likely to be near the margins, edges, or corners of the simplex, where more elements are suppressed than around the middle of the simplex, so that only several entries of the topic are activated. That is why we see some strokes of the digit “8” in Figure 6 (a) rather than holistic representations.

Although the parts-based representation encourages the model to flexibly combine the topical bases, carefully quantized samples can be better representatives of the data, especially when the original data points lie on high-dimensional manifolds. Figure 6 (b) clearly shows the difference of the proposed sampling method from the results in (a). If we use the samples as the topics of sparse PLSI, which would be sparsely activated to recover unseen inputs, the model can confine the estimation in the manifold of the training data.

4.2. Interpolation for Classifying Handwritten Digits

If the quantization is successful, it can be used instead of the whole training dataset or its convex hull. In this section we employ the proposed manifold interpolation method to recover the missing data between neighboring samples. First, given the 10 digit groups we do classification with 10-fold cross validation. Each class has around 1,000 images. We learn manifold samples from each class at different sampling rates with parameters set to: $\alpha = \beta = 1.2$ and $\gamma_1 = \gamma_2 = 0.001$. For a test handwritten digit, we reconstruct it using the proposed manifold preserving interpolation with several pre-defined number of neighbors[2]. Therefore,

[2] Note that in this case we assume the test input is not a mixture of multiple classes. Hence, we do the EM updates and decision of neighbors for each class separately by fixing s in (10), (11), and (12).

the class, where the test vector is best approximated in terms of cross entropy, is assigned as the estimated class label.

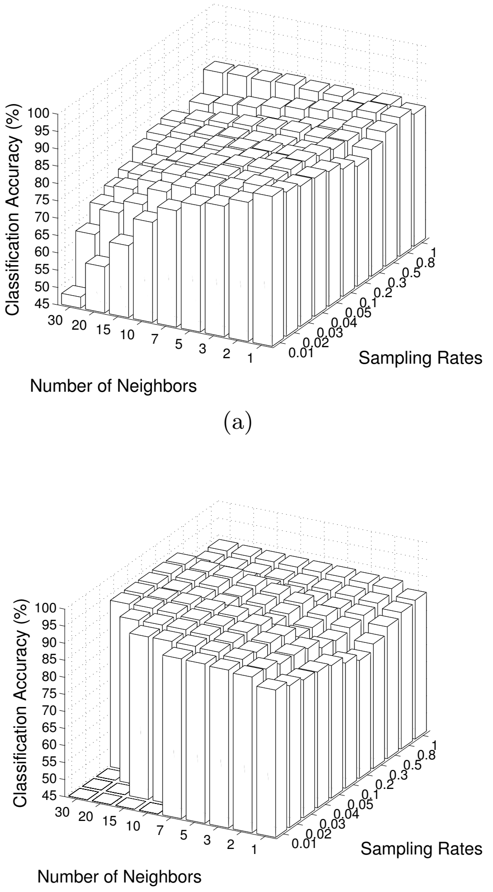
For the comparison, we also conducted a K-Nearest Neighbor (KNN) classification, which uses cross entropy as its divergence measure. During KNN classification, we also consider manifold samples as our training data. For instance, the bars in the back in Figure 7 (a) is the case of 100% sampling rate, where we do the ordinary KNN with the whole training data. We can first check that the proposed quantization performs better than or comparable to (88.2% at sampling rate 1%) the case of using whole data (less than 85%) if we carefully set the number of neighbors.

However, with interpolation we can generally get better classification accuracy, which ranges between 90 to 95% as in Figure 7 (b). Furthermore, we can also resolve the issue of sensitivity to the number of neighbors at the low sampling rates by tying up the neighboring samples to reconstruct the input rather than choosing the best one from here and there. Note that there are cases when the number of neighbors is bigger than that of samples (front-left bars with zero height).

4.3. Quantization of Speech Signals

We further discuss the advantage of the proposed quantization by using speech signals. In this experiment a female speaker is selected from TIMIT speech corpus (Garofolo et al., 1993). Spectrums of a concatenated nine spoken sentences, each of which is 2 to 3 seconds-long, are used as overcomplete training data. The concatenated training signals are converted into the matrix forms, i.e. spectrograms, by using magnitudes of short-time Fourier transform, with 64 ms window size and 32 ms overlaps. Therefore, the training matrix consists of 832 spectra (column vectors), each of which has 513 frequency elements. Now we use the spectrogram matrix X as input vectors to the manifold preserving quantization system. Moreover, we use X as the parameter $P(f|z)$ as they are, and fix them during the process.

Figure 8 shows the sum of the reconstruction errors in terms of cross entropy for the three different systems at seven different sampling rates. For example, when the sampling rate equals to 1%, the number of samples is $8 \approx 0.01 \times 832$. The proposed method produces the samples as a form of sparse linear combination of the whole training points using the selection parameter, $\sum_{t'} X_{f,t'} P(t'|y)$, but provides reconstructions of the input X at the same time as in (6). For the parameters, α, β, γ_1, and γ_2, we once again use the same values as in Section 4.2. We also randomly choose the same number of samples for comparison. After the random sampling, we use the closest sample for each input column as an oracle reconstruction. Lastly, or-



(a)

(b)

Figure 7. Handwritten digits classification results with (a) KNN and (b) the proposed interpolation method.

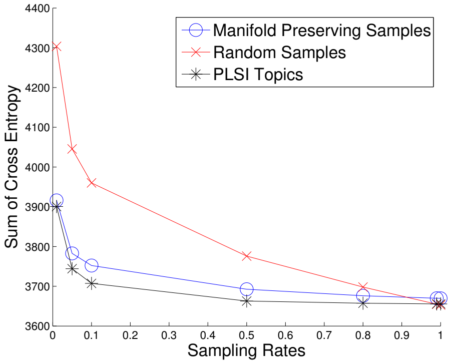


Figure 8. Sum of cross entropy between inputs and the reconstructions from the proposed quantization, oracle random samples, and ordinary PLSI.

dinary PLSI learns same number of topics with the samples. As this is not an experiment for the separation, PLSI's convex hull covers the largest area, so that it provides the best possible reconstruction among the three systems. Note that PLSI does not work well as such, if the inputs are mixture of more than two sources (speakers) as shown in Figure 1. Experiments are repeated 10 times to average out the variance of sampling results.

In the figure, the proposed sampling method can provide representative samples, which recovers input vectors better than the manually chosen closest random samples. All the three methods provide better representation (less cross entropy) as the sampling rate increases, but the proposed sampling provides good performance at low sampling rates, which are better than those of oracle random samples. Also, its results are generally comparable to PLSI topics, which basically can recover the whole convex hull.

4.4. Separation of Crosstalk Using Interpolation

We introduce additional male speaker to build up the crosstalk cancellation problem, on top of the speech from the female speaker. One sentence per a speaker is picked up, and then mixed up. We learn manifold samples at various sampling rates from the spectra of other 9 training sentences of each speaker. They play the role of two sets of pre-learned topics $P_{s=\text{female}}(f|z)$ and $P_{s=\text{male}}(f|z)$. After the EM updates in (10) and (11) plus the neighborhood search (12), we get the converged posterior probabilities $P_t(s, z|f)$, where z is marginalized out to finally get the source-specific posterior probabilities $P_t(s|f)$. For the given t-th input mixture $X^M_{f,t}$, the source s can be recovered by multiplying the resulted posterior, $X^M_{f,t}P_t(s|f)$, and then converting back to the time domain.

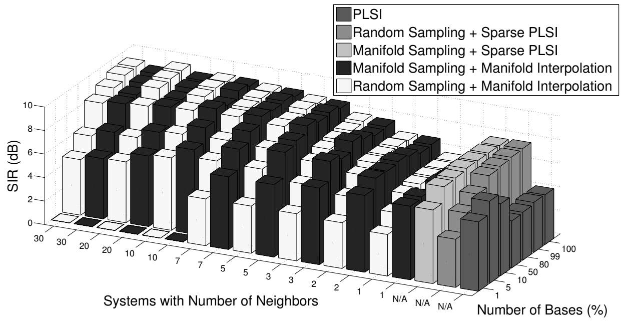


Figure 9. SIR of the crosstalk cancellation results with the proposed quantization and interpolation method compared with random sampling, sparse PLSI, and ordinary PLSI.

Figure 9 shows the separation performance in terms of Signal-to-Interference Ratio (SIR) (Vincent et al., 2006), whose value is zero when the energies of interfering source and the recovered source are same, and can be infinity if the source estimation is perfect. All the algorithms were iterated 100 times. First of all, we use standard PLSI by using the concept in 2.2, but without the sparsity constraint. We set the number of topics, i.e. the corners of PLSI convex hulls, to match the sampling rate. In the shown figure (the right most bars), we can see that PLSI does not give good results with many topics, but its performance is the best (7.0 dB) at around 5% sampling rates (42 topics). On the contrary, sparsity constraints improve the results by around 1 dB (the second and third rightmost bars). It is also noticeable that manifold samples that are only 5% of the entire training data provide almost same separation performance, while random samples start to lost representativeness below 50%.

Both random and manifold samples along with the interpolation techniques further enhance the sparse PLSI results. However, it is also observed that manifold preserving samples are better than random samples at lower sampling rates when they are coupled with interpolation (three frontal bars of manifold interpolation cases). Although as the sampling rate gets higher the merit of manifold sampling vanishes, manifold-preserving interpolation plays a role for the better separation performances (up to about 9.5dB) than both ordinary and sparse PLSI.

5. Conclusion

In this work we proposed a manifold preserving quantization method. By adding a latent variable to the common probabilistic topic model for selecting representatives of overcomplete input, and trying to reduce the reconstruction error at the same time, the method could give better way of compressing the high-dimensional overcomplete data. We showed that the manifold quantization can replace the whole dataset with acceptable loss of approximation power, but with better image classification and speech separation performances compared with existing topic models. On top of that, another model with the explicit neighborhood selection is proposed to compensate the quantization error. This local interpolation technique further improve the classification and separation results whether it is applied to the sampled data or the original entire observations that sometimes does not fully contain the manifold structure of the data.

References

Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

Cao, L. and Fei-Fei, L. Spatially coherent latent topic model for concurrent object segmentation and classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2007.

Févotte, Cédric, Bertin, Nancy, and Durrieu, Jean-Louis. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

Garofolo, John S., Lamel, Lori F., Fisher, William M., Fiscus, Jonathan G., Pallett, David S., Dahlgren, Nancy L., and Zue, Victor. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium, Philadelphia*, 1993.

Hofmann, T. Probablistic latent semantic analysis. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999a.

Hofmann, T. Probablistic latent semantic indexing. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999b.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.

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

Popescul, A., Ungar, L. H., Pennock, D. M., and Lawrence, S. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environment. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2001.

Roweis, S. T. and Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290: 2323–2326, 2000.

Smaragdis, Paris, Raj, Bhiksha, and Shashanka, Madhusudana. A probabilistic latent variable model for acoustic modeling. In *Neural Information Processing Systems Workshop on Advances in Models for Acoustic Processing*, 2006.

Smaragdis, Paris, Shashanka, M., and Raj, B. A sparse non-parametric approach for single channel separation of known sounds. In *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, 2009.

Vincent, E., Fevotte, C., and Gribonval, R. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.