Penalized Probabilistic Clustering

Zhengdong Lu
zhengdon@csee.ogi.edu
Todd K. Leen
tleen@csee.ogi.edu
Department of Computer Science and Engineering, OGI School of Science and Engineering, Oregon Health and Science Institute, Beaverton, OR 97006, U.S.A.

While clustering is usually an unsupervised operation, there are circumstances in which we believe (with varying degrees of certainty) that items A and B should be assigned to the same cluster, while items A and C should not. We would like such pairwise relations to influence cluster assignments of out-of-sample data in a manner consistent with the prior knowledge expressed in the training set. Our starting point is probabilistic clustering based on gaussian mixture models (GMM) of the data distribution. We express clustering preferences in a prior distribution over assignments of data points to clusters. This prior penalizes cluster assignments according to the degree with which they violate the preferences. The model parameters are fit with the expectation-maximization (EM) algorithm. Our model provides a flexible framework that encompasses several other semisupervised clustering models as its special cases. Experiments on artificial and real-world problems show that our model can consistently improve clustering results when pairwise relations are incorporated. The experiments also demonstrate the superiority of our model to other semisupervised clustering methods on handling noisy pairwise relations.

1 Introduction

While clustering is usually executed completely unsupervised, there are circumstances in which we have prior belief (with varying degrees of certainty) that pairs of samples should (or should not) be assigned to the same cluster. These pairwise relations are less informative than direct labeling of the samples, but are often considerably easier to obtain. Indeed, there are many occasions when the pairwise relations can be directly derived from expert knowledge or common sense.

Our interest in such problems was kindled when we tried to manually segment a satellite image by grouping small image clips from the image. It is often hard to assign the image clips to different “groups” since we do not know clearly the characteristic of each group or even how many classes we should have. In contrast, it is much easier to compare two image clips and decide how much they look alike and thus how likely they should be in one cluster. Another interesting example is word sense disambiguation. Ambiguous words like *plant* tend to exhibit only one meaning in one discourse (Yarowsky, 1995). In other words, two *plant*'s in the same discourse probably should be assigned to the same class of sense. This fact is very useful in training unsupervised word sense disambiguation models. The third example is in information retrieval. Cohn, Caruana, and McCallum (2003) suggested that in creating a document taxonomy, the expert critique is often in the form "these two documents shouldn't be in the same cluster." The last example is continuity, which suggests that neighboring pairs of samples in a time series or in an image are likely to belong to the same class of object, is also a source of clustering preferences (Theiler & Gisler, 1997; Ambroise, Dang, & Govaert, 1997). We would like these preferences to be incorporated into the cluster structure so that the assignment of out-of-sample data to clusters captures the concepts that give rise to the preferences expressed in the training data.

Some work has been done on adopting traditional clustering methods, such as K-means, to incorporate pairwise relations (Wagstaff, Cardie, Rogers, & Schroedl, 2001; Basu, Bannerjee, & Mooney, 2002; Klein, Kamvar, & Manning, 2002). These models are based on hard clustering, and the clustering preferences are expressed as hard pairwise constraints that must be satisfied. Wagstaff (2002) and Basu et al. (2004) extended their models to deal with soft pairwise constraints, where each constraint is assigned a weight. The performance of those constrained K-means algorithms is often not satisfactory, largely due to the incapability of K-means to model nonspherical data distribution in each class.

Shental, Bar-Hillel, Hertz, and Weinshall (2003) proposed a gaussian mixture model (GMM) for clustering that incorporates hard pairwise constraints. However, the model cannot be naturally generalized to soft constraints, which are appropriate when our knowledge is only clustering preferences or carries significant uncertainty. Motivated in part to remedy this deficiency, Law, Topchy, and Jain (2004, 2005) proposed another GMM-based model to incorporate soft constraints. In their model, virtual groups are created for samples that are supposed to be in one class. The uncertainty information in pairwise relations is therefore expressed as the soft membership of samples to the virtual group. This modeling strategy is cumbersome to model samples shared by different virtual groups. Moreover, it cannot handle the prior knowledge that two samples are in different clusters. Other efforts to make use of the pairwise relations include changing the metric in feature space in favor of the specified relations (Cohn et al., 2003; Xing, Ng, Jordan, & Russe, 2003) or combining the metric learning with constrained clustering (Bilenko, Basu, & Mooney, 2004).

This letter is a detailed exposition and extension of our previously reported work (Lu & Leen, 2005). We propose a soft clustering algorithm based on GMM that expresses clustering preferences (in the form of pairwise relations) in the prior probability on assignments of data points to clusters. Our algorithm naturally accommodates both hard constraints and soft preferences. In our framework, the preferences are expressed as a Bayesian prior probability that pairs of points should (or should not) be assigned to the same cluster. After training with the expectation-maximization (EM) algorithm, the information expressed as a prior on the cluster assignment of the training data is successfully encoded in the means, covariances, and cluster priors in the GMM. Hence, the model generalizes in a way consistent with the prior knowledge. We call the algorithm *penalized probabilistic clustering* (PPC). Experiments on artificial and real-world data sets demonstrate that PPC can consistently improve the clustering result by incorporating reliable prior knowledge.

The letter is organized as follows. In section 2, we introduce the model and give an artificial example for illustration. In section 3, we discuss the computational complexity and propose two approximation methods for cases where the computation is intractable. Section 4 gives a detailed analysis of the connection of PPC to several other semisupervised clustering models. In section 5, we present the experiments of PPC on a variety of problems, as well as the comparison of PPC to several other semisupervised clustering methods. Section 6 summarizes the letter and points out future research.

2 Model

Penalized probabilistic clustering (PPC) begins with a standard M-component GMM,

$$P(x|\Theta) = \sum_{k=1}^{M} \pi_k \; P(x|\theta_k),$$

with the parameter vector $\Theta = (\pi_1, \ldots, \pi_M, \theta_1, \ldots, \theta_M)$. Here, π_k and θ_k are the prior probability and parameters of the kth gaussian component, respectively. We augment the data set $X = \{x_i\}$, $i = 1, \ldots, N$ with latent cluster assignments $Z = z(x_i)$, $i = 1, \ldots, N$ to form the familiar complete data (X, Z). The complete data likelihood is

$$P(X, Z|\Theta) = P(X|Z, \Theta)P(Z|\Theta), \tag{2.1}$$

where $P(X|Z, \Theta)$ is the probability of X conditioned on Z:

$$P(X|Z, \Theta) = \prod_{i=1}^{N} P(x_i|\theta_{z_i}). \tag{2.2}$$

2.1 Prior Distribution on Cluster Assignments. We incorporate our clustering preferences by manipulating the prior probability $P(Z|\Theta)$. In the standard gaussian mixture model, the prior distribution on cluster assignments Z is trivial:

$$P(Z|\Theta) = \prod_{i=1}^{N} \pi_{z_i}. \tag{2.3}$$

We incorporate our clustering preferences through a weighting function $g(Z)$ that has large values when the assignment of data points to clusters Z conforms to our preferences, and low values when Z conflicts with our preferences. Hence, we write

$$P(Z|\Theta, G) \equiv \frac{(\prod_i \pi_{z_i})g(Z)}{\sum_Z(\prod_j \pi_{z_j})g(Z)} = \frac{1}{\Omega}\left(\prod_i \pi_{z_i}\right)g(Z), \tag{2.4}$$

where $\Omega = \sum_Z(\prod_j \pi_{z_j})g(Z)$ is the normalization constant. The likelihood of the data, given a specific cluster assignment Z, is independent of the cluster assignment preferences, so the complete data likelihood is

$$P(X, Z|\Theta, G) = P(X|Z, \Theta)P(Z|\Theta, G). \tag{2.5}$$

From equations 2.1 through 2.5, the complete data likelihood is

$$P(X, Z|\Theta, G) = P(X|Z, \Theta)\frac{1}{\Omega}\prod_i \pi_{z_i} g(Z) = \frac{1}{\Omega}P(X, Z|\Theta)g(Z), \tag{2.6}$$

where $P(X, Z|\Theta)$ is the complete data likelihood for a standard GMM.

To distinguish the penalized likelihood from the standard likelihood, we introduce the notation $P_s(\cdot)$ to denote the standard likelihood and $P_p(\cdot)$ for penalized likelihood. The data likelihood is the sum of complete data likelihood over all possible Z, that is, $L(X|\Theta) = P_p(X|\Theta, G) = \sum_Z P_p(X, Z|\Theta, G)$, which can be maximized with the EM algorithm. Once the model parameters are fit, we do soft clustering according to the posterior probabilities for new data $P_s(k|x, \Theta)$. (Note that cluster assignment preferences are not expressed for the new data, only for the training data.)

2.2 Pairwise Relations. Pairwise relations provide a special case of the framework discussed above. We specify two types of pairwise relations:

- Link: two samples should be assigned to the same cluster.
- Do-not-link: Two samples should be assigned to different clusters.

The weighting factor given to the cluster assignment configuration Z is

$$g(Z) = \prod_{i \neq j} \exp\left(W_{ij}^{p}\ \delta(z_i, z_j)\right), \tag{2.7}$$

where δ is the Kronecker delta function and W_{ij}^{p} is the weight associated with sample pair (x_i, x_j). This weight satisfies

$$W_{ij}^{p} \in (-\infty, \infty),\ \ W_{ij}^{p} = W_{ji}^{p}.$$

The weight W_{ij}^{p} reflects our preference for assigning x_i and x_j into one cluster.

We use positive W_{ij}^{p} when we prefer to assign x_i and x_j into one cluster (link) and negative W_{ij}^{p} when we prefer to assign them into different clusters (do-not-link). The absolute value $|W_{ij}^{p}|$ reflects the strength of the preference. The prior probability with the pairwise relations is

$$P(Z|\Theta, G) = \frac{1}{\Omega} \prod_{i} \pi_{z_i} \prod_{i \neq j} \exp\left(W_{ij}^{p}\ \delta(z_i, z_j)\right). \tag{2.8}$$

It appears that the $g(Z)$ in equation 2.7 changes asymmetrically with the violation of link and do-not-link: when a link is conformed, $g(Z)$ increases; when a do-not-link is violated, $g(Z)$ decreases. Nevertheless, the *prior probability* given by equation 2.8 decreases under both types of violations.

The PPC model is clearly connected to the standard GMM and the constrained clustering model proposed by Shental et al. (2003). We shall show that both models can be viewed as special cases of PPC with particular W^{p}. The connection between PPC and other semisupervised clustering models is less straightforward and will be discussed in section 4. If $W_{ij}^{p} = 0$, we have no prior knowledge on the assignment relevancy of x_i and x_j. When $W_{ij}^{p} = 0$ for all pairs (i, j), we have $g(Z) = 1$; hence, the complete likelihood reduces to the standard one:

$$P_p(X, Z|\Theta, G) = \frac{1}{\Omega} P_s(X, Z|\Theta) g(Z) = P_s(X, Z|\Theta). \tag{2.9}$$

In the other extreme with $|W_{ij}^{p}| \to \infty$, assignments Z that violate the pairwise relations between x_i and x_j have zero prior probability, since for those assignments,

$$P_p(Z|\Theta, G) = \frac{\prod_k \pi_{z_k} \prod_{i \neq j} \exp(W_{ij}^{p}\ \delta(z_i, z_j))}{\sum_Z \prod_l \pi_{z_l} \prod_{m \neq n} \exp(W_{mn}^{p}\ \delta(z_m, z_n))} \to 0.$$

Then the relations become *hard constraints*, while the relations with $|W_{ij}^p| < \infty$ are called soft preferences. When all the specified pairwise relations are hard constraints, the data likelihood becomes

$$P_p(X, Z|\Theta, G) = \frac{1}{\Omega} \prod_{ij \in \mathcal{L}} \delta(z_i, z_j) \prod_{ij \in \mathcal{N}} (1 - \delta(z_i, z_j)) \prod_{i=1}^{N} \pi_{z_i} P_s(x_i \,|\, \theta_{z_i}), \tag{2.10}$$

where $\mathcal{L}$ is the set of linked sample pairs and $\mathcal{N}$ is the set of do-not-link sample pairs. It is straightforward to verify that equation 2.10 is essentially the same as the complete data likelihood given by Shental et al. (2003). Therefore, the model proposed by Shental et al. (2003) is equivalent to PPC with hard constraints. In appendix A, we give a detailed derivation of equation 2.10 and the equivalence of two models. When only hard constraints are available, we simply implement PPC based on equation 2.10. In the remainder of this letter, we use W^p to denote the prior knowledge on pairwise relations, that is,

$$P_p(X, Z|\Theta, G) \equiv P_p(X, Z|\Theta, W^p) = \frac{1}{\Omega} P_s(X, Z|\Theta) \prod_{i \neq j} \exp\left(W_{ij}^p \; \delta(z_i, z_j)\right) \tag{2.11}$$

2.3 Model Fitting. We use the EM algorithm (Dempster, Laird, & Rubin, 1977) to fit the model parameters Θ:

$$\Theta^* = \arg\max_{\Theta} L(X|\Theta, W^p).$$

The expectation step (E-step) and maximization step (M-step) are

$$\texttt{E-step:}\; Q(\Theta, \Theta^{(t-1)}) = E_{Z|X}(\log P_p(X, Z|\Theta, W^p)|X, \Theta^{(t-1)}, W^p)$$

$$\texttt{M-step:}\; \Theta^{(t)} = \arg\max_{\Theta} Q(\Theta, \Theta^{(t-1)}).$$

In the M-step, the optimal mean and covariance matrix of each component is

$$\mu_k = \frac{\sum_{j=1}^{N} x_j P_p(k|x_j, \Theta^{(t-1)}, W^p)}{\sum_{j=1}^{N} P_p(k|x_j, \Theta^{(t-1)}, W^p)}$$

$$\Sigma_k = \frac{\sum_{j=1}^{N} P_p(k|x_j, \Theta^{(t-1)}, W^p)(x_j - \mu_k)(x_j - \mu_k)^T}{\sum_{j=1}^{N} P_p(k|x_j, \Theta^{(t-1)}, W^p)} \quad .$$

The update of the prior probability of each component is more difficult due to the normalizing constant Ω in the data likelihood,

$$\Omega = \sum_{Z} \left\{ \prod_{k=1}^{N} \pi_{z_k} \prod_{i \neq j} \exp\left(W_{ij}^{p}\ \delta(z_i, z_j)\right) \right\}. \tag{2.12}$$

We need to find

$$\pi \equiv \{\pi_1, \ldots, \pi_m\} = \arg\max_{\pi} \sum_{l=1}^{M} \sum_{i=1}^{N} \log \pi_l P_p(l|x_i, \Theta^{(t-1)}, W^p) - \log \Omega(\pi), \tag{2.13}$$

which, unfortunately, does not have a closed-form solution in general.[1] In this letter, we use a rather crude approximation of the optimal π instead. First, we estimate the values of $\log \Omega(\pi)$ on a grid $H = \{\hat{\pi}^n\}$ on the simplex defined by

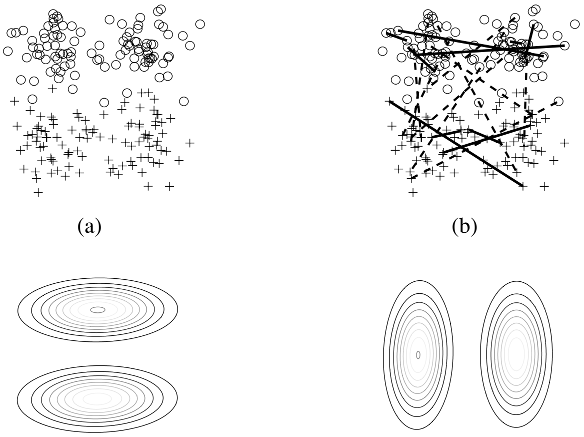
$$\sum_{k=1}^{M} \pi_k = 1,\ \ \pi_k \geq 0.$$

Then in each M-step, we calculate the value of $\sum_{l=1}^{M} \sum_{i=1}^{N} \log \hat{\pi}_l^n P_p(l|x_i, \Theta^{(t-1)}, W^p)$ for each node $\hat{\pi}^n \in H$ and find the node $\hat{\pi}^*$ that maximizes the function defined in equation 2.13:

$$\hat{\pi}^* = \arg\max_{\hat{\pi}^n \in H} \sum_{l=1}^{M} \sum_{i=1}^{N} \log \hat{\pi}_l^n P_p(l|x_i, \Theta^{(t-1)}, W^p) - \log \Omega(\hat{\pi}^n). \tag{2.14}$$

We use $\hat{\pi}^*$ as the approximative solution of equation 2.13. In this letter, the resolution of the grid is set to be 0.01. Although it works very well for all experiments in this letter, we notice that the search over grid will be fairly slow for $M > 5$. Shental, Bar-Hillel, Hertz, and Weinshall (2004) proposed to find optimal π using gradient descent and approximate $\Omega(\pi)$ by pretending all specified relations are nonoverlapping (see section 3.1). Although this method is originally designed for hard constraints, it can be easily adapted for PPC. This will not be covered in this letter.

[1] Shental et al. (2003) pointed out that with a different sampling assumption, a closed-form solution for equation 2.13 exists when only hard links are available. See section 4.


(a) (b)

(c) (d)

Figure 1: The influence of constraint weight on model fitting. (a) Artificial data set. (b) Links (solid lines) and do-not-links (dashed line). (c, d) The probability density contour of two possible fitted models.

It is critically important to note that with a nontrivial W^p, the assignment independence is broken,

$$P_p(z_i, z_j|x_i, x_j, \Theta, W^p \neq P_p(z_i|x_i, \Theta, W^p)P_p(z_j|x_j, \Theta, W^p),$$

which means that the posterior estimation of each sample cannot be done separately. This fact brings an extra computational problem and will be discussed in section 3.

2.4 Selecting the Constraint Weights

2.4.1 Example: How the Weight W_{ij} Affects Clustering. The weight matrix W^p is crucial to the performance of the PPC. Here we give an example demonstrating how the weight of pairwise relations affects the clustering process. Figure 1a shows the two-dimensional data from *two* classes, as indicated by the symbols. Besides the data set, we also have 20 pairs correctly labeled as links and do-not-links, as shown in Figure 1b. We try to fit the data set with a two-component GMM. Figures 1c and 1d give the density contour of the two possible models on the data. Without any pairwise relations specified, we have essentially an equal chance to get each. After incorporating pairwise relations, the EM optimization process is biased to

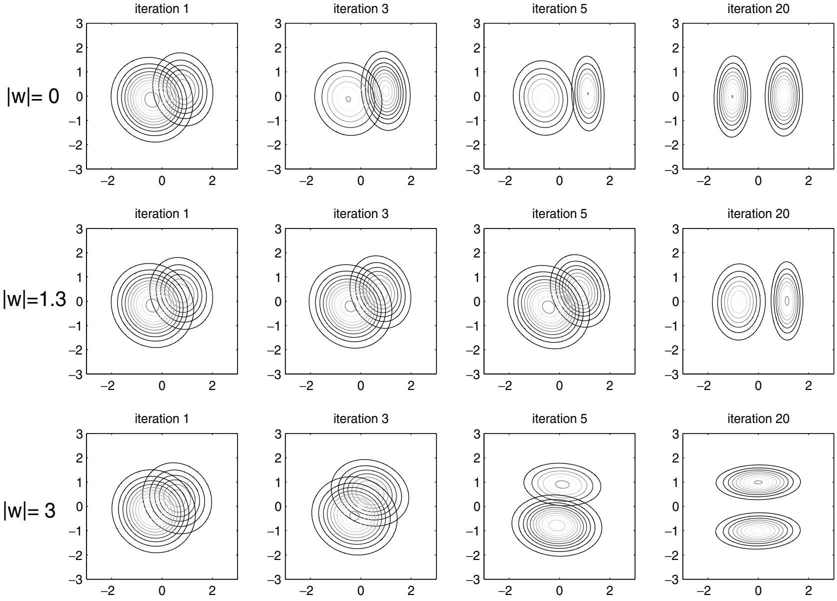


Figure 2: The contour of probability density fit on data with different weight given to pairwise relations. Top row: $w = 0$; middle row: $w = 1.3$; bottom row: $w = 3$.

the correct one. The weights of pairwise relations are given as follows,

$$W_{ij}^{p} = \begin{cases} w & \text{if } (x_i, x_j) \text{ is linked} \\ -w & \text{if } (x_i, x_j) \text{ is do-not-linked} \\ 0 & \text{otherwise,} \end{cases}$$

where $w \geq 0$ measures the certainty of all specified pairwise relations. In Figure 2, we give three runs with the same initial model parameters but different weight given to the specified pairwise relations.

For each run, we give snapshots of the model after 1, 3, 5, and 20 EM iterations. The first row is the run with $w = 0$ (standard GMM). The search ends up with a model that violates our prior knowledge of class membership. The middle row is the run with w set to 1.3. With the same poor initial condition, the model fitting process still goes to the wrong one again, although at a slower pace. In the bottom row, we increase w to 3. This time, the model converges to the one we intend.

2.4.2 Choosing Weight W^p Based on Prior Knowledge. There are some occasions we can translate our prior belief on the relations into the weight

W^p. Here we assume that the pairwise relations are labeled by an oracle but contaminated by flipping noise before they are delivered to us. For each labeled pair (x_i, x_j), there is thus a certainty value $0.5 \leq \gamma_{ij} \leq 1$ equal to the probability that pairwise relation is not flipped, that is, that label is correct.[2] Our prior knowledge would include those specified pairwise relations and their certainty values $\Gamma = \{\gamma_{ij}\}$.

This prior knowledge can be approximately encoded into the weight W^p by letting

$$W^p_{ij} = \begin{cases} \frac{1}{2}\log(\frac{\gamma_{ij}}{1-\gamma_{ij}}) & (x_i, x_j) \text{ is specified as linked} \\ -\frac{1}{2}\log(\frac{\gamma_{ij}}{1-\gamma_{ij}}) & (x_i, x_j) \text{ is specified as do-not-linked} \\ 0 & \text{otherwise.} \end{cases} \tag{2.15}$$

The details of the derivation are in appendix B. It is obvious from equation 2.15 that for a specified pairwise relation (x_i, x_j), the greater the certainty value γ_{ij}, the greater the absolute value of weight W^p_{ij}.

Note that the weight designed this way is not necessarily optimal in terms of classification accuracy, as will be demonstrated by experiment in section 5.1. The reason is twofold. First, equation 2.15 is derived based on a (possibly crude) approximation. Second, gaussian mixture models as classifiers are often considerably biased from true class distribution of data. As a result, even if the PPC prior $P(Z|\Theta, W^p)$ faithfully reflects the truth, it does not necessarily lead to the best classification accuracy. Nevertheless, equation 2.15 gives good initial guidance for choosing the weight. Our experiments in section 5.1 show that this design often yields superior classification accuracy to simply using the hard constraints or ignoring the pairwise relations (standard GMM).

One use for this scheme of weight is when pairwise relations are labeled by domain experts and the certainty values are given at the same time. We might also estimate the flipping noise parameters from historical data or available statistics. For example, we can derive soft pairwise relations based on spatial or temporal continuity among samples. That is, we add soft links to all adjacent pairs of samples, assuming the flipping noise explaining all the adjacent pairs that are actually not in one class. We further assume that the flipping noise of each pair follows the same distribution. Accordingly, we assign uniform weight $w > 0$ to all adjacent pairs. Let q denote the probability that the label on a adjacent pair is flipped. We might be able to estimate q from labeled instances of a similar problem, for example, segmented images or time series. The maximum likelihood (ML) estimation

[2] We consider only the certainty value > 0.5, because a pairwise relation with certainty $\gamma_{ij} < 0.5$ can be equivalently treated as its opposite relation with certainty $1 - \gamma_{ij}$.

of q is given by simple statistics:

$$\tilde{q} = \frac{\text{Number of adjacent pairs that are not in the same class}}{\text{Number of all adjacent pairs}}.$$

We give an application of this idea in section 5.2.

3 Computing the Cluster Posterior

The M-step requires the cluster membership posterior. Computing this posterior is simple for the standard GMM since each data point x_i can be assigned to a cluster independent of the other data points and we have the familiar cluster origin posterior $P_s(z_i = k|x_i, \Theta)$. The pairwise constraints bring extra relevancy in assignment among samples involved. From equation 2.11, if $W_{ij}^p \neq 0$, $P_p(z_i, z_j|x_i, x_j, \Theta, W^p) \neq P_p(z_i|x_i, \Theta, W^p)P_p(z_j|x_j, \Theta, W^p)$. Consequently, the posterior probability of x_i and x_j cannot be estimated separately. This relevancy in the assignment can be formalized as follows:

Definition. *If $W_{ij}^p \neq 0$, we say there is direct assignment relevancy between x_i and x_j, denoted by $x_i R_d x_j$. If $P_p(z_i, z_j|x_i, x_j, \Theta, W^p) \neq P_p(z_i|x_i, \Theta, W^p)P_p(z_j|x_j, \Theta, W^p)$, we say there is assignment relevancy between x_i and x_j , denoted by $x_i R_a x_j$.*

It is clear that R_a is reflexive, symmetric, and transitive. Hence, R_a is an equivalence relation. It can be shown that R_a is the transitive closure of R_d. In other words, two samples have assignment relevancy relation R_a if they can be connected by a path consisting of R_a relations, as illustrated in Figure 3. We call each equivalence class associated with R_a a clique. It is clear that cliques are the smallest sets of samples whose posterior probabilities can be calculated independently. When calculating posterior probabilities, all samples within a clique need to be considered together. In a clique T with size $|T|$, the posterior probability of a given sample $x_i \in T$ is calculated by marginalizing the posterior over the entire clique,

$$P_p(z_i = k|X, \Theta, W^p) = \sum_{Z_T|z_i=k} P_p(Z_T|X_T, \Theta, W^p),$$

with the posterior on the clique given by

$$P_p(Z_T|X_T, \Theta, W^p) = \frac{P_p(Z_T, X_T|\Theta, W^p)}{P_p(X_T|\Theta, W^p)} = \frac{P_p(Z_T, X_T|\Theta, W^p)}{\sum_{Z_T'} P_p(Z_T', X_T|\Theta, W^p)}.$$

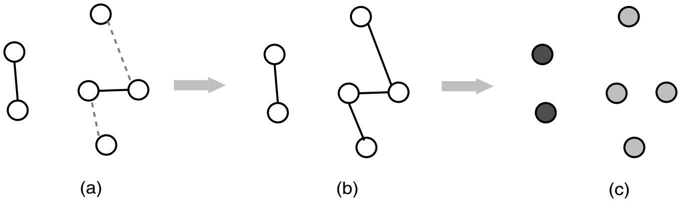


Figure 3: (a) Links (solid line) and do-not-links (dotted line) among six samples. (b) Direct assignment relevancy R_d (solid line) translated from links in a. (c) Equivalence classes defined by assignment relevancy R_a, denoted by shading.

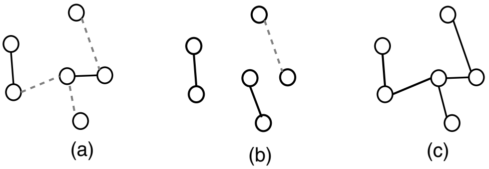


Figure 4: (a) Overlapping pairwise relations, with links (solid line) and do-not-links (dotted line). (b) Nonoverlapping pairwise relations. (c) Only hard links.

Exact calculation of the posterior probability of a sample in clique T requires time complexity $O(M^{|T|})$, where M is the number of components in the mixture model. This calculation can get prohibitively expensive if $|T|$ is large (e.g., 50) for any model size $M \geq 2$. Hence, small-size cliques are required to make the marginalization computationally reasonable.

3.1 Two Special Cases with Easy Inference. Apparently the inference is easy when we limit ourselves to small cliques. Specifically, when $|T| \leq 2$, the pairwise relations are nonoverlapping, as illustrated in Figures 4a and 4b. With nonoverlapping constraints, the posterior probability for the whole data set can be given in closed form with $O(N)$ time complexity. Moreover, the evaluation of the normalization factor $\Omega(\pi)$ is simple:

$$\Omega(\pi) = \left(1 - \sum_{k=1}^{M} \pi_k^2\right)^{N_L} \left(\sum_{k=1}^{M} \pi_k^2\right)^{N_N},$$

where N_L and N_N are, respectively, the number of links and do-not-links. The optimization of π in the M-step can thus be achieved with little cost. Sometimes nonoverlapping relations are a natural choice: they can be generated by picking up sample pairs from sample sets and labeling the relations without replacement. More generally, we can avoid the expensive computation in posterior inference by breaking large cliques into small ones. To do this, we need to deliberately ignore some links or do-not-links. In section 5.2, experiment 3 is an application of this idea.

The second simplifying situation is that we have only hard links ($W_{ij}^{p} = +\infty$ or 0), as illustrated in Figure 4c. In this case, the posterior probability for each sample must be exactly the same as the others in the same clique, so a clique can be treated as a single sample. That is, assume x_i is in clique T. We then have

$$\begin{aligned} P_p(z_i = k|x_i, \Theta, W^p) &= P_p(Z_T = k|x_T, \Theta, W^p) \\ &= \frac{P_p(x_T, Z_T = k|\Theta, W^p)}{\sum_k' P_p(x_T, Z_T = k'|\Theta, W^p)} \\ &= \frac{P_s(x_T, Z_T = k|\Theta)}{\sum_k' P_s(x_T, Z_T = k'|\Theta)} \\ &= \frac{\prod_{j\in T} \pi_k P_s(x_j|\theta_k)}{\sum_{k'}(\prod_{j\in T} \pi_{k'} P_s(x_j|\theta_{k'}))}. \end{aligned}$$

Similar ideas have been proposed independently by Wagstaff et al. (2001), Shental et al. (2003), and Bilenko et al. (2004). This case is useful when we are sure that a group of samples are from one source (Shental et al., 2003).

For more general cases, where the exact inference is computationally prohibitive, we propose to use Gibbs sampling (Neal, 1993) and the mean field approximation (Jaakkola, 2004) to estimate the posterior probability. This will be discussed in sections 3.2 and 3.3.

3.2 Estimation with Gibbs Sampling. For fixed Θ, finding $P_p(Z|\Theta, W^p)$ is a typical inference problem for graphical models. Techniques for approximate inference developed for graphical models can also be used here. In this section, we use Gibbs sampling to estimate the posterior probability in each EM iteration.

In Gibbs sampling, we estimate $P_p(z_i|X, \Theta, W^p)$ as a sample mean,

$$P_p(z_i = k|X, \Theta, W^p) = E(\delta(z_i, k)|X, \Theta, W^p) \approx \frac{1}{S}\sum_{t=1}^{S} \delta\big(z_i^{(t)}, k\big),$$

where the sum is over a sequence of S samples from $P(Z|X, \Theta, G)$ generated by the Gibbs MCMC. The tth sample in the sequence is generated by the usual Gibbs sampling technique:

- Pick $z_1^{(t)}$ from distribution $P_p(z_1|z_2^{(t-1)}, z_3^{(t-1)}, \ldots, z_N^{(t-1)}, X, W^p, \Theta)$.

- Pick $z_2^{(t)}$ from distribution $P_p(z_2|z_1^{(t)}, z_3^{(t-1)}, \ldots, z_N^{(t-1)}, X, W^p, \Theta)$.

  ...
- Pick $z_N^{(t)}$ from distribution $P_p(z_N|z_1^{(t)}, z_2^{(t)}, \ldots, z_{N-1}^{(t)}, X, W^p, \Theta)$.

For pairwise relations it is helpful to introduce some notation. Let Z_{-i} denote an assignment of data points to clusters that leaves out the assignment of x_i. Let $U(i)$ be the indices of the set of samples that participate in a pairwise relation with sample x_i, $U(i) = \{j : W_{ij}^p \neq 0\}$. Then we have

$$P_p(z_i|Z_{-i}, X, \Theta, W^p) \propto P_s(x_i, z_i|\Theta) \prod_{j \in U(i)} \exp(2W_{ij}^p \, \delta(z_i, z_j)). \tag{3.1}$$

The time complexity of each Gibbs sampling pass is $O(NnM)$, where n is the maximum number of pairwise relations a sample can be involved in. When W^p is sparse, the size of $U(i)$ is small. Thus, calculating $P_p(z_i|Z_{-i}, X, \Theta, W^p)$ is fairly cheap, and Gibbs sampling can effectively estimate the posterior probability.

3.3 Estimation with Mean Field Approximation. Another approach to posterior estimation is to use mean field theory (Jaakkola, 2004; Lange, Law, Jain, & Buhmann, 2005). Instead of directly evaluating the intractable $P_p(Z|X, \Theta, W)$, we try to find a tractable mean field approximation $Q(Z)$. To find a $Q(Z)$ close to the true posterior probability $P_p(Z|X, \Theta, W)$, we minimize the Kullback-Leibler divergence between them,

$$\min_Q \mathrm{KL}(Q(Z)|P_p(Z|X, \Theta, W^p)), \tag{3.2}$$

which can be recast into

$$\max_Q [H(Q) + E_Q\{\log P_p(Z|X, \Theta, W^p)\}], \tag{3.3}$$

where $E_Q\{\cdot\}$ denotes the expectation with respect to Q. The simplest family of variational distribution is one where all the latent variables $\{z_i\}$ are independent of each other:

$$Q(Z) = \prod_{i=1}^{N} Q_i(z_i). \tag{3.4}$$

With this $Q(Z)$, the optimization problem in equation 3.3 does not have a closed-form solution and is not a convex problem. Instead, a locally optimal

Q can be found iteratively with the following update equations,

$$Q_i(z_i) \leftarrow \frac{1}{\Omega_i} \exp(E_Q\{\log P_p(Z|X, \Theta, W^p)|z_i\}), \tag{3.5}$$

for all i and $z_i \in \{1, 2, \cdots, M\}$. Here $\Omega_i = \sum_{z_i} \exp(E_Q\{\log P_p(Z|X, \Theta, W^p)|z_i\})$ is the local normalization constant. For the PPC model, we have

$$\exp(E_Q\{\log P_p(Z|X, \Theta, W^p)|z_i\}) = P_s(z_i|x_i, \Theta) \exp\left(\sum_{j\neq i} W^p_{ij} Q_j(z_i)\right).$$

Equation 3.5, collectively for all i, is the mean field equation. Evaluation of mean field equations requires at most $O(NnM)$ time complexity, which is same as the time complexity of one Gibbs sampling pass. Successive updates of equation 3.5 will converge to a local optimum of equation 3.3. In our experiments, the convergence usually occurs after about 20 iterations, which is many fewer than the number of passes required for Gibbs sampling.

4 Related Models

Prior to our work, different authors have proposed several constrained clustering models based on K-means, including the seminal work by Wagstaff and colleagues (Wagstaff et al., 2001; Wagstaff, 2002), and its successor (Basu et al., 2002; Basu, Bilenko, & Mooney, 2004; Bilenko et al., 2004). These models generally fall into two classes. The first class of algorithms (Wagstaff et al., 2001; Basu et al., 2002) keeps the original K-means cost function (reconstruction error) while confining the cluster assignments to be consistent with the specified pairwise relations. The problem can be cast into the following constrained optimization problem,

$$\begin{aligned} &\min_{Z,\mu} \sum_{i=1}^{N} ||x_i - \mu_{z_i}||^2 \\ &\text{subject to } z_i = z_j, \quad \text{if } (x_i, x_j) \in \mathcal{L} \\ &\qquad\qquad\; z_i \neq z_j, \quad \text{if } (x_i, x_j) \in \mathcal{N}, \end{aligned}$$

where $\mu = \{\mu_1, \ldots, \mu_M\}$ is the cluster center. In the second class of algorithms, cluster assignments that violate the pairwise relations are allowed but will be penalized. They employ a modified cost function (Basu et al., 2004):

$$J(\mu, Z) = \frac{1}{2}\sum_{i=1}^{N} ||x_i - \mu_{z_i}||^2 + \sum_{(i,j)\in\mathcal{L}} a_{ij}(z_i \neq z_j) + \sum_{(i,j)\in\mathcal{N}} b_{ij}(z_i = z_j), \tag{4.1}$$

where a_{ij} is the penalty for violating the link between (x_i, x_j) and b_{ij} is the penalty when the violated pairwise relation is a do-not-link. It can be shown that both classes of algorithms are encompassed by PPC as special cases. The particular PPC model we consider has spherical gaussian components with radius shrunk to zero and the weight matrix W^p expands properly. The details are in appendix C.

One weakness shared by the semisupervised K-means algorithms is the limited capability of K-means to model complex data distribution. If data in one class are far from being spherical, it may take a great number of pairwise relations to achieve reasonable classification accuracy (Wagstaff, 2002). Another serious problem lies in the optimization strategy employed by those algorithms to find the optimal assignment within each EM iteration. Due to the extra dependency brought by the pairwise relations, finding the optimal assignment of samples to clusters is not trivial. Evaluating every potential assignments requires $O(M^{|T|})$ time complexity where $|T|$ denotes the size of the biggest clique, which is prohibitively expensive when $|T|$ is big. The greedy search used by these algorithms can return only local optima (Basu et al., 2002, 2004), and the sequential assignment strategy employed by Wagstaff et al. (2001) may lead to the situation where one cannot assign a sample to any cluster because of the conflict with some assigned samples.

To remedy the limited capability of constrained K-means, several authors proposed probabilistic models based on gaussian mixture models. The models proposed by Shental et al. (2003, 2004) address the situation where pairwise relations are hard constraints. The authors partition the whole data set into a number of (maximal) "chunklets" consisting of samples that are (hard) linked to each other.[3] Shental et al. (2003, 2004) discuss two sampling assumptions:

- Assumption 1: Chunklet X_l is generated identically and independently from component k with prior π_k (Shental et al., 2004), and the complete data likelihood is

$$P(X, Z|\Theta, E_\Omega) = \frac{1}{\Omega}\prod_{i\neq j\in\mathcal{N}} (1 - \delta(z_i, z_j)) \cdot \prod_{l=1}^{L}\{\pi_{z_l} \prod_{x_i\in X_l} P_s(x_i|\theta_{z_l})\}, \tag{4.2}$$

[3] If a sample is not linked to any other samples, it comprises a chunklet by itself.

where E_Ω denotes the specified constraints.

- Assumption 2: Chunklet X_i generated from component k with prior $\propto \pi_k^{|X_i|}$, where $|X_i|$ is the number of samples in X_i (Shental et al., 2004). The complete data likelihood is:

$$P(X, Z|\Theta, E_\Omega) = \frac{1}{\Omega} \prod_{ij\in\mathcal{N}} (1 - \delta(z_i, z_j)) \cdot \prod_{l=1}^{L} \{\pi_{z_l}^{|X_l|} \prod_{x_i \in X_l} P_s(x_i|\theta_{z_l})\} \tag{4.3}$$

$$= \frac{1}{\Omega} \prod_{ij\in\mathcal{L}} \delta(z_i, z_j) \prod_{ij\in\mathcal{N}} (1 - \delta(z_i, z_j)) \prod_{i=1}^{N} \pi_{z_i} P_s(x_i|\,\theta_{z_i}). \tag{4.4}$$

In appendix A we show that when using assumption 2, this model (as expressed in equations 4.3 and 4.4) is equivalent to the PPC with only hard constraints (as expressed in equation 2.10). Shental et al. (2004) suggested that assumption 1 might be appropriate, for example, when chunklets are generated from temporal continuity. When pairwise relations are generated by labeling sample pairs picked from a data set, assumption 2 might be more reasonable. Assumption 1 allows a closed-form solution in the M-step (including solution for π) in each EM iteration (Shental et al., 2004). The empirical comparison of the two sampling assumptions will be discussed in section 5.

To incorporate the uncertainty associated with pairwise relations, Law et al. (2004, 2005) proposed to use soft group constraints. To model a link between any sample pair (x_i, x_j), they create a group l and express the strength of the link as the membership of x_i and x_j to group l. This strategy works well for some simple situations, for example, when the pairwise relations are nonoverlapping (as defined in section 3.1). However, it is awkward if samples are shared by multiple groups, which is unavoidable when samples are commonly involved in multiple relations. Another serious drawback of the group constraints model is its inability to model do-not-links. Due to these obvious limitations, we omit the empirical comparison of this model to PPC in the following section.

5 Experiments

The experiments section consists of two parts. In section 5.1, we examine the way the number of constraints affects the clustering results. For each clustering task in this section, we generate artificial pairwise relations based on class labels. In section 5.2, we address real-world problems, where the constraints are derived from our prior knowledge. Also in this section, we demonstrate the approaches to reduce computational complexity, as described in section 3.

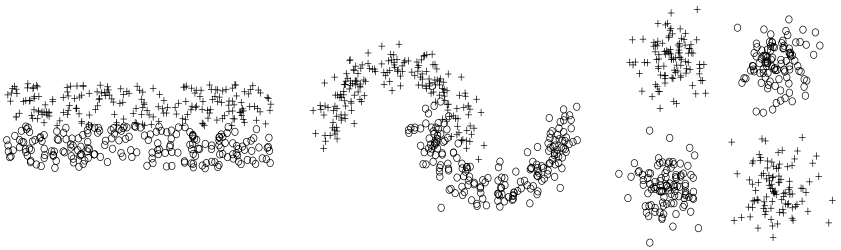
Following are some abbreviations we will use: *soft-PPC* is PPC with soft constraints, *hard-PPC* is PPC with hard constraints (implemented based on equation 2.10), *soft-CKmeans* is the K-means with soft constraints (Basu et al., 2004) and *hard-CKmeans* is the K-means with hard constraints (Wagstaff et al., 2001). The gaussian mixture model with hard constraints (Shental et al., 2003, 2004) will be referred to as constrained-EM.

5.1 Artificial Constraints. In this section, we discuss the influence of pairwise relations on PPC's clustering and compare the result to other semisupervised clustering models. This section presents two experiments. In experiment 1, we consider only correct pairwise relations, as an example of authentic knowledge. Accordingly, we use hard constraints in clustering. In experiment 2, we discuss the situation where pairwise relations contain significant error. We evaluate the performance of soft-PPC and test the weight design strategy described in section 2.4. The result is compared to hard-PPC and other semisupervised clustering models.

5.1.1 Constraint Selection. To avoid the computational burden, we limit our discussion to the nonoverlapping pairwise relations in experiments 1 and 2. As discussed in section 3.1, the nonoverlapping pairwise relations, hard or soft, allow fast solution in the maximization step in each EM iteration. The pairwise relations are generated as follows. We randomly pick two samples from the training set without replacement. If the two have the same class label, we add a link constraint between them; otherwise, we add a do-not-link constraint. Note that the application of PPC is not limited to the nonoverlapping cases. In section 5.2, we discuss more complicated real-world problems where overlapping constraints are necessary, and we also present approaches to solve the computational problems.

5.1.2 Performance Evaluation. We try PPC (with the number of components equal to the number of classes) with various numbers of pairwise relations. For each clustering result, a confusion matrix is built to compare it to true labeling. The classification accuracy is calculated as the ratio of the sum of diagonal elements to the number of all samples.

5.1.3 Experiment 1: Artificial Hard Constraints. This experiment is designed to answer two questions: how the number of pairwise relations affects the clustering result and whether the information in the relations has been successfully encoded into the trained model. To answer the second question, we examine the out-of-sample classification of the gaussian mixture model fit with the aid of the pairwise relations. Toward this end, we divide each data set into a training set (90% of data) and a held-out test set (10% of data). Pairwise relations are generated among samples in



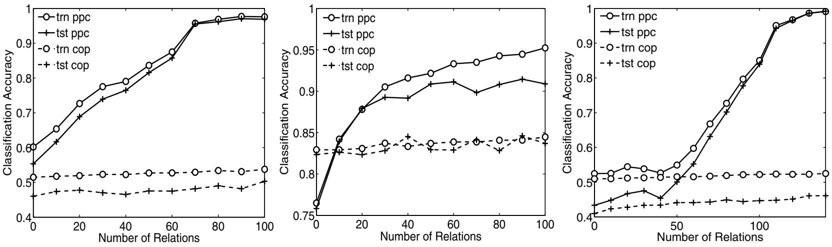
(a) data set1 (b) data set 2 (c) data set 3

Figure 5: The artificial data sets.

the training set. After the density model is fit on the training set and the pairwise relations, it will be applied to the test set. Since the classification on the test set is merely decided by the fit gaussian mixture model, it will reflect the influence of pairwise relations on the trained model. For comparison, we also give results of two other constrained clustering methods: (1) the hard-CKmeans (Wagstaff et al., 2001), for which the accuracy on the test set is given by the nearest-neighbor classification with the cluster centers fit on training set, and (2) constrained-EM (Shental et al., 2004). Since we show in appendix A that constrained-EM with sampling assumption 2 (see section 5) is equivalent to hard-PPC, we need only to consider the constrained-EM with sampling assumption 1. The reported classification accuracy is averaged over 100 different realizations of pairwise relations.

The three two-dimensional artificial data sets shown in Figure 5 are designed to highlight PPC's superior modeling flexibility over constrained K-means.[4] In each example, there are 200 samples in each class. It is clear from Figure 5 that for all three problems, data in each class are nongaussian. So not surprisingly, standard K-means and GMM do not return satisfactory clustering results. Figure 6 compares the clustering result of hard-PPC and hard-CKmeans with various number of pairwise relations. As shown in Figure 6, the accuracy of hard-PPC improves significantly when pairwise relations are incorporated. After enough pairwise relations are added in, we can finally reach close to 100% accuracy on the training data. On the test set, although no pairwise relation is available, we observe significantly improved accuracy as well. For the hard-CKmeans, we do not observe any substantial accuracy improvement on the training set or test set. The classification task of data set 1 is relatively easy for the gaussian mixture and difficult for K-means. The classification accuracy of hard-PPC climbs

[4] Some authors (Xing et al., 2003; Cohn et al., 2003; Bilenko et al., 2004) combined standard or constrained K-means with metric learning based on pairwise relations and reported improvement on classification accuracy. This will not be discussed in this letter.



(a) On data set 1 (b) On data set 2 (c) On data set 3

Figure 6: The performance of hard-PPC and hard-CKmeans with various numbers of relations: trn ppc: accuracy on the training set with hard-PPC; tst ppc: accuracy on the test set from the GMM trained by hard-PPC; trn cop: accuracy on the training set with hard-CKmeans; tst cop: accuracy of K-means on the test set.

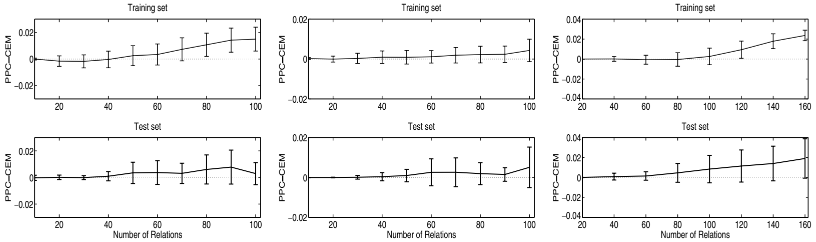
from 60% to close to 100% (on both training and test set) after 70 pairwise relations, whereas the accuracy of hard-CKmeans remains less than 60% even with 100 relations. On data set 2, the hard-PPC accuracy is improved from 75% to close to 95% on the training set and stops at around 90% on the test set. This divergence happens because the two classes in data set 2 are overlapped and thus defy a perfect GMM classifier. Data set 3 is the most difficult since it is highly nongaussian. It takes over 100 pairs for the hard-PPC to reach 95% accuracy, whereas hard-CKmeans never reaches 55%.

The comparison of PPC and constrained-EM is presented in a way to highlight the difference between the classification accuracy of the two methods. Basically, we record the classification from PPC and constrained-EM with the same pairwise relations and initial condition and then calculate

$$\begin{aligned}\Delta\text{Accuracy} = {} & \text{classification accuracy by PPC} \\ & - \text{classification accuracy by CEM}\end{aligned}$$

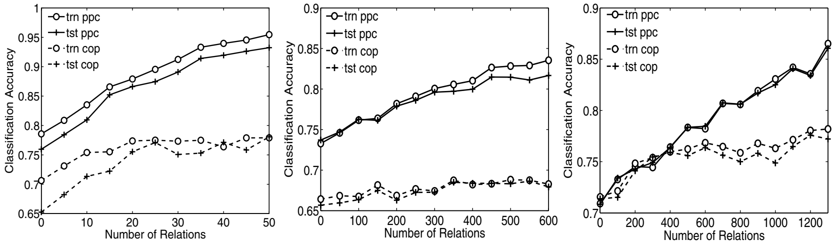
on both the training and test set. In Figure 7, we report the mean and standard deviation of ΔAccuracy estimated over 100 different realizations of pairwise relations. From Figure 7, the difference between PPC and constrained-EM is indistinguishable when the number of relations is small, while PPC is slightly but consistently better than constrained-EM when the relations are abundant.

We perform the same experiments on three UCI data sets: the Iris data set has 150 samples and three classes, with 50 samples in each class; the Waveform data set has 5000 samples and three classes, with around 1700 samples in each class; and the Pendigits data set has four classes (digits 0, 6, 8, 9), each with 750 samples. The results are summarized in



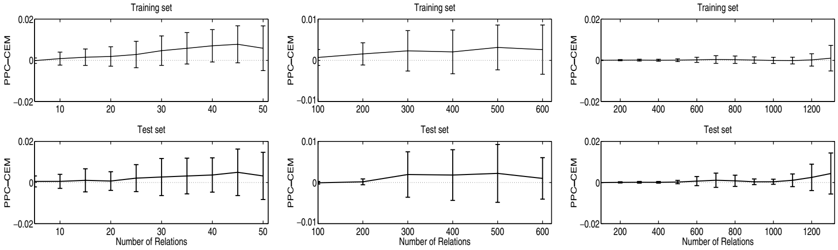
(a) data set 1 (b) data set 2 (c) data set 3

Figure 7: Comparison of PPC and constrained EM on artificial data. With each number of pairwise relations, we show the mean of ΔAccuracy ± standard deviation estimated over 100 random realization of pairwise relations.



(a) Iris data (b) Waveform data (c) Pendigits data

Figure 8: Performance of PPC on UCI data sets with various numbers of relations.



(a) Iris (b) Waveform (c) Pendigits

Figure 9: Comparison of PPC and constrained EM on UCI data sets. With each number of pairwise relations, we show the mean of ΔAccuracy ± standard deviation estimated over 100 random realization of pairwise relations.

Figures 8 and 9. As indicated by Figure 8, hard-PPC can consistently improve its clustering accuracy on the training set when more pairwise constraints are added; also, the effect brought by constraints generalizes to the test set. In contrast, as in the artificial data set case, the increase of accuracy from hard-CKmeans is much less salient than that of hard-PPC. Figure 9 shows that hard-PPC is slightly better than constrained-EM, especially when the number of constraints is large.

5.1.4 Experiment 2: Artificial Soft Constraints. In this experiment, we evaluate the performance of soft-PPC when the specified pairwise relations contain substantial error. The results are compared to hard-PPC, soft-CKmeans, and hard-CKmeans.

The artificial constraints are generated the same way as in the previous experiment. The flipping noise is realized by randomly flipping each pairwise relation with a certain probability $q \leq 0.5$. For the soft-PPC model, the weight W_{ij}^{p} to each specified pairwise relation is given as follows:

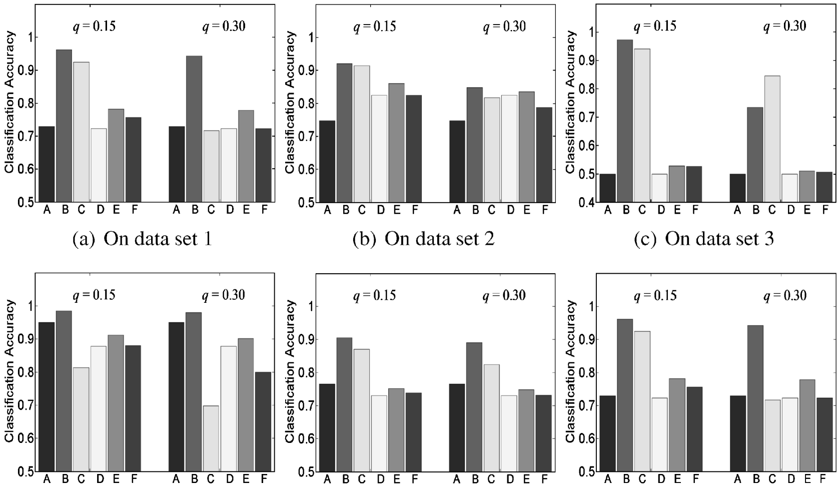
$$W_{ij}^{p} = \begin{cases} \frac{1}{2}\log\left(\frac{1-q}{q}\right) & (z_i, z_j)\text{specified as link} \\ -\frac{1}{2}\log\left(\frac{1-q}{q}\right) & (z_i, z_j)\text{specified as do-not-link.} \end{cases} \tag{5.1}$$

We use w to denote the absolute value of the weight for nontrivial pairs. For soft-PPC, we have $w = \frac{1}{2}\log(\frac{1-q}{q})$. For soft-CKmeans, we give equal weights to all specified constraints. Because there is no guiding rule in the literature on how to choose the weight for the soft-CKmeans model, we simply use the weight that yields the highest classification accuracy.

We present the results on the three artificial data sets and three UCI data sets used in experiment 1. Unlike experiment 1, we use everything available in clustering. On each data set, we randomly generate a number of nonoverlapping pairwise relations to have 50% of the data involved. In this experiment, we try two different noise levels, with q set to 0.15 and 0.3. Figure 10 compares the classification accuracies given by the maximum likelihood (ML) solutions of different models.[5] The accuracy for each model is averaged over 20 random realizations of pairwise relations. On all data sets except artificial data set 3, soft-PPC with the designed weight gives higher accuracy than hard-PPC ($w = \infty$) and standard GMM ($w = 0$) on both noise levels. On artificial data set 3, when $q = 0.3$, hard-PPC gives the best classification accuracy.[6] Soft-PPC apparently gives superior classification accuracy to the K-means models on all six data sets, even though the weight of soft-CKmeans is optimized. Figure 10 also shows that it can be harmful to use hard constraints when pairwise relations

[5] We choose the one with the highest data likelihood among 100 runs with different random initialization. For K-means models, including soft-CKmeans and hard-CKmeans, we use the solutions with the smallest value of cost function.

[6] Further experiment shows that on these data, soft-PPC with the optimal w (> the one suggested by equation 5.1) is still slightly better than hard-PPC.



(a) On data set 1

(b) On data set 2

(c) On data set 3

(d) On Iris data

(e) On Waveform data

(f) On Pendigits data

Figure 10: Classification accuracy with noisy pairwise relations. We use all the data in clustering. (A) Standard GMM. (B) Soft-PPC. (C) Hard-PPC. (D) Standard K-means. (E) Soft-CKmeans with optimal weight. (F) Hard-CKmeans.

are noisy, especially when the noise is significant. Indeed, as shown by Figures 10d and 10f, hard-PPC can yield accuracy even worse than standard GMM.

5.2 Real-World Problems. In this section, we present two examples where pairwise constraints are from domain experts or common sense. Both examples are about image segmentation based on gaussian mixture models. In the first problem, experiment 3, hard pairwise relations are derived from image labeling done by a domain expert. In the second problem, soft pairwise relations are generated based on spatial continuity.

5.2.1 Experiment 3: Hard Do-Not-Links from Partial Class Information. The experiment in this section shows the application of pairwise constraints on partial class information. For example, consider a problem with six classes $A, B, \ldots, F$. The classes are grouped into several class sets $C_1 = \{A, B, C\}, C_2 = \{D, E\}, C_3 = \{F\}$. The samples are partially labeled in the sense that we are told which class set a sample is from but not which specific class it is from. We can logically derive a do-not-link constraint between any pair of samples known to belong to different class sets, while no link constraint can be derived if each class set has more than one class in it.

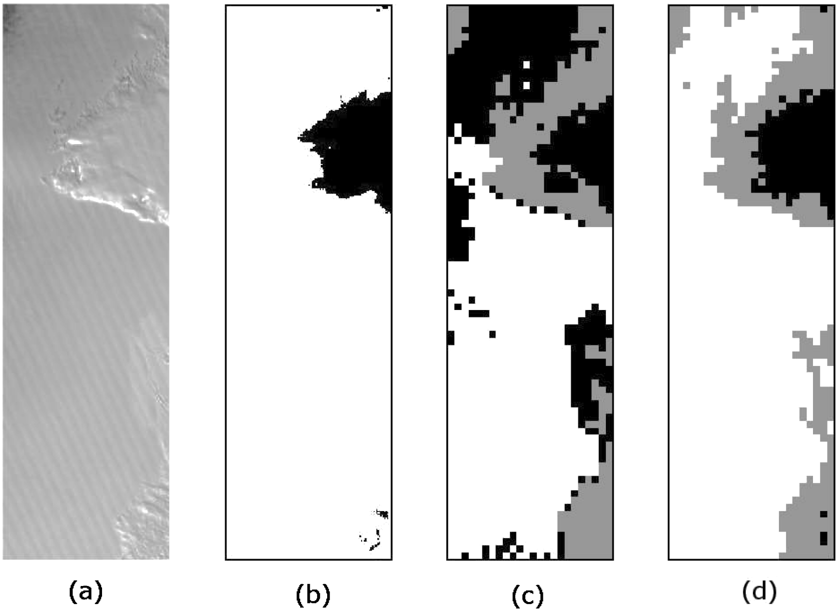


Figure 11: (a) Gray-scale image from the first spectral channel 1. (b) Partial label given by experts. Black pixels denote nonsnow area, and white pixels denote snow area. Clustering result of standard GMM (c) and PPC (d). Panels c and d are colored according to image blocks' assignment.

Figure 11a is a 120 × 400 region from the Greenland ice sheet from NASA Langley DAAC (Srivastava, Oza, & Stroeve, 2005).[7] Each pixel has intensities from seven spectrum bands. This region is labeled into snow area and nonsnow area, as indicated in Figure 11b. The snow area may contain samples from several classes of interest: ice, melting snow, and dry snow, while the nonsnow area can be bare land, water, or cloud. The labeling from the expert contains incomplete but useful information for further segmentation of the image. To segment the image, we first divide it into $5 \times 5 \times 7$ blocks (175 dim vectors). We use the first 50 principal components as feature vectors. Our goal is then to segment the image into (typically more than 2) areas by clustering those feature vectors. With PPC, we can encode the partial class information into do-not-links.

[7] We use the first seven moderate resolution imaging spectroradiometer (MODIS) channels with bandwidths as follows (in nm): channel 1: 620–670; channel 2: 841–876; channel 3: 459–479; channel 4: 545–565; channel 5: 1230–1250; channel 6: 1628–1652; channel 7: 2105–2155.

For hard-PPC, we use half of the data samples for training and the rest for test. Hard do-not-link constraints (only on training set) are generated as follows. For each block in the nonsnow area, we randomly choose (without replacement) six blocks from the snow area to build do-not-link constraints. By doing this, we achieve cliques with size seven (1 nonsnow block + 6 snow blocks). As in section 5.1, we apply the model fit with hard-PPC to the test set and combine the clustering results on both data sets into a complete picture. Clearly, the clustering task is nontrivial for any $M > 2$. Typical clustering results of three-component standard GMM and three-component PPC are shown as Figures 11c and 11d, respectively. Standard GMM gives a clustering that is clearly in disagreement with the human labeling in Figure 11b. The hard-PPC segmentation makes far fewer misassignments of snow areas (tagged white and gray) to nonsnow (black) than does the GMM. The hard-PPC segmentation properly labels almost all of the nonsnow regions as nonsnow. Furthermore, the segmentation of the snow areas into the two classes (not labeled) tagged white and gray in Figure 11d reflects subtle differences in the snow regions captured by the gray-scale image from spectral channel 1, as shown in Figure 11a.

5.2.2 Experiment 4: Soft Links from Continuity. In this section, we present an example where soft constraints come from continuity. As in the previous experiment, we try to do image segmentation based on clustering. The image is divided into blocks and rearranged into feature vectors. We use a GMM to model those feature vectors, with each gaussian component representing one texture. However, standard GMM often fails to give good segmentations because it cannot make use of the spatial continuity of image, which is essential in many image segmentation models, such as random field (Bouman & Shapiro, 1994). In our algorithm, the spatial continuity is incorporated as the soft link preferences with uniform weight between each block and its neighbors. As described in section 2.4, the weight w of the soft link can be given as

$$w = \frac{1}{2} \log\left(\frac{1-q}{q}\right), \tag{5.2}$$

where q is the ratio of softly linked adjacent pairs that are not in the same class. Usually q is given by an expert or estimated from segmentation results of similar images. In this experiment, we assume we already know the ratio q, which is calculated from the label of the image.

The complete data likelihood is

$$P_p(X, Z|\Theta, W^p) = \frac{1}{\Omega} P_s(X, Z|\Theta) \prod_i \prod_{j \in U(i)} \exp(w\, \delta(z_i, z_j)), \tag{5.3}$$

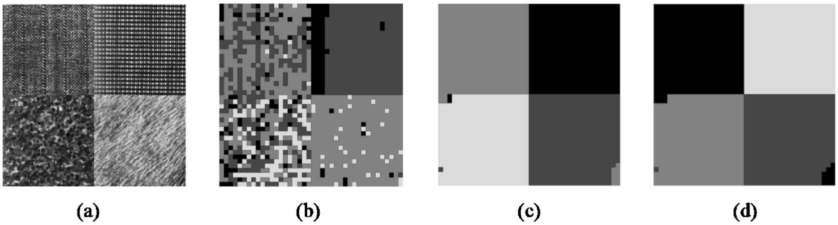


Figure 12: (a) Texture combination. (b) Clustering result of standard GMM. (c) Clustering result of soft-PPC with Gibbs sampling. (d) Clustering result of soft-PPC with mean field approximation. Panels b to d are shaded according to the blocks assignments to clusters.

where $U(i)$ means the neighbors of the ith block. The EM algorithm can be roughly interpreted as iterating on two steps: (1) estimating the texture description (parameters of mixture model) based on segmentation and (2) segmenting the image based on the texture description given by step 1. Since exact calculation of the posterior probability is intractable due to the large clique containing all samples, we have to resort to approximation methods. In this experiment, both the Gibbs sampling (see section 3.2) and the mean field approximation (see section 3.3) are used for posterior estimation. For Gibbs sampling, equation 3.1 is reduced to

$$P_p(z_i | Z_{-i}, X, \Theta, W^p) \propto P_s(x_i, z_i | \Theta) \prod_{j \in U(i)} \exp(2w \; \delta(z_i, z_j)).$$

The mean field equation, 3.5, is reduced to

$$Q_i(z_i) \leftarrow \frac{1}{\Omega_i} P_s(x_i, z_i | \Theta) \prod_{j \in U(i)} \exp(2w \; Q_j(z_i)).$$

The image shown in Figure 12a is built from four Brodatz textures.[8] This image is divided into 7 × 7 blocks and then rearranged to 49-dim vectors. We use those vectors' first five principal components as the associated feature vectors. A typical clustering result of four-component standard GMM is shown in Figure 12b. For soft-PPC, the soft links with weight w calculated from equation 5.2 are added between each block and its four neighbors. Figures 12c and 12d are the clustering result of four-component soft-PPC with respectively Gibbs sampling and mean field approximation. One run with Gibbs sampling takes around 160 minutes on a PC with Pentium 4,

[8] Downloaded from http://sipi.usc.edu/services/database/Database.html, April 2004.

2.0 GHZ processor, whereas the algorithm using the mean field approximation takes only 3.1 minutes. Although mean field approximation is about 50 times faster than Gibbs sampling, the clustering results are comparable according to Figure 12. Compared to the result given by standard GMM, soft-PPC with both approximation methods achieves significantly better segmentation after incorporating spatial continuity.

6 Conclusion and Future Work

We have proposed a probabilistic clustering model that incorporates prior knowledge in the form of pairwise relations between samples. Unlike previous work in semisupervised clustering, our model formulates clustering preferences as a Bayesian prior over the assignment of data points to clusters and so naturally accommodates both hard constraints and soft preferences. Unlike many semisupervised learning methods (Szummer & Jaakkola, 2001; Zhou & Schölkopf, 2004; Zhu, Lafferty, & Ghahramani, 2003) addressing labeled subsets, PPC returns a fitted parametric density model and thus can deal with unseen data. Experiments on different data sets have shown that pairwise relations can consistently improve the performance of the clustering process.

Despite its success, PPC has limitations. First, PPC often needs a substantial proportion of samples involved in pairwise relations to give good results. Indeed, if we have the number of relations fixed and keep adding samples without any new relations, the algorithm will finally degenerate into unsupervised learning (clustering). To overcome this, one can instead build a semisupervised model based on discriminative models such as a neural network or gaussian process classifier and use the pairwise relations in the form of hint (Sill & Abu-Mostafa, 1996) or observation (Zhu et al., 2003). Second, since PPC is based on the gaussian mixture model, it works well when the data in each class can be approximated by a gaussian distribution. When this condition is not satisfied, PPC could lead to poor results. One way to alleviate this situation is to use multiple clusters to model one class, an interesting direction for future exploration. Third, in choosing the weight matrix W^p, although our design works well on some data sets, it is not clear how to set the weight for a more general situation.

In this letter, we implement hard constraints using equation 2.10. Alternatively, we can approximate hard constraints by using large $|W_{ij}|$ for every constrained pair (x_i, x_j). Indeed, from equations 2.5 to 2.8, when a constraint with large weight is violated in assignments Z, the prior probability $P(Z|\Theta, W^p)$ will be close to zero. The value of $P(Z|\Theta, W^p)$ with such a Z can be made arbitrarily small by increasing the corresponding weight. This is convenient when we want to model soft and hard relations at the same time. This situation is not covered in this letter, but remains an interesting direction for future exploration.

To address the computational difficulty caused by large cliques, we propose two approximation methods: Gibbs sampling and mean field approximation. We also observe that Gibbs sampling can be fairly slow for large cliques. One way to address this problem is to use fewer sampling passes (and thus a cruder approximate inference) in the early phase of EM training and gradually increase the number of sampling passes (and a finer approximation) when EM is close to convergence. By doing this, we may be able to achieve a much faster algorithm without sacrificing too much precision. For the mean field approximation, the bias brought by the independence assumption among $Q_i(\cdot)$ could be severe for some problems. We can ameliorate this, as suggested by Jaakkola (2004), by retaining more substructure of the original graphical model (for PPC, it is expressed in W^p), while still keeping the computation tractable.

Appendix A: Hard Constraints Limit

In this appendix, we prove that when $|W_{ij}| \to \infty$ for each specified pair (x_i, x_j), the complete likelihood of PPC can written as in equation 2.10, and thus equivalent to the model proposed by Shental et al. (2003).

In the model proposed by Shental et al. (2003), the complete likelihood is written as

$$\begin{aligned}P(X, Z|\Theta, E_\Omega) &= \frac{1}{\Omega}\prod_{c_i}\delta_{y_{c_i}}\prod_{a_i^1 \neq a_i^2}(1-\delta(y_{a_i^1}, y_{a_i^2}))\prod_{i=1}^{N}P_s(z_i|\Theta)P_s(x_i|z_i,\Theta)\\ &= \frac{1}{\Omega}\prod_{c_i}\delta_{y_{c_i}}\prod_{a_i^1 \neq a_i^2}(1-\delta(y_{a_i^1}, y_{a_i^2}))P_s(X, Z|\Theta),\end{aligned}$$

where E_Ω stands for the pairwise constraints, $\delta_{y_{c_i}}$ is 1 iff all the points in the chunklet (the clique of samples connected with only hard links) c_i have the same label, and (a_i^1, a_i^2) is the index of the sample pair with hard do-not-link between them. This is equivalent to

$$P(X, Z|\Theta, E_\Omega) = \begin{cases}\frac{1}{\Omega}P_s(X, Z|\Theta) & Z \text{ satisfies all the constraints}\\ 0 & \text{otherwise}\end{cases}. \tag{A.1}$$

In the corresponding PPC model with hard constraints, we have

$$W_{ij}^p = \begin{cases}+\infty & i \text{ and } j \text{ is linked}\\ -\infty & i \text{ and } j \text{ is do-not-linked}\,.\\ 0 & \text{no relation}\end{cases} \tag{A.2}$$

According to equation 2.5 and A.1, to prove

$$P(X, Z|\Theta, E_\Omega) = P_p(X, Z|\Theta, W^p),$$

we need only to prove

$$P_p(Z|\Theta, W^p) = 0,$$

for all the Z that violate the constraints, that is,

$$P_p(Z|\Theta, W^p) = \frac{\prod_k \pi_{z_k} \prod_{i\neq j} \exp\left(W^p_{ij}\ \delta(z_i, z_j)\right)}{\sum_Z \prod_l \pi_{z_l} \prod_{m\neq n} \exp\left(W^p_{mn}\ \delta(z_m, z_n)\right)} = 0.$$

First, let us assume Z violates one link between pair (α, β) $\left(W^p_{\alpha\beta} = +\infty\right)$. We have

$$z_\alpha \neq z_\beta \Rightarrow \delta(z_\alpha, z_\beta) = 0 \Rightarrow \exp(W^p_{\alpha\beta}\ \delta(z_\alpha, z_\beta)) = 1.$$

We assume the constraints are consistent. In other words, there is at least one Z that satisfies all the constraints. We can denote one such Z by Z^*. We also assume each component has a positive prior probability. It is straightforward to show that

$$P_p(Z^*|\Theta, W^p) > 0.$$

Then it is easy to show

$$\begin{aligned}
P_p(Z|\Theta, W^p) &= \frac{\prod_k \pi_{z_k} \prod_{i\neq j} \exp\left(W^p_{ij}\ \delta(z_i, z_j)\right)}{\sum_Z \prod_l \pi_{z_l} \prod_{m\neq n} \exp\left(W^p_{mn}\ \delta(z_m, z_n)\right)} \\
&\leq \frac{\prod_k \pi_{z_k} \prod_{i\neq j} \exp\left(W^p_{ij}\ \delta(z_i, z_j)\right)}{\prod_k \pi_{z_k^*} \prod_{i\neq j} \exp\left(W^p_{mn}\ \delta(z_i^*, z_j^*)\right)} \\
&= \left(\prod_k \frac{\pi_{z_k}}{\pi_{z_k^*}} \prod_{(i,j)\neq(\alpha,\beta)} \frac{\exp\left(W^p_{ij}\ \delta(z_i, z_j)\right)}{\exp\left(W^p_{ij}\ \delta(z_i^*, z_j^*)\right)}\right) \frac{\exp\left(2W^p_{\alpha\beta}\ \delta(z_\alpha, z_\beta)\right)}{\exp\left(2W^p_{\alpha\beta}\ \delta(z_\alpha^*, z_\beta^*)\right)} \\
&= \left(\prod_k \frac{\pi_{z_k}}{\pi_{z_k^*}} \prod_{(i,j)\neq(\alpha,\beta)} \frac{\exp\left(W^p_{ij}\ \delta(z_i, z_j)\right)}{\exp\left(W^p_{ij}\ \delta(z_i^*, z_j^*)\right)}\right) \frac{1}{\exp\left(2W^p_{\alpha\beta}\ \delta(z_\alpha^*, z_\beta^*)\right)}.
\end{aligned}$$

Since Z^* satisfies all the constraints, we must have

$$\prod_{(i,j)\neq(\alpha,\beta)} \frac{\exp\left(W_{ij}^p\ \delta(z_i, z_j)\right)}{\exp\left(W_{ij}^p\ \delta(z_i^*, z_j^*)\right)} \leq 1.$$

So we have

$$P_p(Z|\Theta, W^p) \leq \left(\prod_k \frac{\pi_{z_k}}{\pi_{z_k^*}}\right) \frac{1}{\exp\left(2W_{\alpha\beta}^p\ \delta(z_\alpha^*, z_\beta^*)\right)}.$$

When

$$W_{\alpha\beta}^p \to +\infty,$$

we have

$$\frac{1}{\exp\left(2W_{\alpha\beta}^p\ \delta(z_\alpha^*, z_\beta^*)\right)} \to 0$$

and then

$$P_p(Z|\Theta, W^p) \leq \left(\prod_k \frac{\pi_{z_k}}{\pi_{z_k^*}}\right) \frac{1}{\exp\left(2W_{\alpha\beta}^p\ \delta(z_\alpha^*, z_\beta^*)\right)} \to 0. \tag{A.3}$$

The do-not-link case can be proved in a similar way.

Appendix B: Prior for Noisy Pairwise Relations

In this appendix, we show how to derive weight W from the certainty value γ_{ij} for each pair (x_i, x_j). Let E denote those original (noise-free) labeled pairwise relations and $\tilde{E}$ the noisy version delivered to us. If we know the original pairwise relations E, we only have to consider the cluster assignments that are consistent with E and neglect the others, that is, the prior probability of Z is

$$P(Z|\Theta, E) = \begin{cases} \frac{1}{\Omega_E} P_s(Z|\Theta) & Z \text{ is consistent with } E \\ 0 & \text{otherwise} \end{cases},$$

where Ω_E is the normalization constant for E: $\Omega_E = \sum_{Z:\ \text{consistent with } E} P_s(Z|\Theta)$. Since we know $\tilde{E}$ and the associated certainty values $\Gamma = \{\gamma_{ij}\}$, we know

$$P(Z|\Theta, \tilde{E}, \Gamma) = \sum_E P(Z|\Theta, E, \tilde{E}, \Gamma) P(E|\tilde{E}, \Gamma) \tag{B.1}$$

$$= \sum_E P(Z|\Theta, E) P(E|\tilde{E}, \Gamma). \tag{B.2}$$

Let $E(Z) \equiv$ the unique E that is consistent with Z. From equation B.2, we know

$$\begin{aligned} P(Z|\Theta, \tilde{E}, \Gamma) &= P_p(Z|\Theta, E(Z)) P(E(Z)|\tilde{E}, \Gamma) \\ &= \frac{1}{\Omega_E} P_s(Z|\Theta) P(E(Z)|\tilde{E}, \Gamma) = \frac{1}{\Omega_E} P(E(Z)|\tilde{E}, \Gamma) P_s(Z|\Theta). \end{aligned}$$

If we ignore the variation of Ω_E over E, we can get an approximation of $P(Z|\Theta, \tilde{E}, \Gamma)$, denoted as $P_a(Z|\Theta, \tilde{E}, \Gamma)$:

$$\begin{aligned} P_a(Z|\Theta, \tilde{E}, \Gamma) &= \frac{1}{\Omega_a} P_s(Z|\Theta) P(E(Z)|\tilde{E}, \Gamma) \\ &= \frac{1}{\Omega_a} P_s(Z|\Theta) \prod_{i<j} \gamma_{ij}^{H_{ij}(\tilde{E}, z_i, z_j)} (1 - \gamma_{ij})^{1 - H_{ij}(\tilde{E}, z_i, z_j)}, \end{aligned}$$

where Ω_a is the new normalization constant: $\Omega_a = \sum_Z P_s(Z|\Theta)$ $P(E(Z)|\tilde{E}, \Gamma)$ and

$$H_{ij}(\tilde{E}, z_i, z_j) = \begin{cases} 1 & (z_i, z_j) \text{ is consistent with } \tilde{E} \\ 0 & \text{otherwise} \end{cases}.$$

We argue that $P_a(Z|\Theta, \tilde{E}, \Gamma)$ is equal to a PPC prior probability $P_p(Z|\Theta, W^p)$ with

$$W_{ij}^p = \begin{cases} \frac{1}{2} \log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) & (z_i, z_j) \text{ is specified as must-linked in } \tilde{E} \\ -\frac{1}{2} \log\left(\frac{\gamma_{ij}}{1-\gamma_{ij}}\right) & (z_i, z_j) \text{ is specified as cannot-linked in } \tilde{E} \\ 0 & \text{otherwise.} \end{cases} \tag{B.3}$$

This can be easily proven by verifying

$$\frac{P_p(Z|\Theta, W^p)}{P_a(Z|\Theta, \tilde{E}, \Gamma)} = \frac{\Omega_a}{\Omega_w} \prod_{i<j, W_{ij^p} \neq 0} \gamma_{ij}^{\text{sign}(W_{ij}^p) - 1} (1 - \gamma_{ij})^{-\text{sign}(W_{ij}^p)} = \text{constant}.$$

Since both $P_a(Z|\Theta, \tilde{E}, \Gamma)$ and $P_p(Z|\Theta, W^p)$ are normalized, we know

$$P_a(Z|\Theta, \tilde{E}, \Gamma) = P_p(Z|\Theta, W^p).$$

Appendix C: From PPC to Constrained K-Means

In this appendix, we show how to derive K-means model with soft and hard constraints from PPC.

C.1 From PPC to K-Means with Soft Constraints. The adopted cost function for K-means with soft constraints is

$$J(\mu, Z) = \frac{1}{2}\sum_{i=1}^{N} ||x_i - \mu_{z_i}||^2 + \sum_{(i,j)\in\mathcal{L}} a_{ij}(z_i \neq z_j) + \sum_{(i,j)\in\mathcal{N}} b_{ij}(z_i = z_j), \tag{C.1}$$

where μ_k is the center of the kth cluster. Equation 4.1 can be rewritten as

$$J(\mu, Z) = \frac{1}{2}\sum_{i=1}^{N} ||x_i - \mu_{z_i}||^2 - \sum_{ij} W_{ij}^p \delta(z_i, z_j) + C, \tag{C.2}$$

with $C = -\sum_{(i,j)\in\mathcal{L}} a_{ij}$ is a constant and

$$W_{ij}^p = \begin{cases} a_{ij} & (i,j) \in \mathcal{L} \\ -b_{ij} & (i,j) \in \mathcal{N} \\ 0 & \text{otherwise.} \end{cases} \tag{C.3}$$

The clustering process includes minimizing the cost function $J(\mu, Z)$ over both the model parameters $\mu = \{\mu_1, \mu_2, \ldots, \mu_M\}$ and cluster assignment $Z = \{z_1, z_2, \ldots, z_N\}$. The optimization is usually done iteratively with a modified Linde-Buzo-Gray (LBG) algorithm. Assume we have the PPC model with the matrix W^p the same as in equation C.2. We further constrain each gaussian component to be spherical with radius σ. The complete data likelihood for the PPC model is

$$P(X, Z|\Theta, W^p) = \frac{1}{\Omega}\prod_{i=1}^{N}\left\{\pi_{z_i} \exp\left(-\sum_{i=1}^{N}\frac{||x_i - \mu_{z_i}||^2}{2\sigma^2}\right)\right\}$$
$$\prod_{mn} \exp\left(W_{mn}^p \delta(z_m, z_n)\right), \tag{C.4}$$

where Ω is the normalizing constant and μ_k is the mean of the kth gaussian component. To build its connection to the cost function in equation C.2, we consider the following scaling:

$$\sigma \to \alpha\sigma, \qquad W_{ij}^{p} \to W_{ij}^{p}/\alpha^2. \tag{C.5}$$

The complete data likelihood with the scaling parameters α is

$$P_\alpha(X, Z|\Theta, W^p) = \frac{1}{\Omega(\alpha)} \prod_{i=1}^{N} \left\{ \pi_{z_i} \exp\left(-\sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2\alpha^2\sigma^2} \right) \right\} \prod_{mn} \exp\left(\frac{W_{mn}^{p}}{\alpha^2} \delta(z_m, z_n) \right). \tag{C.6}$$

It can be shown that when $\alpha \to 0$, the maximum data likelihood will dominate the data likelihood:

$$\lim_{\alpha\to 0} \frac{\max_Z P_\alpha(X, Z|\Theta, W^p)}{\sum_Z P_\alpha(X, Z|\Theta, W^p)} = 1. \tag{C.7}$$

To prove equation C.7, we first show that when α is small enough, we have

$$\arg\max_Z P_\alpha(X, Z|\Theta, W^p) = Z^* \equiv \arg\min_Z \left\{ \sum_{i=1}^{N} \frac{||x_i - \mu_{z_i^*}||^2}{2} - \sum_{mn} W_{mn}^{p} \delta\left(z_m^*, z_n^*\right) \right\}. \tag{C.8}$$

Proof of Equation C.8. Assume Z' is any cluster assignment different from Z^*. We only need to show that when α is small enough,

$$P_\alpha(X, Z^*|\Theta, W^p) > P_\alpha(X, Z'|\Theta, W^p). \tag{C.9}$$

To prove equation C.9, we notice that

$$\begin{aligned} &\log P_\alpha(X, Z^*|\Theta, W^p) - \log P_\alpha(X, Z'|\Theta, W^p) \\ &\quad = \sum_{i=1}^{N} \left(\log \pi_{z_i^*} - \log \pi_{z_i'} \right) + \frac{1}{\alpha^2} \left\{ \sum_{i=1}^{N} \left(\frac{||x_i - \mu_{z_i'}||^2}{2} - \frac{||x_i - \mu_{z_i^*}||^2}{2} \right) \right. \\ &\qquad \left. - \sum_{mn} W_{mn}^{p} \left(\delta\left(z_m', z_n'\right) - \delta\left(z_m^*, z_n^*\right) \right) \right\}. \end{aligned} \tag{C.10}$$

Since $Z^* = \arg\min_Z \left\{ \sum_{i=1}^N \frac{\|x_i - \mu_{z_i^*}\|^2}{2} - \sum_{mn} W_{mn}^p \delta\big(z_m^*, z_n^*\big) \right\}$, we have

$$\sum_{i=1}^N \left(\frac{\|x_i - \mu_{z_i'}\|^2}{2} - \frac{\|x_i - \mu_{z_i^*}\|^2}{2} \right) - \sum_{mn} W_{mn}^p \big(\delta\big(z_m', z_n'\big) - \delta\big(z_m^*, z_n^*\big)\big) > 0. \tag{C.11}$$

Let $\varepsilon = \sum_{i=1}^N \left(\frac{\|x_i - \mu_{z_i'}\|^2}{2} - \frac{\|x_i - \mu_{z_i^*}\|^2}{2} \right) - \sum_{mn} W_{mn}^p \big(\delta\big(z_m', z_n'\big) - \delta\big(z_m^*, z_n^*\big)\big)$. We can see that when α is small enough,

$$\begin{aligned}&\log P_\alpha(X, Z^*|\Theta, W^p) - \log P_\alpha(X, Z'|\Theta, W^p)\\ &\quad = \sum_{i=1}^N \big(\log \pi_{z_i^*} - \log \pi_{z_i'}\big) + \frac{\varepsilon}{\alpha^2} > 0.\end{aligned} \tag{C.12}$$

It is obvious from equation C.12 that for any Z' different from Z^*,

$$\begin{aligned}&\lim_{\alpha \to 0} \log P_\alpha(X, Z^*|\Theta, W^p) - \log P_\alpha(X, Z'|\Theta, W^p)\\ &\quad = \lim_{\alpha \to 0} \sum_{i=1}^N \big(\log \pi_{z_i^*} - \log \pi_{z_i'}\big) + \frac{\varepsilon}{\alpha^2}\\ &\quad = +\infty,\end{aligned}$$

or equivalently

$$\lim_{\alpha \to 0} \frac{P_\alpha(X, Z'|\Theta, W^p)}{P_\alpha(X, Z^*|\Theta, W^p)} = 0, \tag{C.13}$$

which proves equation C.7. As the result of equation C.7, when optimizing the model parameters, we can equivalently maximize $\max_Z P_\alpha(X, Z|\Theta, W^p)$ over Θ. It is then a joint optimization problem:

$$\max_{\Theta, Z} P_\alpha(X, Z|\Theta, W^p).$$

Following the same thought, we find the soft posterior probability of each sample (as in conventional mixture model) becomes hard membership (as in K-means). This fact can be simply proved as follows. The posterior probability of sample x_i to component k is

$$P_\alpha(z_i = k|X, \Theta, W^p) = \frac{\sum_{Z|z_i=k} P_\alpha(X, Z|\Theta, W^p)}{\sum_Z P_\alpha(X, Z|\Theta, W^p)}.$$

From equation C.7, it is easy to see that

$$\lim_{\alpha\to 0} P_\alpha(z_i = k|X, \Theta, W^p) = \begin{cases} 1 & z_i^* = k \\ 0 & \text{otherwise.} \end{cases} \tag{C.14}$$

The negative logarithm of the complete likelihood P_α is then:

$$\begin{aligned} J_\alpha(\Theta, Z) &= -\log P_\alpha(X, Z|\Theta, W^p) \\ &= -\sum_{i=1}^{N} \log \pi_{z_i} + \sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2\alpha^2} - \sum_{mn} \frac{W_{mn}^p}{\alpha^2}\delta(z_m, z_n) + \log(\Omega(\alpha)) \\ &= -\sum_{i=1}^{N} \log \pi_{z_i} + \frac{1}{\alpha^2}\left(\sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2} - \sum_{mn} W_{mn}^p \delta(z_m, z_n)\right) + C, \end{aligned}$$

where $C = \log \Omega(\alpha)$ is a constant. It is obvious that when $\alpha \to 0$, we can neglect the term $-\sum_{i=1}^{N} \log \pi_{z_i}$. Hence, the only model parameters left for adjusting are the gaussian means μ. We only have to consider the new cost function

$$\tilde{J}_\alpha(\mu, Z) = \frac{1}{\alpha^2}\left(\sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2} - \sum_{mn} W_{mn}^p \delta(z_m, z_n)\right), \tag{C.15}$$

the optimization of which is obviously equivalent to equation C.1. So we can conclude that when $\alpha \to 0$ in equation C.5, the PPC model shown in equation 3.4 becomes a K-means model with soft constraints.

C.2 From PPC to K-Means with Hard Constraints (COPK-Means). COPK-means is a hard clustering algorithm with hard constraints. The goal is to find a set of cluster centers μ and clustering result Z that minimizes the cost function

$$\sum_{i=1}^{N} ||x_i - \mu_{z_i}||^2, \tag{C.16}$$

while subject to the constraints

$$z_i = z_j, \quad \text{if } (x_i, x_j) \in \mathcal{L} \tag{C.17}$$

$$z_i \neq z_j, \quad \text{if } (x_i, x_j) \in \mathcal{N}. \tag{C.18}$$

Assume we have the PPC model with soft relations represented with the matrix W^p such that

$$W^p_{ij} = \begin{cases} w & (x_i, x_j) \in \mathcal{L} \\ -w & (x_i, x_j) \in \mathcal{N} \\ 0 & \text{otherwise,} \end{cases} \tag{C.19}$$

where $w > 0$. We further constrain each gaussian component to be spherical with radius σ. The complete data likelihood for PPC model is

$$P(X, Z|\Theta, W^p) = \frac{1}{\Omega} \prod_{i=1}^{N} \left\{ \pi_{z_i} \exp\left(-\sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2\sigma^2} \right) \right\} \prod_{(m,n)\in\mathcal{L}} \exp(w\delta(z_m, z_n)) \prod_{(m',n')\in\mathcal{N}} \exp(-w\delta(z_{m'}, z_{n'})), \tag{C.20}$$

where μ_k is the mean of the kth gaussian component. There are infinite ways to get equations C.16 to C.18 from equation C.20, but we consider the following scaling with factor β:

$$\sigma \to \beta\sigma, \quad W^p_{ij} \to W^p_{ij}/\beta^3. \tag{C.21}$$

The complete data likelihood with the scaled parameters is

$$P_\beta(X, Z|\Theta, W^p) = \frac{1}{\Omega(\beta)} \prod_{i=1}^{N} \left\{ \pi_{z_i} \exp\left(-\sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2\beta^2\sigma^2} \right) \right\} \prod_{(m,n)\in\mathcal{L}} \exp\left(\frac{w}{\beta^3}\delta(z_m, z_n) \right) \prod_{(m',n')\in\mathcal{N}} \exp\left(-\frac{w}{\beta^3}\delta(z_{m'}, z_{n'}) \right). \tag{C.22}$$

As established in previous section, when $\beta \to 0$, the maximum data likelihood will dominate the data likelihood

$$\lim_{\beta\to 0} \frac{\max_Z P_\beta(X, Z|\Theta, W^p)}{\sum_Z P_\beta(X, Z|\Theta, W^p)} = 1.$$

As a result, when optimizing the model parameters Θ, we can equivalently maximize $\max_Z P_\beta(X, Z|\Theta, W^p)$. Also, the soft posterior probability (as in the conventional mixture model) becomes hard membership (as in K-means).

The negative logarithm of the complete likelihood P_β is then:

$$J_\beta(\Theta, Z) = -\sum_{i=1}^{N} \log \pi_{z_i} + C + \frac{1}{\beta^2}\left(\sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2} + \frac{1}{\beta}\left(\sum_{(m',n')\in\mathcal{N}} w\delta(z_{m'}, z_{n'}) - \sum_{(m,n)\in\mathcal{L}} w\delta(z_m, z_n)\right)\right), \tag{C.23}$$

where $C = \log \Omega(\beta)$ is a constant. It is obvious that when $\beta \to 0$, we can neglect the term $-\sum_{i=1}^{N} \log \pi_{z_i}$. Hence, we only have to consider the new cost function

$$\tilde{J}_\beta(\mu, Z) = \frac{1}{\beta^2}\left(\sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2} + \frac{1}{\beta}\left(\sum_{(m',n')\in\mathcal{N}} w\delta(z_{m'}, z_{n'}) - \sum_{(m,n)\in\mathcal{L}} w\delta(z_j, z_k)\right)\right), \tag{C.24}$$

the minimization of which is obviously equivalent to the following equation since we can neglect the constant factor $\frac{1}{\beta^2}$:

$$\tilde{\tilde{J}}_\beta(\mu, Z) = \sum_{i=1}^{N} \frac{||x_i - \mu_{z_i}||^2}{2} + \frac{w}{\beta} J_c(Z), \tag{C.25}$$

where $J_c(Z) = \sum_{(m',n')\in\mathcal{N}} \delta(z_{m'}, z_{n'}) - \sum_{(m,n)\in\mathcal{L}} \delta(z_m, z_n)$ is the cost function term from pairwise constraints.

Let $S_Z = \{Z | z_i = z_j \text{ if } W_{ij}^p > 0; z_i \neq z_j \text{ if } W_{ij}^p < 0;\}$. We assume the pairwise relations are consistent, that is, $S_Z \neq \emptyset$. Obviously, all Z in S_Z achieve the same minimum value of the term $J_c(Z)$, that is

$$\forall Z \in S_Z, Z' \in S_Z \quad J_c(Z) = J_c(Z')$$
$$\forall Z \in S_Z, Z'' \notin S_Z \quad J_c(Z) < J_c(Z'').$$

It is obvious that when $\beta \to 0$, any Z that minimizes $\tilde{\tilde{J}}_\beta(\mu, Z)$ must be in S_Z. So the minimization of equation C.22 can be finally casted into the following form,

$$\min_{Z,\mu} \sum_{i=1}^{N} ||x_i - \mu_{z_i}||^2$$

subject to $Z \in S_Z$,

which is apparently equivalent to equations C.16 to C.18. So we can conclude that $\beta \to 0$ in equation C.21, and the PPC model shown in equation C.20 becomes a K-means model with hard constraints.

Acknowledgments

We thank Ashok Srivastava of NASA Ames Research Center for providing satellite image data and the reviewers for comments leading to a stronger letter. This work was funded by NASA Collaborative Agreement NCC 2-1264 and NSF grant OCI-0121475.

References

Ambroise, C., Dang, M., & Govaert, G. (1997). Clustering of spatial data by the EM algorithm. In A. Soares, J. Gómez-Hernández, & R. Froidevaux (Eds.), *Geostatistics for environmental applications* (vol. 3, pp. 493–504). Norwell, MA: Kluwer.

Basu, S., Bannerjee, A., & Mooney, R. (2002). Semi-supervised clustering by seeding. In C. Sammut & A. Hoffmann (Eds.), *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 19–26). San Francisco: Morgan Kaufmann.

Basu, S., Bilenko, M., & Mooney, R. J. (2004). A probabilistic framework for semisupervised clustering. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 59–68). New York: ACM.

Bilenko, M., Basu, S., & Mooney, R. (2004). Integrating constraints and metric learning in semi-supervised clustering. In C. Brodley (Ed.), *Proceedings of the Twenty-First International Conference on Machine Learning* (pp. 11–18). New York: ACM.

Bouman, C., & Shapiro, M. (1994). A multiscale random field model for Bayesian image segmentation. *IEEE Transactions on Image Processing*, *3*, 162–177.

Cohn, D., Caruana, R., & McCallum, A. (2003). *Semi-supervised clustering with user feedback* (Tech. Rep. TR2003-1892). Ithaca, NY: Cornell University.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Jaakkola, T. (2004). Tutorial on variational approximation methods. In D. Saad & M. Opper (Eds.), *Advanced mean field methods: Theory and practice* (pp. 129–160). Cambridge, MA: MIT Press.

Klein, D., Kamvar, S., & Manning, C. (2002). From instance level to space-level constraints: Making the most of prior knowledge in data clustering. In C. Sammut, & A. Hoffmann (Eds.), *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 307–313). San Francisco: Morgan Kaufmann.

Lange, T., Law, M., Jain, A., & Buhmann, J. (2005). Learning with constrained and unlabelled data. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 730–737). Los Alamitos, CA: IEEE Computer Society Press.

Law, M., Topchy, A., & Jain, A. (2004). Clustering with soft and group constraints. In A. Fred, T. Caelli, R. Duin, A. Campilho, & D. Ridder (Eds.), *Joint IAPR International Workshop on Syntactical and Structural Pattern Recognition and Statistical Pattern Recognition* (pp. 662–670). Berlin: Springer-Verlag.

Law, M., Topchy, A., & Jain, A. (2005). Model-based clustering with probabilistic constraints. In *Proceedings of SIAM Data Mining* (pp. 641–645). Philadelphia: SIAM.

Lu, Z., & Leen, T. (2005). Semi-supervised learning with penalized probabilistic clustering. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems, 17* (pp. 849–856). Cambridge, MA: MIT Press.

Neal, R. (1993). *Probabilistic inference using Markov chain Monte Carlo methods* (Tech. Rep. CRG-TR-93-1). Toronto: Computer Science Department, Toronto University.

Shental, N., Bar-Hillel, A., Hertz, T., & Weinshall, D. (2003). *Computing gaussian mixture models with EM using side-information* (Tech. Rep. 2003-43). Leibniz: Leibniz Center for Research in Computer Science.

Shental, N., Bar-Hillel, A., Hertz, T., & Weinshall, D. (2004). Computing gaussian mixture models with EM using equivalence constraints. In L. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 16* (pp. 505–512). Cambridge, MA: MIT Press.

Sill, J., & Abu-Mostafa, Y. (1996). Monotonicity hints. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in neural information processing systems, 8* (pp. 634–640). Cambridge, MA: MIT Press.

Srivastava, A., Oza, N. C., & Stroeve, J. (2005). Virtual sensors: Using data mining techniques to efficiently estimate remote sensing spectra. *IEEE Transactions on Geoscience and Remote Sensing, 43*(3), 590–599.

Szummer, M., & Jaakkola, T. (2001). Partially labeled classification with Markov random walks. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14*. Cambridge, MA: MIT Press.

Theiler, J., & Gisler, G. (1997). A contiguity-enhanced K-means clustering algorithm for unsupervised multispectral image segementation. In B. Javidi & D. Psaltis (Eds.), *Proceedings of SPIE* (Vol. 3159, pp. 108–118). Bellingham, WA: SPIE.

Wagstaff, K. (2002). *Intelligent clustering with instance-level constraints*. Unpublished doctoral, dissertation, Cornell University.

Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S. (2001). Constrained K-means clustering with background knowledge. In C. Brodley & A. Danyluk (Eds.), *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 577–584). San Francisco: Morgan Kaufmann.

Xing, E., Ng, A., Jordan, M., & Russe, S. (2003). Distance metric learning with applications to clustering with side information. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 505–512). Cambridge, MA: MIT Press.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics* (pp. 189–196). Morristown, NJ: Association for Computational Linguistics.

Zhou, D., & Schölkopf, B. (2004). Learning from labeled and unlabeled data using random walks. In C. Rasmussen, H. Buelthoff, M. Giese, & B. Scholköpf (Eds.), *26th Annual Meeting of the German Association for Pattern Recognition (DAGM 2004)*. Berlin: Springer.

Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). *Semi-supervised learning: From Gaussian field to gaussian processes* (Tech. Rep. CMU-CS-03-175). Pittsburgh, PA: Computer Science Department, CMU.

Received October 21, 2005; accepted July 27, 2006.