High Dimensional Data Clustering

Charles Bouveyron[1,2], Stéphane Girard[1], and Cordelia Schmid[2]

[1] LMC-IMAG, BP 53, Université Grenoble 1, 38041 Grenoble Cedex 9, France
charles.bouveyron@imag.fr, stephane.girard@imag.fr

[2] INRIA Rhône-Alpes, Projet Lear, 655 av. de l'Europe, 38334 Saint-Ismier Cedex, France
cordelia.schmid@inrialpes.fr

Summary. Clustering in high-dimensional spaces is a recurrent problem in many domains, for example in object recognition. High-dimensional data usually live in different low-dimensional subspaces hidden in the original space. This paper presents a clustering approach which estimates the specific subspace and the intrinsic dimension of each class. Our approach adapts the Gaussian mixture model framework to high-dimensional data and estimates the parameters which best fit the data. We obtain a robust clustering method called High-Dimensional Data Clustering (HDDC). We apply HDDC to locate objects in natural images in a probabilistic framework. Experiments on a recently proposed database demonstrate the effectiveness of our clustering method for category localization.

Key words: Model-based clustering, high-dimensional data, dimension reduction, dimension reduction, parsimonious models.

1 Introduction

In many scientific domains, the measured observations are high-dimensional. For example, visual descriptors used in object recognition are often high-dimensional and this penalizes classification methods and consequently recognition. Popular clustering methods are based on the Gaussian mixture model and show a disappointing behavior when the size of the training dataset is too small compared to the number of parameters to estimate. To avoid overfitting, it is therefore necessary to find a balance between the number of parameters to estimate and the generality of the model. In this paper we propose a Gaussian mixture model which determines the specific subspace in which each class is located and therefore limits the number of parameters to estimate. The Expectation-Maximization (EM) algorithm [5] is used for parameter estimation and the intrinsic dimension of each class is determined automatically with the scree test of Cattell. This allows to derive a robust clustering method in high-dimensional spaces, called High Dimensional Data Clustering (HDDC). In order to further limit the number of parameters, it is possible to make additional assumptions on the model. We can for example assume that classes are spherical in their subspaces or fix some parameters to be common between classes.

We evaluate HDDC on a recently proposed visual recognition dataset [4]. We compare HDDC to standard clustering methods and to the state of the art results. We show that our approach outperforms existing results for object localization.

This paper is organized as follows. Section 2 presents the state of the art on clustering of high-dimensional data. In Section 3, we describe our parameterization of the Gaussian mixture model. Section 4 presents our clustering method, *i.e.* the estimation of the parameters and of the intrinsic dimensions. Experimental results for our clustering method are given in Section 5.

2 Related work on high-dimensional clustering

Many methods use global dimensionality reduction and then apply a standard clustering method. Dimension reduction techniques are either based on *feature extraction* or *feature selection*. Feature extraction builds new variables which carry a large part of the global information. The most known method is Principal Component Analysis (PCA) which is a linear technique. Recently, many non-linear methods have been proposed, such as Kernel PCA and non-linear PCA. In contrast, feature selection finds an appropriate subset of the original variables to represent the data. Global dimension reduction is often advantageous in terms of performance, but loses information which could be discriminant, *i.e.* clusters are often hidden in different subspaces of the original feature space and a global approach cannot capture this. It is also possible to use a parsimonious model [7] which reduces the number of parameters to estimate. It is for example possible to fix some parameters to be common between classes. These methods do not solve the problem of high dimensionality because clusters are usually hidden in different subspaces and many dimensions are irrelevant. Recent methods determine the subspaces for each cluster. Many subspace clustering methods use heuristic search techniques to find the subspaces. They are usually based on grid search methods and find dense clusterable subspaces [8]. The approach "mixtures of Probabilistic Principal Component Analyzers" [10] proposes a latent variable model and derives an EM based method to cluster high-dimensional data. Bocci *et al.* [1] propose a similar method to cluster dissimilarity data. In this paper, we introduce an unified approach for class-specific subspace clustering which includes these two methods and allows additional regularizations.

3 Gaussian mixture models for high-dimensional data

Clustering divides a given dataset $\{x_1, ..., x_n\}$ of n data points into k homogeneous groups. Popular clustering techniques use Gaussian Mixture Models (GMM), which assume that each class is represented by a Gaussian probability density. Data $\{x_1, ..., x_n\} \in \mathbb{R}^p$ are then modeled with the density $f(x, \theta) = \sum_{i=1}^{k} \pi_i \phi(x, \theta_i)$, where ϕ is a multi-variate normal density with parameter $\theta_i = \{\mu_i, \Sigma_i\}$ and π_i are mixing proportions. This model estimates full covariance matrices and therefore the number of parameters is very large in high dimensions. However, due to the *empty*

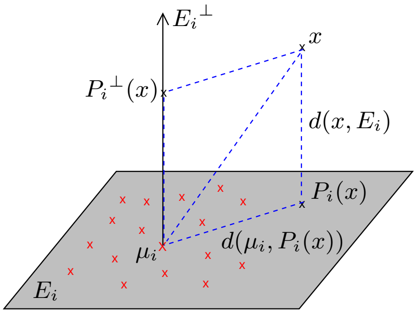


Fig. 1. The class-specific subspace $\mathbb{E}_i$.

space phenomenon we can assume that high-dimensional data live in subspaces with a dimensionality lower than the dimensionality of the original space. We therefore propose to work in low-dimensional class-specific subspaces in order to adapt classification to high-dimensional data and to limit the number of parameters to estimate.

3.1 The family of Gaussian mixture models

We remind that class conditional densities are Gaussian $\mathcal{N}(\mu_i, \Sigma_i)$ with means μ_i and covariance matrices Σ_i, $i = 1, ..., k$. Let Q_i be the orthogonal matrix of eigenvectors of Σ_i, then $\Delta_i = Q_i^t \, \Sigma_i \, Q_i$ is a diagonal matrix containing the eigenvalues of Σ_i. We further assume that Δ_i is divided into two blocks:

$$\Delta_i = \left(\begin{array}{cc} \boxed{\begin{array}{ccc} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{array}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{array}{ccc} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{array}} \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c} a \\ \ddots \\ a \end{array}}\right\} \quad d_i \\ \left.\vphantom{\begin{array}{c} b \\ \ddots \\ b \end{array}}\right\} (p - d_i) \end{array}$$

where $a_{ij} > b_i$, $\forall j = 1, ..., d_i$. The class specific subspace $\mathbb{E}_i$ is generated by the d_i first eigenvectors corresponding to the eigenvalues a_{ij} with $\mu_i \in \mathbb{E}_i$. Outside this subspace, the variance is modeled by the single parameter b_i. Let $P_i(x) = \tilde{Q}_i \tilde{Q}_i{}^t (x - \mu_i) + \mu_i$ be the projection of x on $\mathbb{E}_i$, where $\tilde{Q}_i$ is made of the d_i first columns of Q_i supplemented by zeros. Figure 1 summarizes these notations.

The mixture model presented above will be in the following referred to by $[a_{ij} b_i Q_i d_i]$. By fixing some parameters to be common within or between classes, we obtain a family of models which correspond to different regularizations. For example, if we fix the first d_i eigenvalues to be common within each class, we obtain the more restricted model $[a_i b_i Q_i d_i]$. The model $[a_i b_i Q_i d_i]$ is often robust and gives satisfying results, *i.e.* the assumption that each matrix Δ_i has only two different eigenvalues is in many cases an efficient way to regularize the estimation of Δ_i. In this paper, we focus on the models $[a_{ij}b_iQ_id_i]$, $[a_{ij}bQ_id_i]$, $[a_ib_iQ_id_i]$, $[a_ibQ_id_i]$ and $[abQ_id_i]$.

3.2 The decision rule

Classification assigns an observation $x \in \mathbb{R}^p$ with unknown class membership to one of k classes $C_1, ..., C_k$ known *a priori*. The optimal decision rule, called *Bayes decision rule*, affects the observation x to the class which has the *maximum* posterior probability $P(x \in C_i|x) = \pi_i\phi(x,\theta_i)/\sum_{l=1}^{k}\pi_l\phi(x,\theta_l)$. Maximizing the posterior probability is equivalent to minimizing $-2\log(\pi_i\,\phi(x,\theta_i))$. For the model $[a_{ij}b_iQ_id_i]$, this results in the decision rule δ^+ which assigns x to the class minimizing the following cost function $K_i(x)$:

$$K_i(x) = \|\mu_i - P_i(x)\|^2_{\Lambda_i} + \frac{1}{b_i}\|x - P_i(x)\|^2 + \sum_{j=1}^{d_i}\log(a_{ij}) + (p-d_i)\log(b_i) - 2\log(\pi_i),$$

where $\|.\|_{\Lambda_i}$ is the Mahalanobis distance associated with the matrix $\Lambda_i = \tilde{Q}_i\Delta_i\tilde{Q}_i^{\,t}$. The posterior probability can therefore be rewritten as follows: $P(x \in C_i|x) = 1/\sum_{l=1}^{k}\exp\left(\frac{1}{2}(K_i(x) - K_l(x))\right)$. It measures the probability that x belongs to C_i and allows to identify dubiously classified points.

We can observe that this new decision rule is mainly based on two distances: the distance between the projection of x on $\mathbb{E}_i$ and the mean of the class; and the distance between the observation and the subspace $\mathbb{E}_i$. This rule assigns a new observation to the class for which it is close to the subspace and for which its projection on the class subspace is close to the mean of the class. If we consider the model $[a_ib_iQ_id_i]$, the variances a_i and b_i balance the importance of both distances. For example, if the data are very noisy, *i.e.* b_i is large, it is natural to balance the distance $\|x - P_i(x)\|^2$ by $1/b_i$ in order to take into account the large variance in $\mathbb{E}_i^{\perp}$.

Remark that the decision rule δ^+ of our models uses only the projection on $\mathbb{E}_i$ and we only have to estimate a d_i-dimensional subspace. Thus, our models are significantly more parsimonious than the general GMM. For example, if we consider 100-dimensional data, made of 4 classes and with common intrinsic dimensions d_i equal to 10, the model $[a_ib_iQ_id_i]$ requires the estimation of 4 015 parameters whereas the full Gaussian mixture model estimates 20 303 parameters.

4 High Dimensional Data Clustering

In this section we derive the EM-based clustering framework for the model $[a_{ij}b_iQ_id_i]$ and its sub-models. The new clustering approach is in the following referred to by High-Dimensional Data Clustering (HDDC). By lack of space, we do not present proofs of the following results which can be found in [2].

4.1 The clustering method HDDC

Unsupervised classification organizes data in homogeneous groups using only the observed values of the p explanatory variables. Usually, the parameters are estimated

by the EM algorithm which repeats iteratively E and M steps. If we use the parameterization presented in the previous section, the EM algorithm for estimating the parameters $\theta = \{\pi_i, \mu_i, \Sigma_i, a_{ij}, b_i, Q_i, d_i\}$, can be written as follows:
– E step: this step computes at the iteration q the conditional posterior probabilities $t_{ij}^{(q)} = P(x_j \in C_i^{(q)} | x_j)$ according to the relation:

$$t_{ij}^{(q)} = 1 / \sum_{l=1}^{k} \exp\left(\frac{1}{2}(K_i^{(q-1)}(x_j) - K_l^{(q-1)}(x_j))\right), \tag{1}$$

where K_i is defined in Paragraph 3.2.
– M step: this step maximizes at the iteration q the conditional likelihood. Proportions, means and covariance matrices of the mixture are estimated by:

$$\hat{\pi}_i^{(q)} = \frac{n_i^{(q)}}{n},\ \hat{\mu}_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ij}^{(q)} x_j,\ n_i^{(q)} = \sum_{j=1}^{n} t_{ij}^{(q)}. \tag{2}$$

$$\hat{\Sigma}_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ji}^{(q)} (x_j - \hat{\mu}_i^{(q)})(x_j - \hat{\mu}_i^{(q)})^t. \tag{3}$$

The estimation of HDDC parameters is detailed in the following subsection.

4.2 Estimation of HDDC parameters

Assuming for the moment that parameters d_i are known and omitting the index q of the iteration for the sake of simplicity, we obtain the following closed form estimators for the parameters of our models:
– Subspace $\mathbb{E}_i$: the d_i first columns of Q_i are estimated by the eigenvectors associated with the d_i largest eigenvalues λ_{ij} of $\hat{\Sigma}_i$.
– Model $[a_{ij} b_i Q_i d_i]$: the estimators of a_{ij} are the d_i largest eigenvalues λ_{ij} of $\hat{\Sigma}_i$ and the estimator of b_i is the mean of the $(p - d_i)$ smallest eigenvalues of $\hat{\Sigma}_i$ and can be written as follows:

$$\hat{b}_i = \frac{1}{(p - d_i)} \left(\mathrm{Tr}(\hat{\Sigma}_i) - \sum_{j=1}^{d_i} \lambda_{ij} \right). \tag{4}$$

– Model $[a_i b_i Q_i d_i]$: the estimator of b_i is given by (4) and the estimator of a_i is the mean of the d_i largest eigenvalues of $\hat{\Sigma}_i$:

$$\hat{a}_i = \frac{1}{d_i} \sum_{j=1}^{d_i} \lambda_{ij}, \tag{5}$$

– Model $[a_i b Q_i d_i]$: the estimator of a_i is given by (5) and the estimator of b is:

$$\hat{b} = \frac{1}{(np - \sum_{i=1}^{k} n_i d_i)} \left(n\,\mathrm{Tr}(\hat{W}) - \sum_{i=1}^{k} n_i \sum_{j=1}^{d_i} \lambda_{ij} \right), \tag{6}$$

where $\hat{W} = \sum_{i=1}^{k} \hat{\pi}_i \hat{\Sigma}_i$.

– Model $[abQ_id_i]$: the estimator of b is given by (6) and the estimator of a is:

$$\hat{a} = \frac{1}{\sum_{i=1}^{k} n_i d_i} \sum_{i=1}^{k} n_i \sum_{j=1}^{d_i} \lambda_{ij}. \tag{7}$$

4.3 Intrinsic dimension estimation

We also have to estimate the intrinsic dimensions of each subclass. This is a difficult problem with no unique technique to use. Our approach is based on the eigenvalues of the class conditional covariance matrix $\hat{\Sigma}_i$ of the class C_i. The jth eigenvalue of $\hat{\Sigma}_i$ corresponds to the fraction of the full variance carried by the jth eigenvector of $\hat{\Sigma}_i$. We estimate the class specific dimension d_i, $i = 1, ..., k$, with the empirical method scree-test of Cattell [3] which analyzes the differences between eigenvalues in order to find a break in the scree. The selected dimension is the one for which the subsequent differences are smaller than a threshold. In our experiments, the threshold is chosen by cross-validation. We also compared to the probabilistic criterion BIC [9] which gave very similar results.

5 Experimental results

In this section, we use our clustering method HDDC to recognize and locate objects in natural images. Object category recognition is one of the most challenging problems in computer vision. Recent methods use local image descriptors which are robust to occlusions, clutters and geometric transformations. Many of these approaches form clusters of local descriptors as an initial step; in most cases clustering is achieved with k-means, diagonal or spherical GMM and EM estimation – with or without PCA to reduce the dimension. Dorko and Schmid [6] select discriminant clusters based on the likelihood ratio and use the most discriminative ones for recognition. Bag-of-keypoint methods [11] represent an image by a histogram of cluster labels and learn a Support Vector Machine classifier.

5.1 Protocol and data

We use an approach similar to Dorko and Schmid [6]. Local descriptors of dimension 128 are extracted from the training images (see [6] for details) and then are organized into k groups by a clustering method ($k = 200$ in our experiments). We then compute the discriminative capacity of the class C_i for a given object category O through the posterior probability $R_i = P(C_i \in O|C_i)$. This probability is estimated by $R_i = \left[(\Psi^t \Psi)^{-1} \Psi^t \Phi \right]_i$, where $\Phi_j = P(x_j \in O|x_j)$ and $\Psi_{jl} = P(x_j \in C_l|x_j)$. Learning can be either supervised or weakly supervised. In the supervised framework, the objects are segmented using bounding boxes and only the descriptors located inside the bounding boxes are labeled as positive in the learning step. In the

	HDDC $[**Q_id_i]$				GMM			Pascal
Learning	$[a_{ij}b_i]$	$[a_{ij}b]$	$[a_ib_i]$	$[a_ib]$	PCA+diag.	Diagonal	Spherical	Best of [4]
Supervised	0.172	0.181	**0.183**	0.175	0.177	0.161	0.150	0.112
Weakly-sup.	0.145	0.147	0.142	**0.148**	0.120	0.110	0.106	/

Table 1. Object localization on the database *Pascal test2*: mean of the average precision on the four object categories. Best results are highlighted.

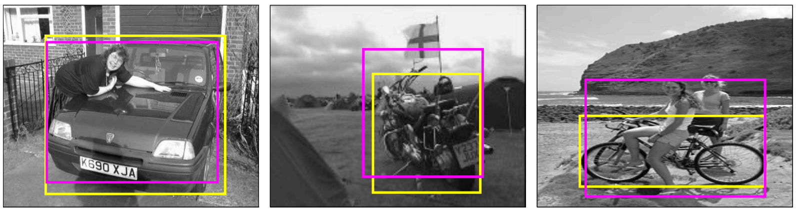
weakly-supervised scenario, the object are not segmented and all descriptors from images containing the object are labeled as positive. Note that in this case many descriptors from the background are labeled as positive. In both cases, we consider that $P(x_j \in O|x_j) = 1$ if x_j is positive and $P(x_j \in O|x_j) = 0$ otherwise. For each descriptor of a test image, the probability that this point belongs to the object O is then given by $P(x_j \in O|x_j) = \sum_{i=1}^{k} R_i P(x_j \in C_i|x_j)$ where the posterior probability $P(x_j \in C_i|x_j)$ is obtained by the decision rule associated to the clustering method (see Paragraph 3.2 for HDDC).

We compare the HDDC clustering method to the following classical clustering methods: diagonal Gaussian mixture model, spherical Gaussian mixture model, and data reduction with PCA combined with a diagonal Gaussian mixture model. The diagonal GMM has a covariance matrix defined by $\Sigma_i = \mathrm{diag}(\sigma_{i1}, ..., \sigma_{ip})$ and the spherical GMM is characterized by $\Sigma_i = \sigma_i Id$. For all the models the parameters were estimated via the EM algorithm. The EM estimation used the same initialization based on k-means for both HDDC and classical methods.

The object category database used in our experiments is the *Pascal* dataset [4] which contains four categories: motorbikes, bicycles, people and cars. There are 684 training images and two test sets: *test1* and *test2*. We evaluate our method on the set *test2*, which is the most difficult of the two test sets and contains 956 images. There are on average 250 descriptors per image. From a computational point of view, the localization step is very fast. For the learning step, computing time mainly depends of the number of groups k and is equal on average to 2 hours on a recent computer. To locate an object in a test image, we compute for each descriptor the probability to belong to the object. We then predict the bounding box based on the arithmetic mean and the standard deviation of descriptors. In order to compare our results with those of the Pascal Challenge [4], we used its evaluation criterion "average precision" which is the area under the precision-recall curve computed for the predicted bounding boxes (see [4] for further details).

5.2 Object localization results

Table 1 presents localization results for the dataset *Pascal test2* with supervised and weakly-supervised training. First of all, we observe that HDDC performs better than standard GMM within the probabilistic framework described in Section 5.1 and particularly in the weakly-supervised framework. This indicates that our clustering method identifies relevant clusters for each object category. In addition, HDDC



(a) car (b) motorbike (c) bicycle

Fig. 2. Object localization on on the database *Pascal test2*: predicted bounding boxes with HDDC are in red and true bounding boxes are in yellow.

provides better localization results than the state of the art methods reported in the Pascal Challenge [4]. Note that the difference between the results obtained in the supervised and in the weakly-supervised framework is not very high. This means that HDDC efficiently identifies discriminative clusters of each object category even with weak supervision. Weakly-supervised results are promising as they avoid time consuming manual annotation. Figure 2 shows examples of object localization on test images with the model $[a_i b_i Q_i d_i]$ of HDDC and supervised training.

Acknowledgments

This work was supported by the French department of research through the *ACI Masse de données* (Movistar project).

References

1. Bocci, L., Vicari, D., Vichi, M.: A mixture model for the classification of three-way proximity data. Computational Statistics and Data Analysis, **50**, 1625–1654 (2006).
2. Bouveyron, C., Girard, S., Schmid, C.: High-Dimensional Data Clustering. Technical Report **1083M**, LMC-IMAG, Université J. Fourier Grenoble 1 (2006).
3. Cattell, R.: The scree test for the number of factors. Multivariate Behavioral Research, **1**, 245–276 (1966).
4. D'Alche Buc, F., Dagan, I., Quinonero, J.: The 2005 Pascal visual object classes challenge. Proceedings of the first PASCAL Challenges Workshop, Springer (2006).
5. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, **39**, 1–38 (1977).
6. Dorko, G., Schmid, C.: Object class recognition using discriminative local features. Technical Report **5497**, INRIA (2004).
7. Fraley, C., Raftery, A.: Model-based clustering, discriminant analysis and density estimation. Journal of American Statistical Association, **97**, 611–631 (2002).
8. Parsons, L., Haque, E., Liu, H.: Subspace clustering for high dimensional data: a review. SIGKDD Explor. Newsl. **6**, 90–105 (2004).
9. Schwarz, G.: Estimating the dimension of a model. Annals of Statistics, **6**, 461–464 (1978).
10. Tipping, M., Bishop, C.: Mixtures of probabilistic principal component analysers. Neural Computation, 443–482 (1999).
11. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories. Technical Report **5737**, INRIA (2005).