Mixed dimensions estimation and clustering in high dimensional noisy point clouds

G. Haro, G. Randall and G. Sapiro

A framework for the regularized and robust estimation of non-uniform dimensionality and density in high dimensional noisy data is introduced in this work. This leads to learning stratifications, that is, mixture of manifolds representing different characteristics and complexities in the data set. The basic idea relies on modeling the high dimensional sample points as a process of Translated Poisson mixtures, with regularizing restrictions, leading to a model which incorporates the modeling of noise.

High dimensional data often 'hides' an intrinsic structure or an embedded manifold which allows to represent the original data points in a lower dimensional space. This so called dimensionality reduction then needs an estimate of the intrinsic dimension. However, in the presence of noise, the intrinsic dimension is difficult to estimate and many popular geometric dimension estimators give a dimension significantly larger than the actual one (this is related to the scale of the estimation). Clearly, noise brings points outside of the underlying manifold and into the higher dimensional embedding space. In these cases, spectral dimension estimators may be preferred because they give more information about the importance of each dimension and the 'noisy' dimensions could be inferred from that information. Our proposed framework naturally incorporates the notion of noise in the estimation.

Levina and Bickel, [3], proposed a geometric and probabilistic method which estimates the local dimension and density of a point cloud data. This dimension estimator is equivalent to the one proposed in [5] in the context of dynamical systems. In this work we show how the presence of noise can be naturally incorporated in this model in order to obtain a dimension estimator which is robust to noise. Then, we extend this result to the more general case where the point cloud data is a sampling of two or more manifolds of different dimensions and densities (a stratification), and cluster the noisy data points according to these parameters. In our work, we do not assume linear subspaces, and we simultaneously estimate the soft clustering and the intrinsic dimension and density of the clusters while being robust to noise and outliers.

The approach in [3] is based on the idea that if we sample an m-dimensional manifold with T points, the proportion of points that fall into a ball around a point x_t is $\frac{k}{T} \approx \rho(x_t)V(m)R_k(x_t)^m$. The given point cloud, embedded in high dimensions D, is $X = \{x_t \in \mathbb{R}^D; t = 1, \ldots, T\}$, k is the number of points inside the ball, $\rho(x_t)$ is the local sampling density at point x_t, $V(m)$ is the volume of the unit sphere in $\mathbb{R}^m$, and $R_k(x_t)$ is the Euclidean distance from x_t to its k-th nearest neighbor (kNN). The inhomogeneous process $N(R, x_t)$, which counts the number of points falling into a small D-dimensional sphere $B(R, x_t)$ of radius R centered at x_t, is a binomial process. Under certain assumptions it can approximated by a Poisson process and the rate λ of the counting process $N(R, x_t)$ can be expressed as $\lambda(R, x_t) = \rho(x_t)V(m)mR^{m-1}$. The local intrinsic dimension estimator at each point x_t is obtained from the Maximum Likelihood estimator based on a Poisson distribution with this rate.

Usually, point samples are contaminated with noise, thus the point process that we observe is not a simple sampling of a low dimensional manifold but a perturbation of this sample process. This can be modeled, as we perform in our proposed framework, with a Translated Poisson Process [4], where an underlying (unobservable) point process is translated to an output (observable) point process. The input and output spaces of the points are not necessarily the same or even of the same dimension. More concretely, an input point at location x in the input space X is randomly translated to a location z in the output space Z, according to a conditional probability density $f(z|x)$, called the *transition density*. We consider the particular case where each point is translated independently of the others and there are no deletions or insertions in the translation process. A critical observation, [4], is that a translated Poisson process with an integrable intensity function $\{\lambda(x)\colon x \in X\}$ is also a Poisson process with intensity $\mu(z) = \int_X f(z|x)\lambda(x)dx$.

Since the intensity of the Poisson process in our model is parametrized by the Euclidean distances of the points (and not by the points themselves), we consider a random translation in the distances. This means that we do not observe the original distances but noisy distances. Let $f(s|r)$ be the transition density which defines the random process which translates a distance r in the input space to a distance s in the observable space. If $\lambda(r, x_t)$ is the local rate of the Poisson process which defines the counting process in the input space, then $\mu(s)$, the intensity of the Poisson process in the output space, is given by $\mu(s, x_t) = \int_0^{R'} f(s|r)\rho(x_t)V(m)mr^{m-1}dr$.

We consider $R' > R$ since points originally at distance greater than R from x_t can be placed within a distance less than R after the translation process.

Maximizing the Likelihood of the new Translated Poisson process, we obtain the following expression for the local dimension $m(x_t)$ at point x_t when we use the k nearest neighbors (k-NN) instead of the points within distance less to R,

$$m(x_t) = \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \frac{\int_0^{R'} f(R_i(x_t)|r) r^{m-1} \log \frac{R_k(x_t)}{r} dr}{\int_0^{R'} f(R_i(x_t)|r) r^{m-1} dr} \right]^{-1}, \tag{1}$$

where, by an abuse of notation, we have identified $m = m(x_t)$ in the right hand side. Note that this expression reduces to the Levina and Bickel estimator [3] in the particular case that $f(s|r) = \delta(s-r)$, i.e., there is no translation of the original points. This corresponds to the ideal case with no noise. Equation (1) is a nonlinear recursive expression in m which is difficult to solve. We approximate it by an easier to compute closed expression. Since the translation density is modeling the effect of noise, the effective support of $f(s|r)$ is going to be concentrated around s. Then, we can substitute r^{m-1} in (1) by its Taylor expansion around R_i. Thus, we obtain

$$m(x_t) \approx \left[\frac{1}{k-1} \sum_{i=1}^{k-1} \frac{\int_0^{R'} f(R_i|r) \log \frac{R_k}{r} dr}{\int_0^{R'} f(R_i|r) dr} \right]^{-1}. \tag{2}$$

With this approximation, which is fully studied in our work, the estimator (2) still reduces to the estimator in [3] when $f(R_i|r) = \delta(R_i - r)$.

These estimators are local since they come from a Maximum Likelihood (ML) in each point x_t based on a Translated Poisson distribution modeling the counting process in the local ball $B(R, x_t)$. We propose to compute a ML on the whole point cloud data at the same time (and not one for each point independently), based on a Translated Poisson Mixture Model, which models the presence of noise and permits to have different classes (each one with their own dimension and sampling density). This technique automatically gives a soft clustering according to dimensionality and density, with an estimation of both quantities for each class. A preliminary version of this work was presented in [1] and a regularized version together with asymptotic results in [2]. These techniques are particular cases of the more general Translated Poisson model introduced in this work in order to handle noise.

The Translated Poisson Mixture Model (TPMM) is solved with an EM algorithm and the dimension estimator m^j of class j is

$$m^j = \left[\frac{\sum_{t=1}^{T} h^j(x_t) m(x_t)^{-1}}{\sum_{t=1}^{T} h^j(x_t)} \right]^{-1} \tag{3}$$

where $h^j(x_t)$ is the probability that point x_t belongs to class j. Then, (3) is a weighted harmonic mean of local dimension estimators of points belonging to class j.

The TPMM algorithm and its regularized versions have been applied to synthetic noisy point clouds and have proved to give a more accurate estimation of the different dimensions, compared to the Poisson Mixture Model of [1] or its regularized version in [2], while obtaining a good clustering according to dimensionality and density. We have also tested these algorithms clustering real data in computer vision applications (scanned digits, faces under varying pose and illumination, different activities and motion in video), again improving the clustering and the dimension estimation with the translated version. These results and the underlying theory will be presented at the workshop.

References

[1] G. Haro, G. Randall, and G. Sapiro. Stratification learning: Detecting mixed density and dimensionality in high dimensional point clouds. In *Advances in NIPS 19, Vancouver, Canada*, 2006.

[2] G. Haro, G. Randall, and G. Sapiro. Regularized mixed dimensionality and density learning in computer vision. In *Proceedings of 1st Workshop on Component Analysis Methods for Classification, Clustering, Modeling and Estimation Problems in Computer Vision, in conjunction with CVPR*, Minneapolis, June 2007.

[3] E. Levina and P.J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in NIPS 17, Vancouver, Canada*, 2005.

[4] D. L. Snyder and M. I. Miller. *Random Point Processes in Time and Space*. Springer-Verlag, 1991.

[5] F. Takens. On the numerical determination of the dimension of an attractor. *Lecture notes in mathematics. Dynamical systems and bifurcations*, 1125:99–106, 1985.