Manifold learning is based on the hypothesis that data in high dimensional Euclidean spaces usually lie in the vicinity of a low dimensional submanifold.

Can this hypothesis be tested using limited data?
Low dimensional manifolds with bounded volume and curvature

Let $\mathcal{G}_e = \mathcal{G}_e(k, V, \tau)$ be the family of Riemannian $k$–submanifolds of the unit ball in $\mathbb{R}^n$, with volume $\leq V$ and curvature $\leq \kappa$. 

Low curvature

high curvature
A positive result

Let $\mathcal{P}$ be a probability distribution supported in the unit ball from which data $x_1, \ldots, x_s$ is drawn i.i.d. If $s$ is greater than

$$C \left( \min \left( \left( \frac{1}{\epsilon^2} \right) \log^4 \left( \frac{N_p}{\epsilon} \right), N_p \right) \frac{N_p}{\epsilon^2} + \frac{1}{\epsilon^2} \log \frac{1}{\delta} \right)$$

where $N_p$ is

$$V \left( C \left( k \max \left( \frac{1}{\epsilon}, \kappa \right) \right) \right)^k$$

Then, independent of ambient dimension $n$,

$$\mathbb{P} \left[ \sup_{\mathcal{G}} \left| \frac{1}{s} \sum_{i=1}^{s} d(\mathcal{M}, x_i)^2 - \int d(\mathcal{M}, x)^2 d\mathcal{P}(x) \right| < \epsilon \right] > 1 - \delta$$
K-means

In particular, this improves the best known upper bound on the sample complexity of k-means from $O\left(\frac{k^2 + \log \frac{1}{\delta}}{\epsilon^2}\right)$ to

$$O\left( k \min \left( k, \frac{\log^4 \frac{k}{\epsilon}}{\epsilon^2} \right) + \log \frac{1}{\delta} \right)$$

Aspects of the proof:
1. Estimates for the volumes of balls in Riemannian manifolds
2. Bounding the Fat-Shattering dimension using Random Projections onto a low-dimensional subspace.