Robust Faces Manifold Modeling: Most Expressive Vs. Most Sparse Criterion

Presented by
Xiaoyang Tan
x.tan@nuaa.edu.cn ;
http://parnec.nuaa.edu.cn/xtan

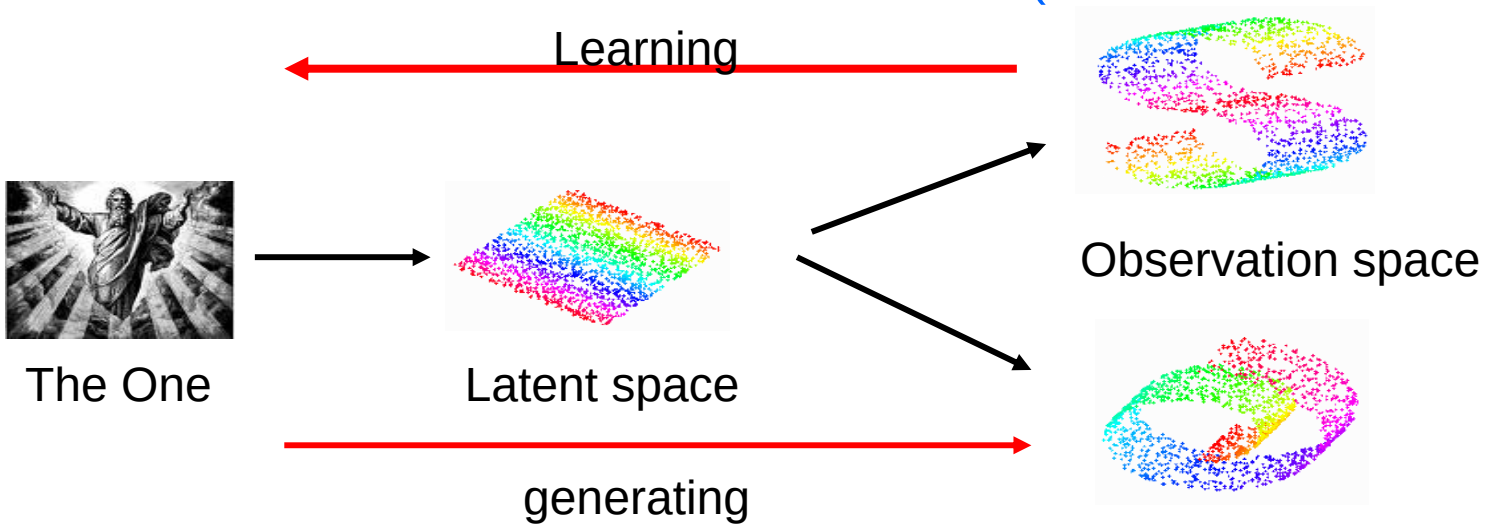
南京航空航天大学
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

In collaboration with
Lishan Qiao , Wenjuan Gao & Jun Liu
Sept., 2009

ParNeC

- High dimensional data everywhere everyday
  - Images, documents, web pages, genes…
  - Complicated but useful
- The problem: to learn some meaningful structure/model from them (or, manifold learning).
  - For visualization, classification, regression…
  - But hidden among a couple of sparsely and irregularly sampled data, possibly polluted by noise, outlier, etc.
  - Is that possible? – largely positive.

- A generative point of view
  - Two-step generation process
    - Generated by a simple low-dimensional process
      - Few degree of freedom & little noise
    - Observed by a complex observation process
  - no structure in high-dim by itself, Low-dim process lends it to the high-dim data.
    - Ill posed inverse problem (few HD data ->general struc.)
    - structure latent –we will never see it (but can re-embed it)



- Classical solutions for manifold learning

- Two approaches
  - Deterministic methods – nonparametric
  - Probabilistic methods – parametric

  - Share common characteristics, but make different assumptions
    - impose different bias on the manifold to be learned
    - work well only when such conditions are satisfied.

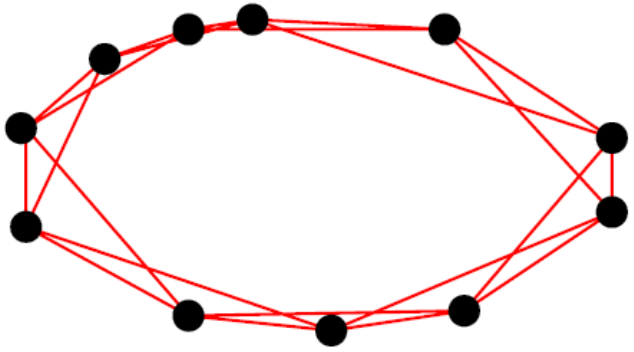
- Deterministic (geometric) approach
  - Manifold should be densely & regularly sampled
  - Data located on the manifold - no noise
  - Flexible – can be any shape of distribution
    - ISOMAP (Tenenbaum, et al, 00)
    - LLE (Roweis, Saul, 00)
    - Laplacian Eigenmaps (Belkin, Niyogi, 01)
    - Local Tangent Space Alignment (Zhang, Zha, 02)
    - Hessian Eigenmaps (Donoho, Grimes, 02)
    - Diffusion Maps (Coifman, Lafon, et al, 04)

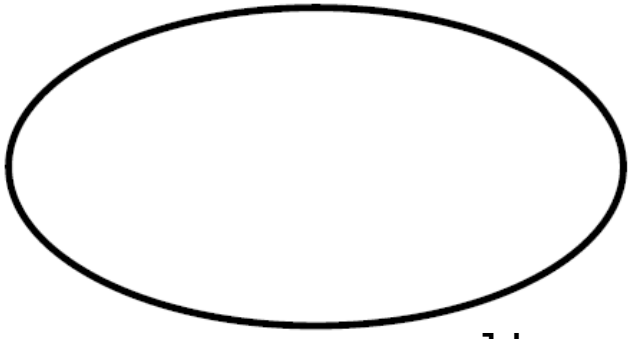
*ISOMAP & Hessian Eigenmaps has strong asymptotic guarantees.

- Probabilistic approach
  - Manifold can be sparsely / irregularly sampled
  - Data is “close to” the manifold - noisy
  - types of shape of distribution - Limited
    - Local PCA
    - Principal curve & surface (Hastie 89)
    - Manifold chart (Brand 03)
    - Global coordination(Roweis,Saul 02,Teh, Saul 02,Verbeek 02,06)
    - GPLVM (Lawrence 03,05)
    - …

- How they works? Follow a common proc.
- Step 1, observation model – our bias about the structure to be learned
  - Neighborhood graph (Deterministic methods)
  - Finite Mixture Model/GP/DP/BP (Probabilistic methods)
- Step 2, embedding
  - Spectral techniques (Deterministic methods)
  - Different alignment methods for Local coordinate systems (Probabilistic methods)

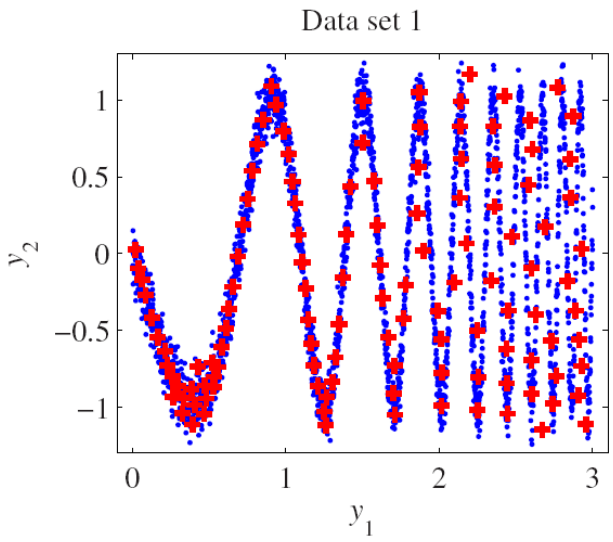
- Neighborhood graph
  - Connecting neighboring points in the space
  - How you connect them ~ what kind of topology to be preserved.
  - To be or not to be a neighborhood – user's needs and preferences (not necessarily Eucldiean distances). –> k-rule or epsilon-rule

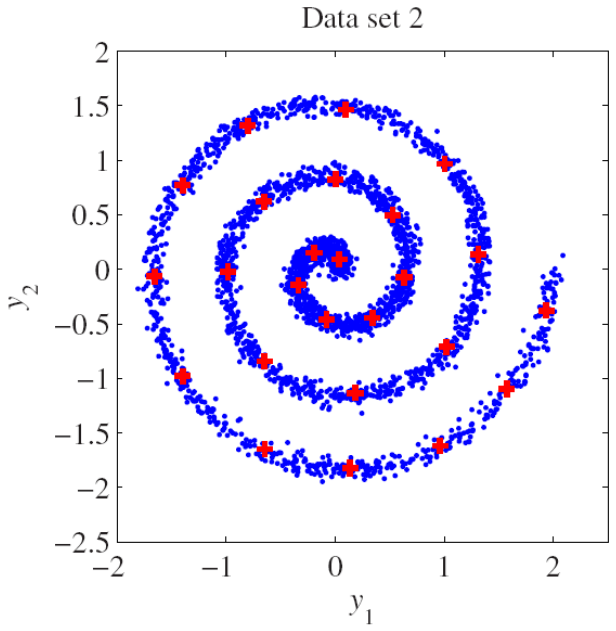




J.Lee et al. NDR

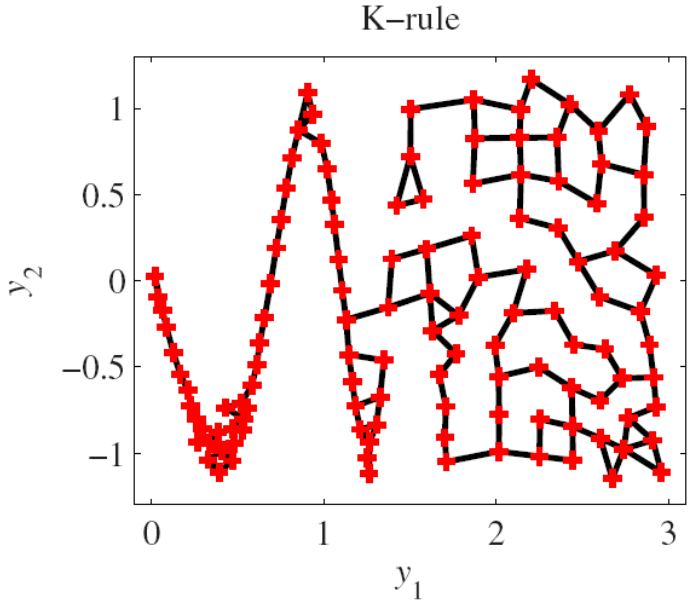
- K-ary

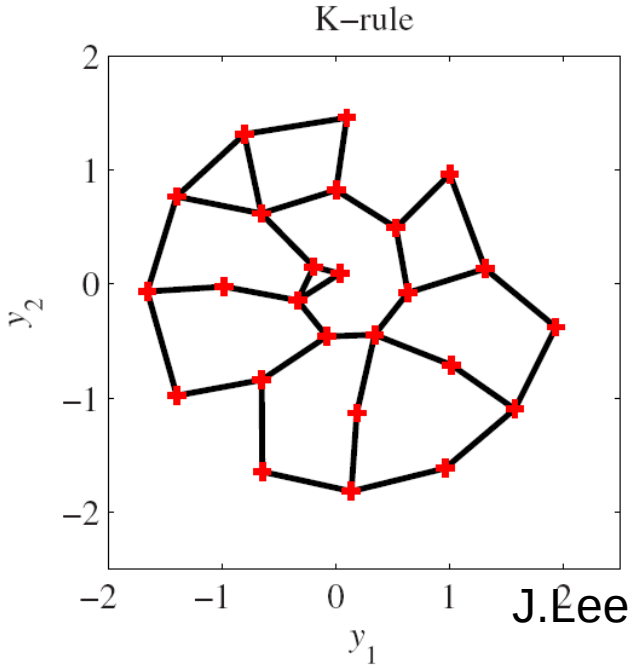




•Uncontrollable patch size (outlier sensitive)

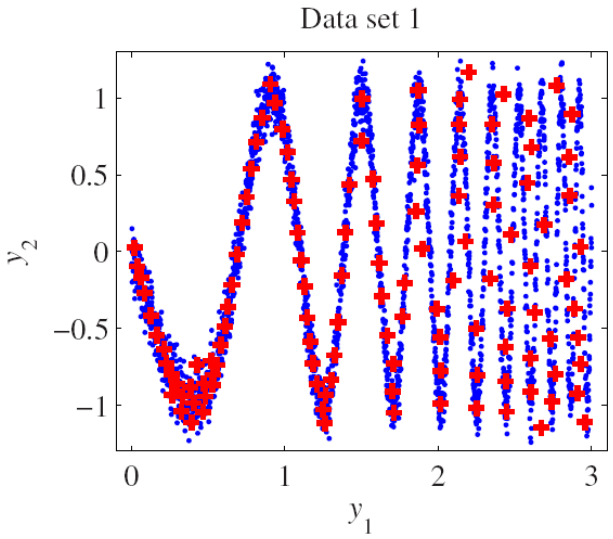
•Easy in practice

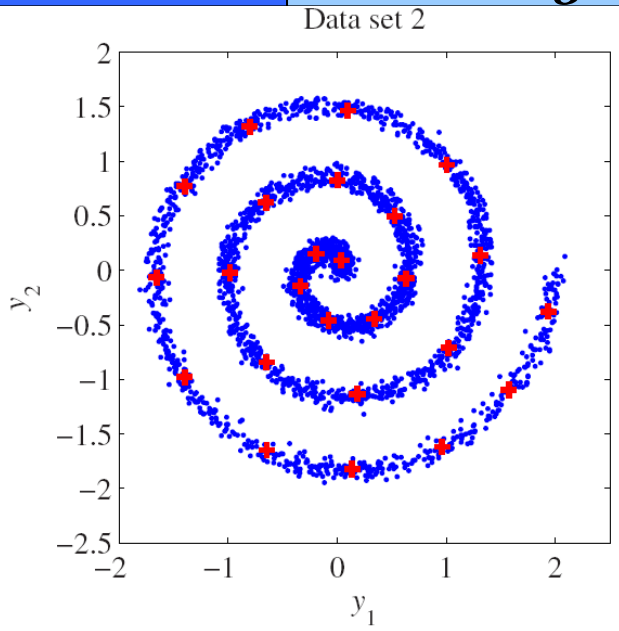




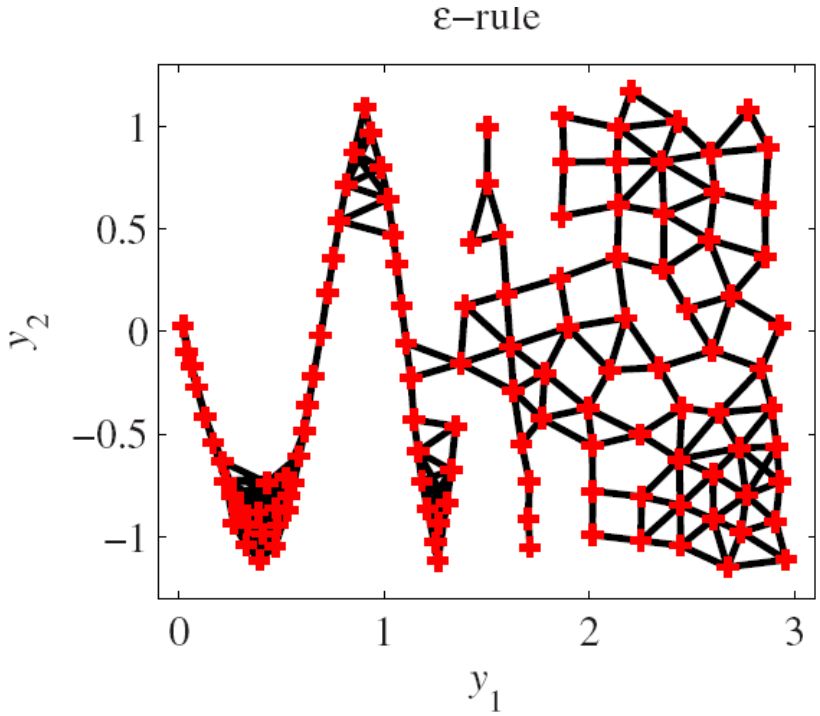
J.Lee et al. NDR

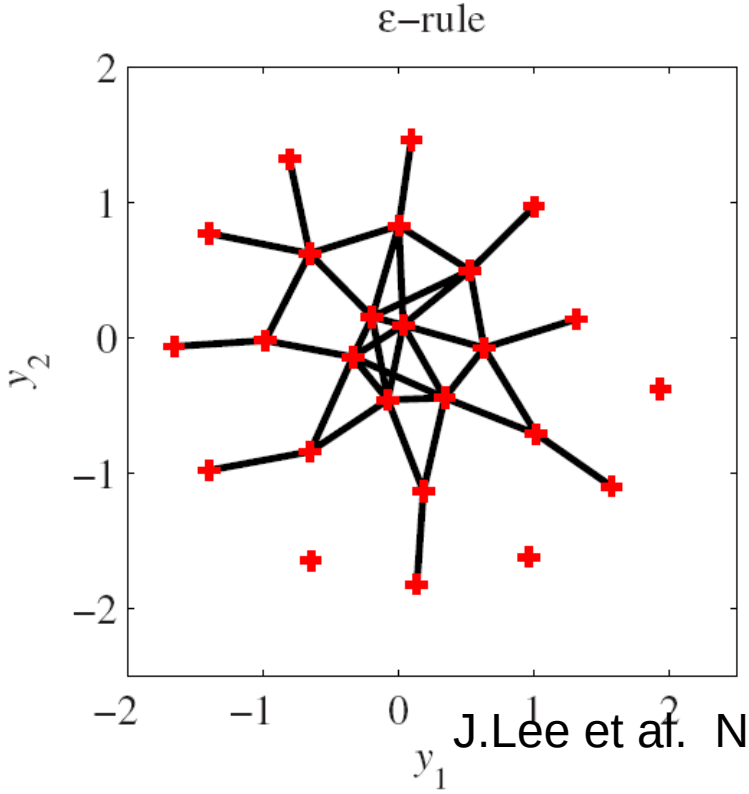
$\varepsilon-rule$





- Difficult to set
- Good only in uniform distr. regions.





J.Lee et al. NDR

- Drawbacks of K-rule and epsilon-rule
  - Lack of flexibility
    - a fixed global parameter to determine the neighborhoods for all the data
  - Lack of discriminative power
    - Tends to put samples from different classes into the same patch, thus enlarging the within-class variations.
- How to overcome these shortcomings?

- Neighborhood graph as sparse linear model
  - For each point *y* in the observed space, seek a linear model with two requirements:
    - *y is approximated by a linear model y' =Aw as accurately as possible, A is the dictionary*.
    - *w is sparse (many zero entries) and robust to noise (bad fit for noise)*
  - This leads to the l0-norm minimization problem

$$(l^0) : \hat{w}_{l0} = \arg\min_{w} ||y - Aw||_2^2 + \lambda ||w||_0$$

  - By solving it, we get both neighbors and weights
    - adaptively and robustly, with sparsity property exploited.

- Neighborhood graph as sparse linear model

$$(l^0) : \hat{w}_{l0} = \arg\min_w \|y - Aw\|_2^2 + \lambda\|w\|_0$$

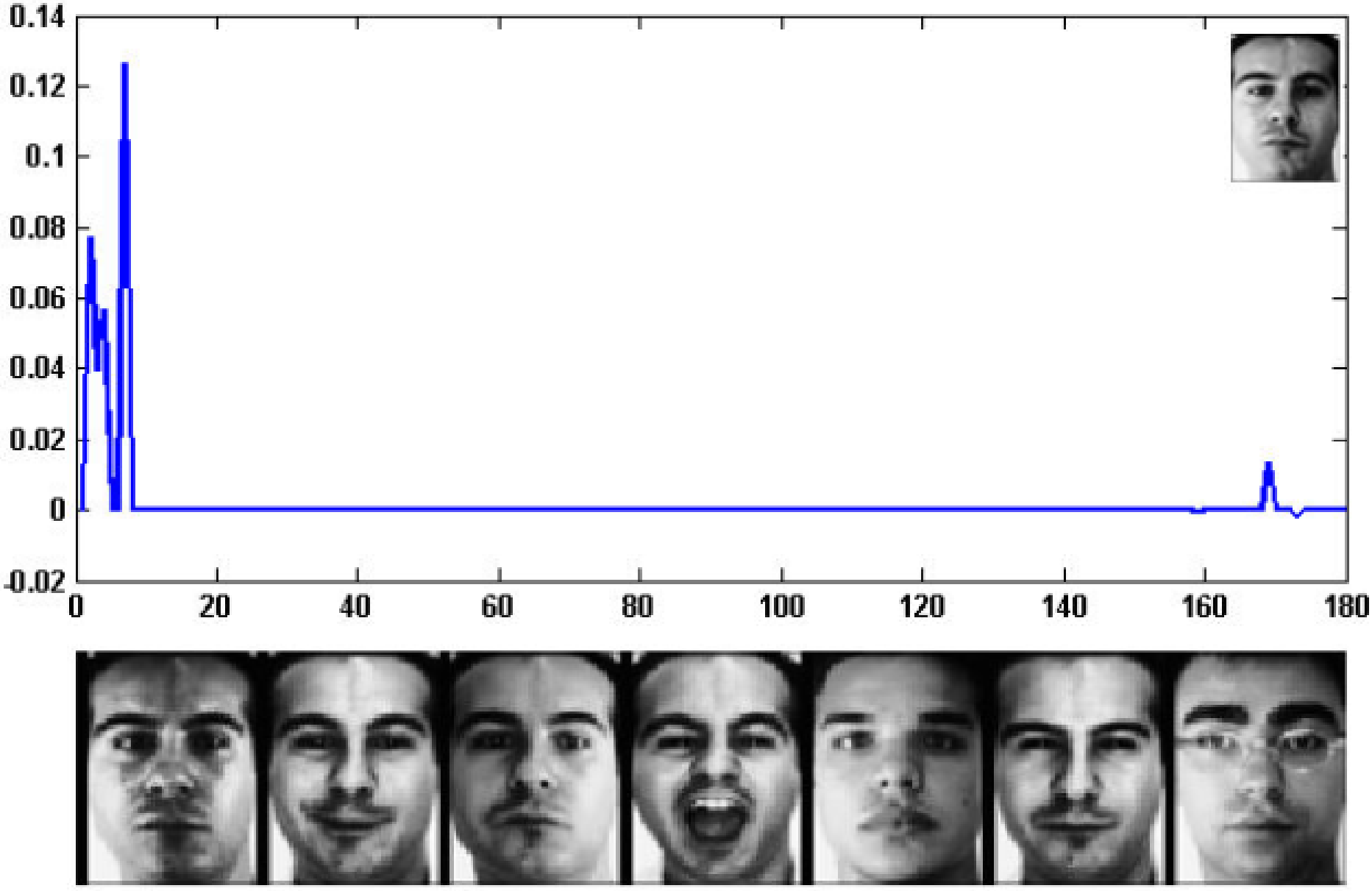
- Under certain conditions, l0-norm is equivalent to l1-norm (LASSO)

$$(l^1) : \hat{w}_{l_1} = \arg\min_w \|y - Aw\|_2^2 + \lambda\|w\|_1$$

In this paper, solved it by a slightly different but essentially equivalent form

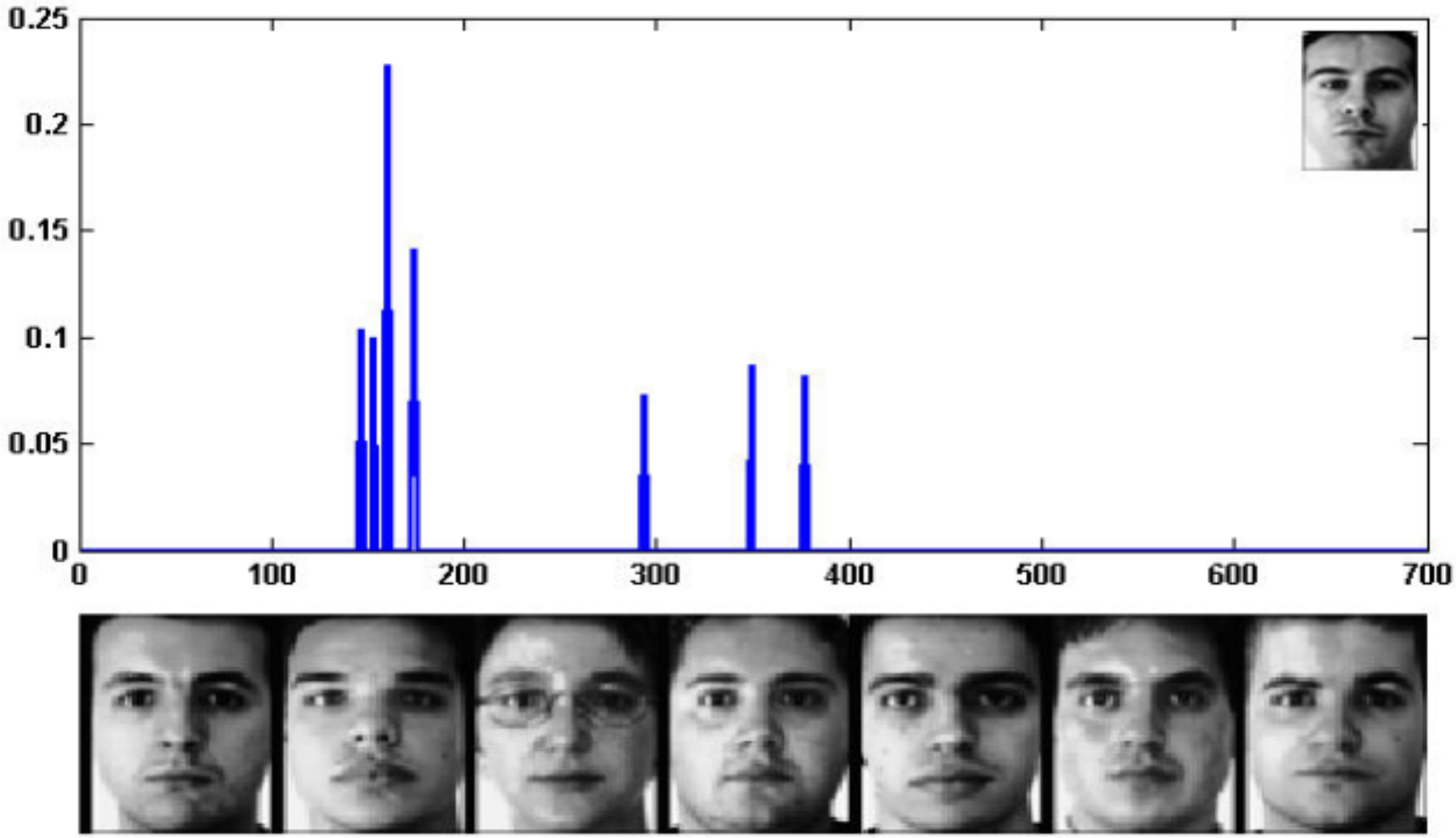
$$(l^{1**}) : \hat{w}_{l_1} = \arg\min_w \|y' - \Theta w\|_2^2$$
$$s.t. \|w\|_1 \leq \tau$$

The Proposed Method | L1-norm Neighborhood Graph



L1-norm neighborhood and its weights

Sparse, Adaptive, Discriminative, Outlier-insensitive



conventional K neighborhood and its weights

Put samples from different class into one patch

- Two practical issues:
- 1. Can we always obtain a sparse solution using LASSO?
- 2. Can we always recovery the real sparse pattern that supposed to preserved?

NO! – largely depends on your design matrix A

- Can we always obtain a sparse solution?
- RIP (restricted isometry property)
  - Hard to check but always ok for a Gaussian random matrix (Candes 06, Baraniuk 07)
  - If not fulfilled, multiply the design matrix by a Gaussian random matrix

$$(l^{1*}) : \hat{w}_{l_1} = \arg\min_{w} \|\Phi y - \Phi A w\|_2^2 + \lambda \|w\|_1$$

- Can we always recovery the real sparse pattern?
  - sign(w')= (+1, -1, 0, 0, +1, -1, -1, 0, 0,…)
  - If the prob. of recovered sparse pattern equals true sparse pattern converges to 1 -> sign consistency

- neighborhood stability (Meinshausen 06) or Irrepresentable condition (Zhao & Yu 06)
  - Hard to check.
  - If not fulfilled, use a two step procedures (Meinshausen & Yu 07)
    - Step 1, run LASSO with a smaller lambda
    - Step 2, remove small absolute coefficient

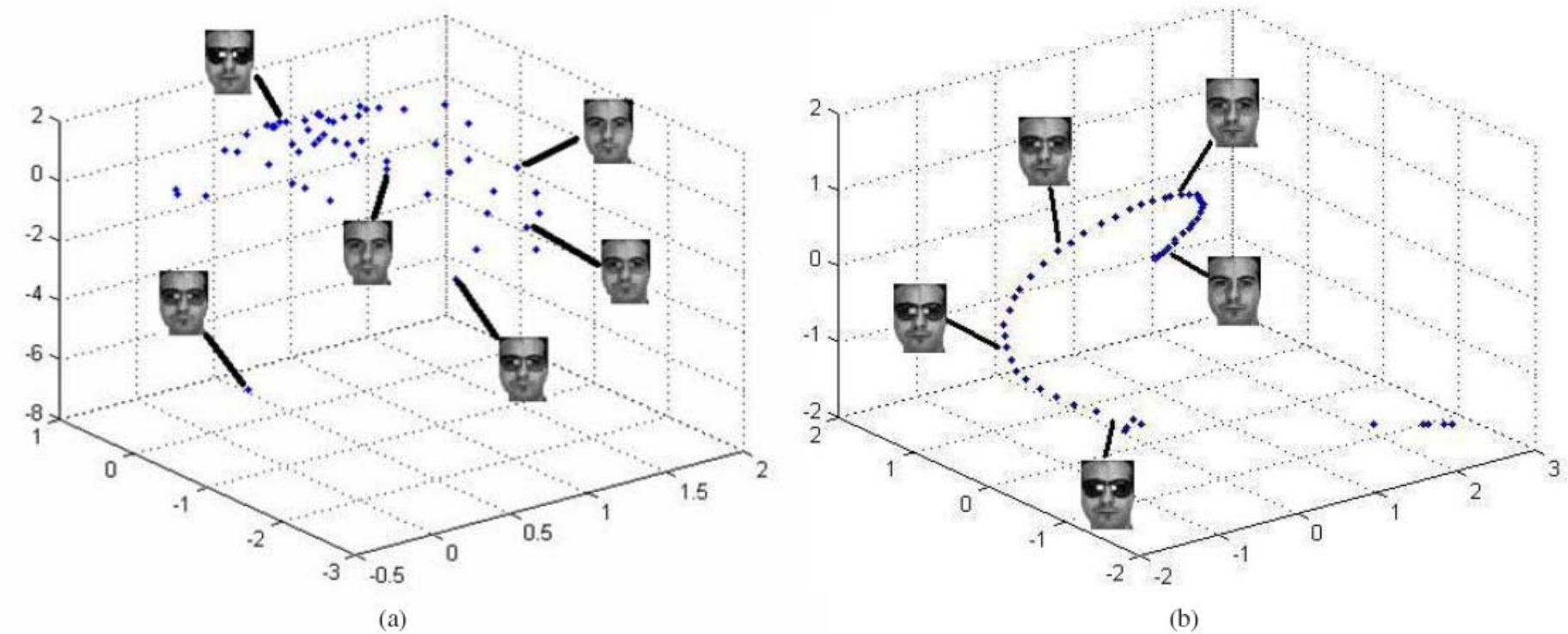
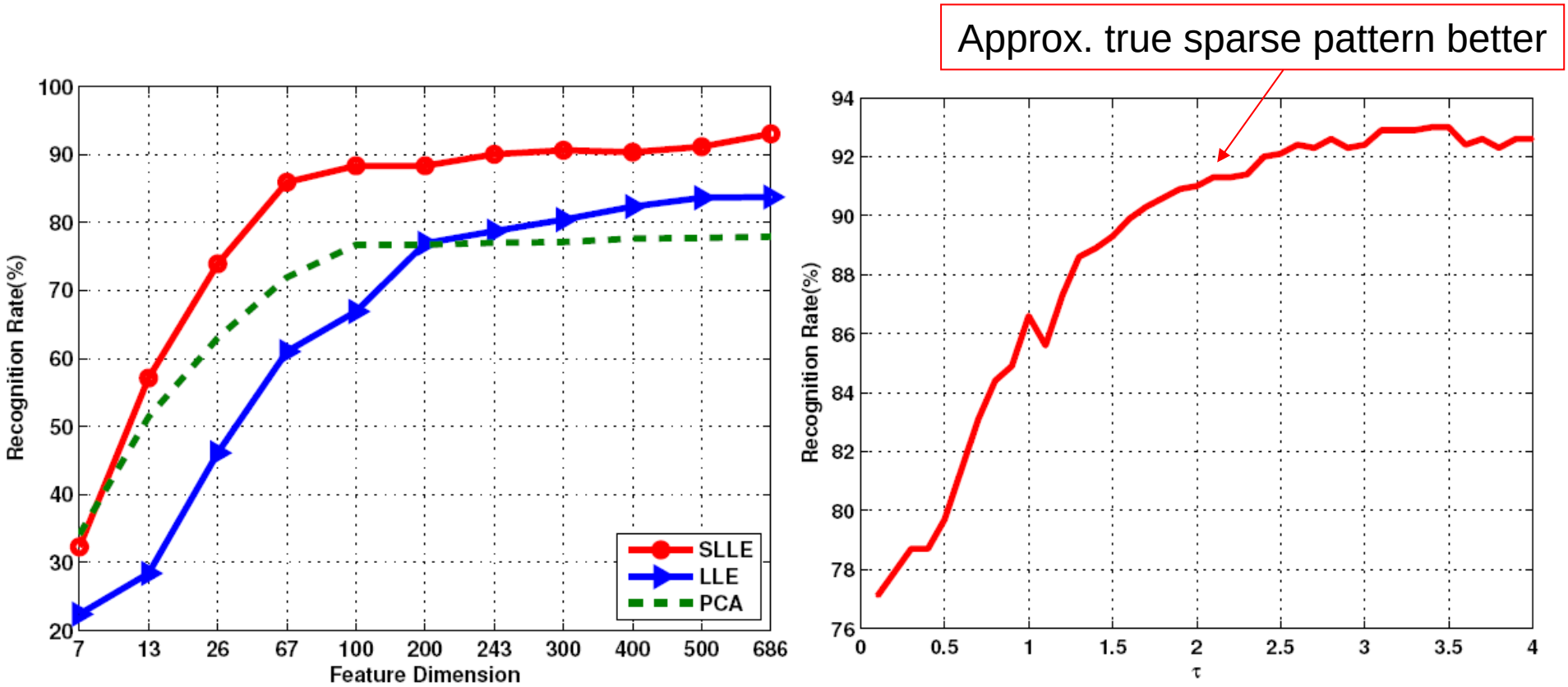


Figure 3. The face manifold respectively modeled with LLE and SLLE on AR with partial occlusions when there exists noise: (a) LLE and (b) SLLE.

Dealing with outliers

Figure 5. Some AR samples from one subject. Left: training images; Right: test images.



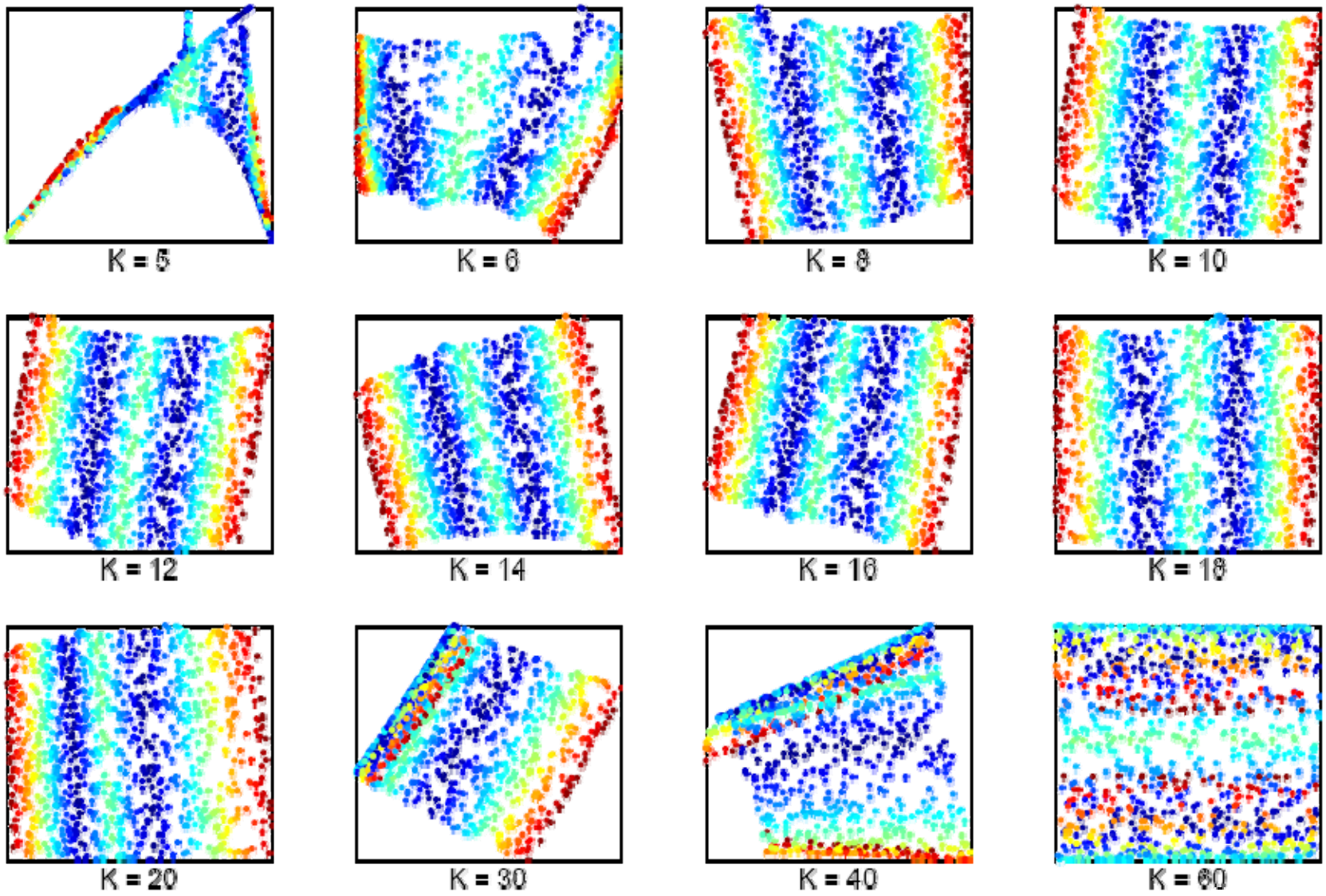
Face Recognition on the manifold

Conclusions and Future work

- l1-norm neighborhood graph
  - Sparsity, adaptive, discriminative, robust
  - Give some analysis on practical issues.

- weakness:
  1. Lambda/tau is hard to set as well, but generally smaller/larger one leads to better predication performance
  2. May need a few samples from the same class to encourage its discriminative capability
  3. Could be very slow for large data set (no obvious way to accelerate it, k-d tree is no use here)
  4. overall geometric property of neighborhood graph is unclear (it uses non-local information).

Thanks!



K-rule LLE – very sensitive to the value of k

- When new sample comes
  - Get its sparse representation in the observation space
  - Reconstruct itself in the low-dimensional space by fixing the weights.

- Where the discriminative power of sparse reconstruction comes from?
  - R. V. E. Elhamifar. Sparse subspace clustering. CVPR, 2009.

Theorem 1 *Let* $Y \in \mathbb{R}^{D\times N}$ *be a matrix whose columns are drawn from a union of* n *independent linear subspaces. Assume that the points within each subspace are in general position. Let* $\boldsymbol{y}$ *be a new point in subspace* i*. The solution to the* ℓ_1 *problem in* (9) $\boldsymbol{s} = \Gamma^{-1}[\boldsymbol{s}_1^\top, \boldsymbol{s}_2^\top, \cdots, \boldsymbol{s}_n^\top]^\top \in \mathbb{R}^N$ *is block sparse,* i.e. $\boldsymbol{s}_i \neq 0$ *and* $\boldsymbol{s}_j = 0$ *for all* $j \neq i$.

  - If the number of samples of each subjects are so large that the self-express condition is satisfied, any new test sample will only be recovered by the samples with the same identity.