Non-Local Manifold Parzen Windows

Yoshua Bengio, Hugo Larochelle and Pascal Vincent

Département d'informatique et de recherche opérationnelle
Université de Montréal

July 15^{th}, 2005

Plan

Plan

About this talk...

- **What :** density estimation of high dimensional continuous data, lying on a lower dimensional manifold

About this talk...

- **What :** density estimation of high dimensional continuous data, lying on a lower dimensional manifold
- **How :**
  - using the Manifold Parzen Windows model
  - learning the model's parameters with a neural network

About this talk...

- **What :** density estimation of high dimensional continuous data, lying on a lower dimensional manifold
- **How :**
  - using the Manifold Parzen Windows model
  - learning the model's parameters with a neural network
- **Why :**

About this talk...

- **What :** density estimation of high dimensional continuous data, lying on a lower dimensional manifold
- **How :**
  - using the Manifold Parzen Windows model
  - learning the model's parameters with a neural network
- **Why :**
  - ... because my supervisor wants me to work on that

About this talk...

- **What :** density estimation of high dimensional continuous data, lying on a lower dimensional manifold
- **How :**
  - using the Manifold Parzen Windows model
  - learning the model's parameters with a neural network
- **Why :**
  - ... because my supervisor wants me to work on that
  - ... to publish papers

About this talk...

- **What :** density estimation of high dimensional continuous data, lying on a lower dimensional manifold
- **How :**
  - using the Manifold Parzen Windows model
  - learning the model's parameters with a neural network
- **Why :**
  - ... because my supervisor wants me to work on that
  - ... to publish papers
  - but mostly to use and make a point about ***non-local learning***

Manifold Parzen Windows (Vincent and Bengio, 2003)

- Extension of the Parzen Windows model (mixture of spherical Gaussians, centered on the training points)
- The Gaussians are parametrized so that most of the density is situated on the underlying manifold

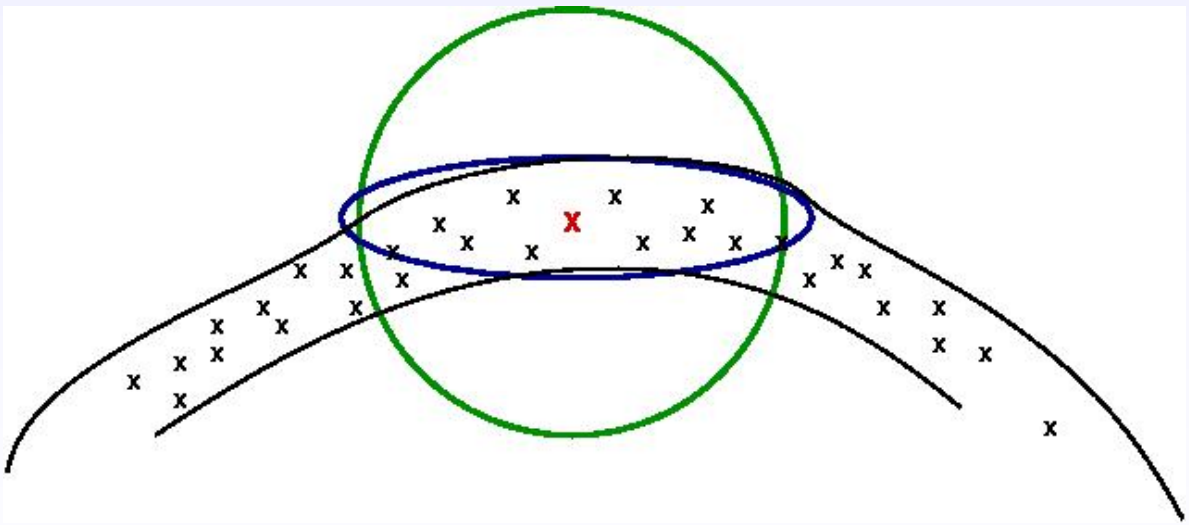


FIG.: Parzen Windows vs Manifold Parzen Windows

Manifold Parzen Windows (Vincent and Bengio, 2003)

- Density estimator :

$$p(x) = \frac{1}{n} \sum_{t=1}^{n} \mathcal{N}(x; \mu(x_t), \Sigma(x_t))$$

Manifold Parzen Windows (Vincent and Bengio, 2003)

- Density estimator :

$$p(x) = \frac{1}{n} \sum_{t=1}^{n} \mathcal{N}(x; \mu(x_t), \Sigma(x_t))$$

- Parametrization :

$$\Sigma(x_t) = \sigma^2_{noise}(x_t) I + \sum_{j=1}^{d} s_j(x_t) v_j(x_t) v_j(x_t)'$$

Manifold Parzen Windows (Vincent and Bengio, 2003)

- Density estimator :

$$p(x) = \frac{1}{n} \sum_{t=1}^{n} \mathcal{N}(x; \mu(x_t), \Sigma(x_t))$$

- Parametrization :

$$\Sigma(x_t) = \sigma^2_{noise}(x_t) I + \sum_{j=1}^{d} s_j(x_t) v_j(x_t) v_j(x_t)'$$

- Training :
  - $\mu(x_t) = x_t$ is fixed
  - for each x_t, use principal eigenvalues ($s_j(x_t)$) and eigenvectors ($v_j(x_t)$) of k nearest neighbors covariance matrix
  - $\sigma_{noise}(x_t)$ is an hyper-parameter

Non-Local Manifold Parzen Windows

- In Manifold Parzen Windows, $\mu(x_t)$, $\sigma_{noise}(x_t)$, $s_j(x_t)$ and $v_j(x_t)$ are stored in memory for every training point x_t

Non-Local Manifold Parzen Windows

- In Manifold Parzen Windows, $\mu(x_t)$, $\sigma_{noise}(x_t)$, $s_j(x_t)$ and $v_j(x_t)$ are stored in memory for every training point x_t
- In Non-Local Manifold Parzen Windows, $\mu(x_t)$,$\sigma_{noise}(x_t)$, $s_j(x_t)$ and $v_j(x_t)$ are functions of x_t, modeled by a neural network

Non-Local Manifold Parzen Windows

- In Manifold Parzen Windows, $\mu(x_t)$, $\sigma_{noise}(x_t)$, $s_j(x_t)$ and $v_j(x_t)$ are stored in memory for every training point x_t
- In Non-Local Manifold Parzen Windows, $\mu(x_t)$,$\sigma_{noise}(x_t)$, $s_j(x_t)$ and $v_j(x_t)$ are functions of x_t, modeled by a neural network
- The neural network can capture global information about the underlying manifold, and share it among all training points

Non-Local Manifold Parzen Windows

- In Manifold Parzen Windows, $\mu(x_t)$, $\sigma_{noise}(x_t)$, $s_j(x_t)$ and $v_j(x_t)$ are stored in memory for every training point x_t
- In Non-Local Manifold Parzen Windows, $\mu(x_t)$,$\sigma_{noise}(x_t)$, $s_j(x_t)$ and $v_j(x_t)$ are functions of x_t, modeled by a neural network
- The neural network can capture global information about the underlying manifold, and share it among all training points
- The neural network is trained using stochastic gradient descent on the average negative log-likelihood of the training set

Plan

Informal definitions

- What is *local learning* :
  - A learning algorithm is said to be local if it uses mostly nearby points of x to make a prediction at x
  - Examples : k nearest neighbors, SVM, most popular dimensionality reduction algorithms, Manifold Parzen Windows

Informal definitions

- What is *local learning* :
  - A learning algorithm is said to be local if it uses mostly nearby points of x to make a prediction at x
  - Examples : k nearest neighbors, SVM, most popular dimensionality reduction algorithms, Manifold Parzen Windows
- What is *non-local learning* :
  - A learning algorithm is said to be non-local if it is able to use information from training points far from x to generalize at x

Toy example

- We are trying to learn a density using this training set :

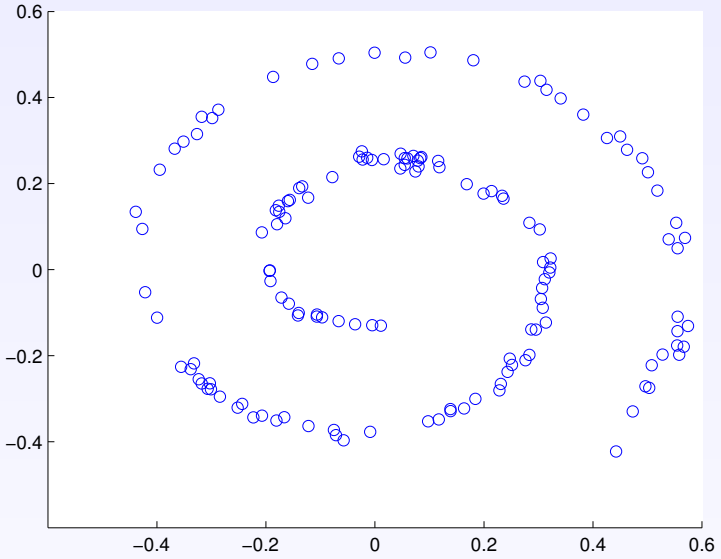


FIG.: Samples from a spiral distribution

Toy example

- We are trying to learn a density using this training set :

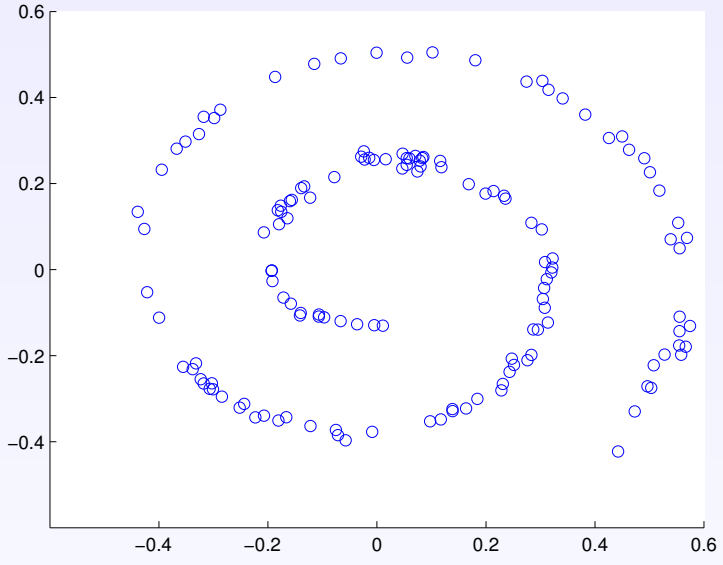
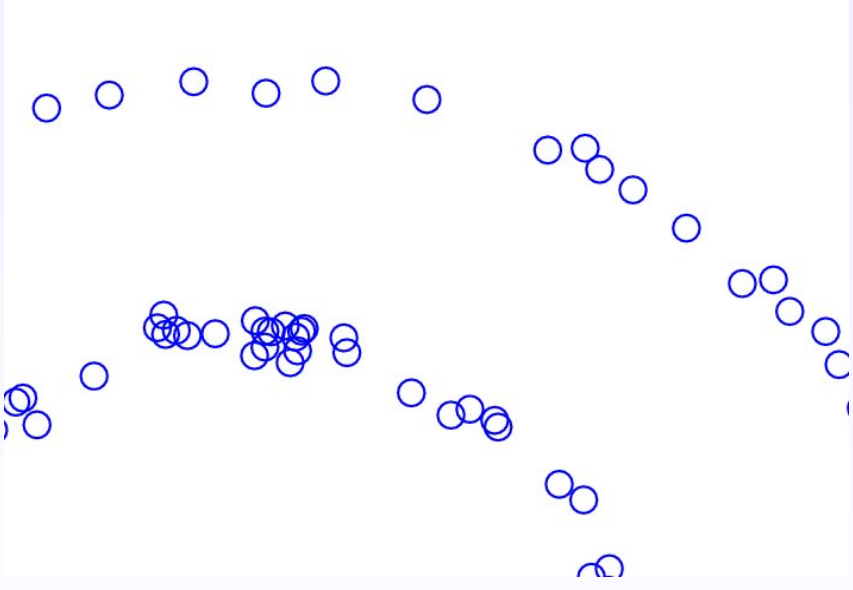


FIG.: Samples from a spiral distribution

- Let's train a Manifold Parzen Windows model, and look at the first principal direction of variance of the training point gaussians

Toy example

Because the training of Manifold Parzen Windows uses only local information, some of the principal directions of variance of badly estimated.



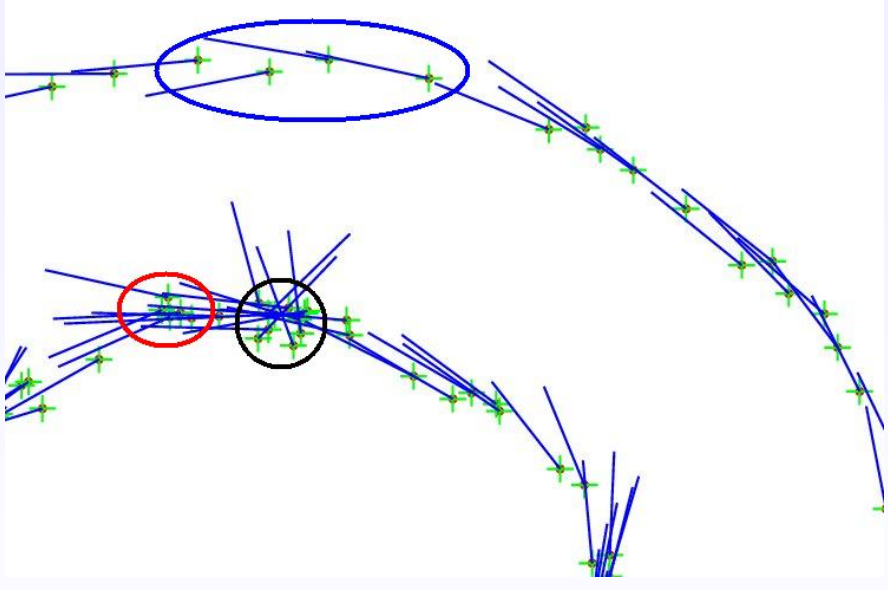


FIG.: On the left : training points. On the right : first principal direction of variance.

Real life examples

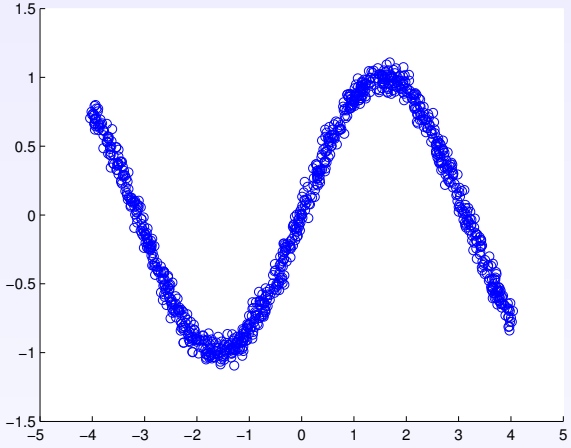
A lot of large scale, real life problems are likely to befenit from non-local learning :

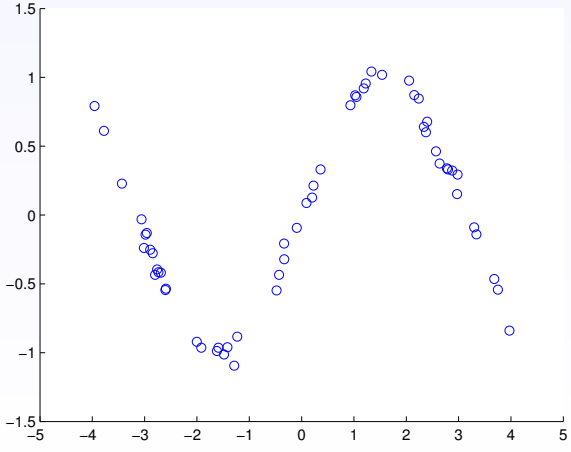
- **Vision :** the pixels at a certain position from very different images share the same properties with respect to certain transformations (e.g. translation, rotation) ;
- **Natural Language Processing :** words that are very different in some aspect still usually share a lot of properties (e.g. two nouns, even if they have very different meanings, will still obey to the same grammatical rules)

Plan

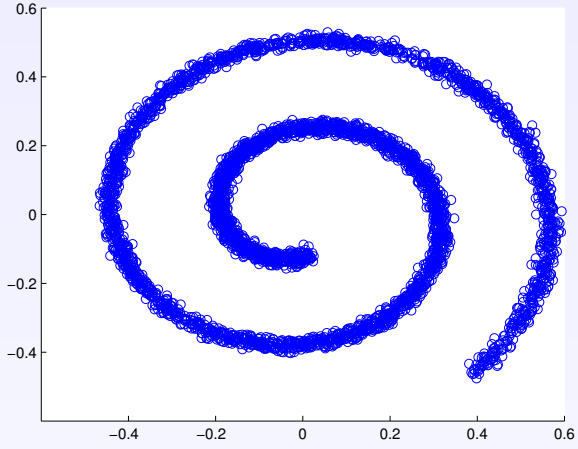
Toy 2D data experiments

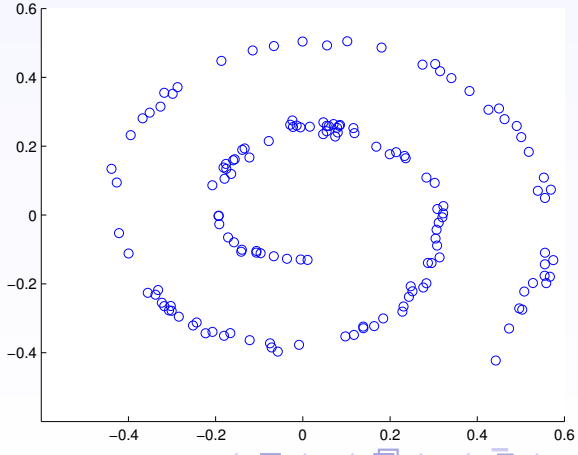
Sinusoidal distribution





Spiral distribution





Toy 2D data experiments

- Results :

Algorithm	sinus	spiral
Non-Local MP	**1.144**	**-1.346**
Manifold Parzen	1.345	-0.914
Gauss Mix Full	1.567	-0.857
Parzen Windows	1.841	-0.487

TAB.: *Average out-of-sample negative log-likelihood on two toy problems, for Non-Local Manifold Parzen, a Gaussian mixture with full covariance, Manifold Parzen and Parzen Windows. The non-local algorithm dominates all the others.*

Toy 2D data experiments

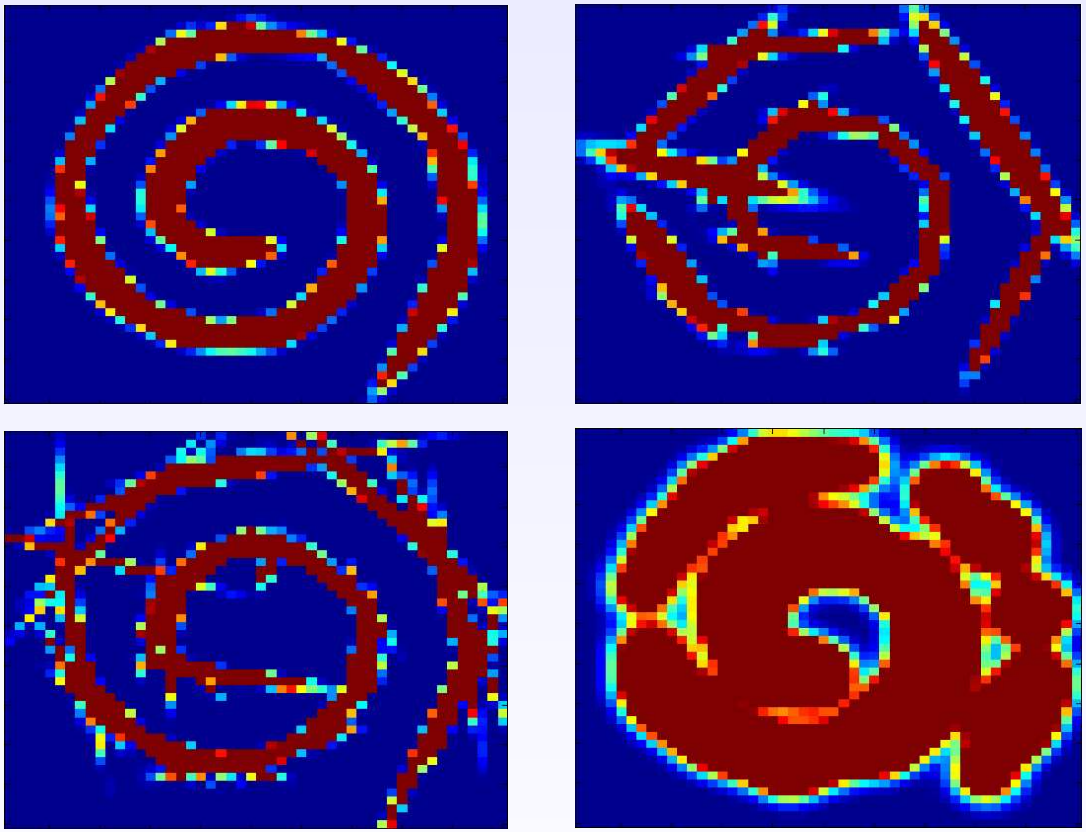
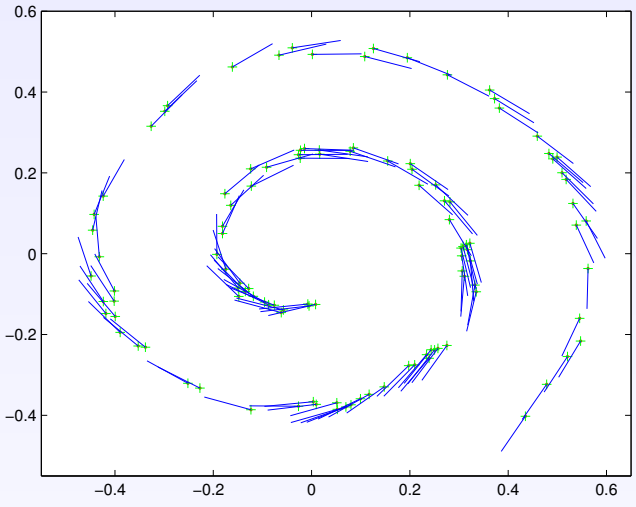
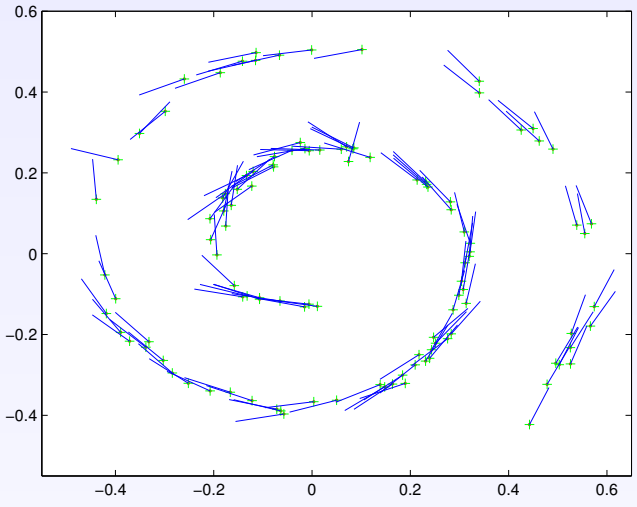


FIG.: *From left to right, top to bottom, densities learned by Non-Local Manifold Parzen, a Gaussian mixture with full covariance, Manifold Parzen and Parzen Windows.*

Toy 2D data experiments



(a) Non-Local Manifold Parzen



(b) Manifold Parzen

FIG.: *Illustration of the learned principal directions for Non-Local Manifold Parzen and local Manifold Parzen, for the spiral distribution data set.*

Experiments on rotated digits

- 729 first examples in the USPS digit recognition training set

Experiments on rotated digits

- 729 first examples in the USPS digit recognition training set
- Add two rotated versions (0.1 and 0.2 radians) of each of those examples

Experiments on rotated digits

- 729 first examples in the USPS digit recognition training set
- Add two rotated versions (0.1 and 0.2 radians) of each of those examples
- Train on digits 2 to 9, test on rotated 1 digits

Experiments on rotated digits

- 729 first examples in the USPS digit recognition training set
- Add two rotated versions (0.1 and 0.2 radians) of each of those examples
- Train on digits 2 to 9, test on rotated 1 digits
- For NLMP, allow gaussians to be centered on original, unrotated 1 digits

Experiments on rotated digits

- 729 first examples in the USPS digit recognition training set
- Add two rotated versions (0.1 and 0.2 radians) of each of those examples
- Train on digits 2 to 9, test on rotated 1 digits
- For NLMP, allow gaussians to be centered on original, unrotated 1 digits
- For (Manifold) Parzen Windows, do as usual, by including the unrotated 1 digits in the training set

Experiments on rotated digits

- 729 first examples in the USPS digit recognition training set
- Add two rotated versions (0.1 and 0.2 radians) of each of those examples
- Train on digits 2 to 9, test on rotated 1 digits
- For NLMP, allow gaussians to be centered on original, unrotated 1 digits
- For (Manifold) Parzen Windows, do as usual, by including the unrotated 1 digits in the training set
- The number of principal directions of variance was set to one

Experiments on rotated digits

- Results :

Algorithm	Validation	Test
Non-Local MP	**-73.10**	**-76.03**
Manifold Parzen	65.21	58.33
Parzen Windows	77.87	65.94

TAB.: *Average Negative Log-Likelihood on the digit rotation experiment, when testing on a digit class (1's) for Non-Local Manifold Parzen, Manifold Parzen, and Parzen Windows. The non-local algorithm is clearly superior.*

Experiments on rotated digits

- We can use the predicted principal direction of variance to rotate an image, by making small steps

Experiments on rotated digits

- We can use the predicted principal direction of variance to rotate an image, by making small steps
- To illustrate non-local learning capacity, we rotate a sample of the digit 1 in the inverse direction as seen in the training set !

Experiments on rotated digits

- We can use the predicted principal direction of variance to rotate an image, by making small steps
- To illustrate non-local learning capacity, we rotate a sample of the digit 1 in the inverse direction as seen in the training set !

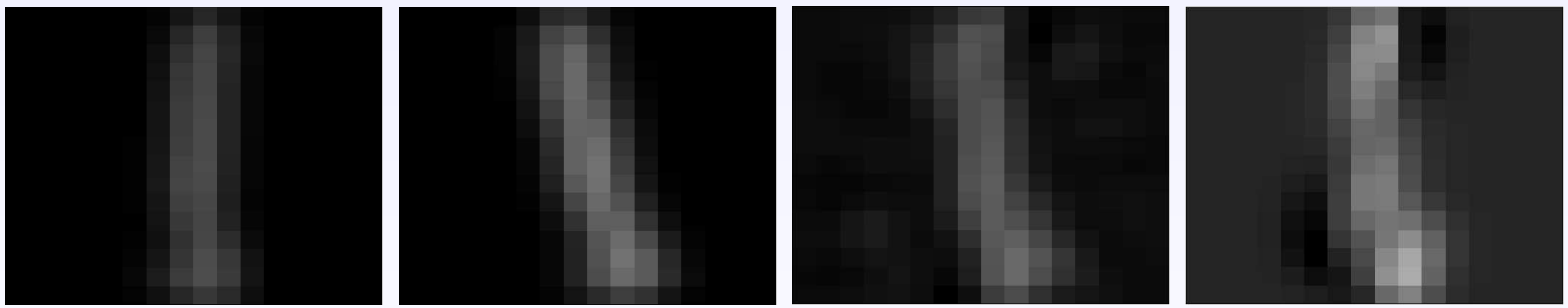


FIG.: *From left to right : original image of a digit 1 ; rotated analytically by -0.2 radians ; rotation predicted using Non-Local MP ; rotation predicted using MP. Rotations are obtained by following the tangent vector in small steps.*

Experiments on digit recognition task

- Digit recognition on the USPS dataset

Algorithm	Valid.	Test
SVM	1.2%	4.68%
Parzen Windows	1.8%	5.08%
Manifold Parzen	0.9%	4.08%
Non-local MP	**0.6%**	**3.54%**

TAB.: *Classification error obtained on USPS with SVM, Parzen Windows and Local and Non-Local Manifold Parzen Windows classifiers.*

Plan

Conclusion

- We developped a non-local version of Manifold Parzen Windows

Conclusion

- We developped a non-local version of Manifold Parzen Windows
- This model is able to better estimate the density of data lying on a lower dimensional manifold, by sharing information about it's structure among all training points

Conclusion

- We developped a non-local version of Manifold Parzen Windows
- This model is able to better estimate the density of data lying on a lower dimensional manifold, by sharing information about it's structure among all training points
- We showed the capacity of non-local learning to generalize far from training examples

Conclusion

THANK YOU !

Vincent, P. and Bengio, Y. (2003).
Manifold parzen windows.
In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA. MIT Press.