Representation of High Dimensional Data

Ross Goroshin

June 11, 2013

NEW YORK UNIVERSITY

Classification: Invariance, not Independence

Image Classifier Demo – Chromium

Image Classifier Demo × | outdoor image - Google S ×

horatio.cs.nyu.edu

openSUSE | Latest Headlines | Chromium | Imported

Image Classifier Demo | Demo | About | Terms

Predicted objects:

1. Home Theater, Home Theatre (0.21)
2. Studio Couch, Day Bed (0.08)
3. Window Shade (0.08)
4. Washbasin, Handbasin, Washbowl, Lavabo, Wash-Hand Basin (0.07)
5. Golfcart, Golf Cart (0.05)

Other objects:

Remove

Predicted objects:

1. Cliff, Drop, Drop-Off (0.26)
2. Promontory, Headland, Head, Foreland (0.22)
3. Canoe (0.18)
4. Paddle, Boat Paddle (0.11)
5. Valley, Vale (0.05)

Other objects:

Remove

Predicted objects:

1. American Lobster, Northern Lobster, Maine Lobster, Homarus Americanus (0.58)
2. Harvester, Reaper (0.15)
3. Racer, Race Car, Racing Car (0.07)
4. Tractor (0.05)
5. Thresher, Thrasher, Threshing Machine (0.04)

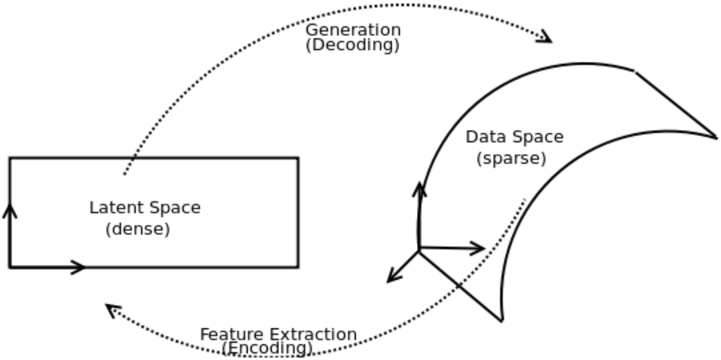
Other objects:

Remove

Dimensionality of Data & Statistical Dependence

- Suppose we have a 42 second video played at 24 frames/second, with a resolution of 1000 by 1000 pixels
- In theory each pixel can vary independently from frame to frame, which implies that there are $\approx 10^9$ degrees of freedom
- If all of these pixels were to vary independently of one another, the picture would not be very interesting
- Moreover if you were to describe the content of the video to a friend over the telephone, it is doubtful that the term 'pixel' would ever be mentioned in the conversation

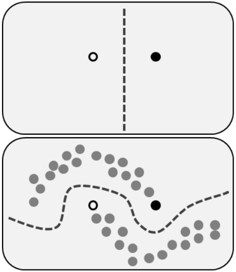
Dimensionality of Data & Statistical Dependence



- This illustration is representative of many processes
- However, dependence can be introduced without increasing the dimensionality
- Latent representation is NOT unique for generative processes of interest

Semisupervised Learning

- Sometimes, we really only care about class lables
- Leverage vast quantities of unlabled data
- Assume that x is the data and y are the lables
- Usually, $p(x)$ contains information about $p(y|x)$



Relationship Between Approaches

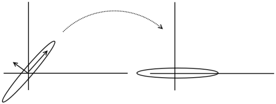
Algorithm	Model	Encode	Decode	Relate Enc. & Dec.
PCA	Global Linear	✓	✓	$W_D = W_E^T$
ICA	Global Linear	✓	✓	$W_D = W_E^T$
Sparse Coding	Local Linear	✓($)	✓	$W_D = W_E^T$
PSD & LISTA	Local Linear	✓	✓	Learned W_E
DrLIM	Nonlinear	✓	X	Enc. Only
Auto-Encoders	Nonlinear	✓	✓	Learned W_E & W_D

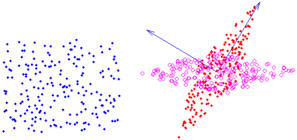
First Attempt at Independence: PCA

- Assume you have $x = As$ where each $x_i \in \mathbb{R}^D$ (s blue, x red)
- PCA assumes that there are $M \leq D$ linearly interdependent combinations of the input space variables which are responsible for most of the variance of the data

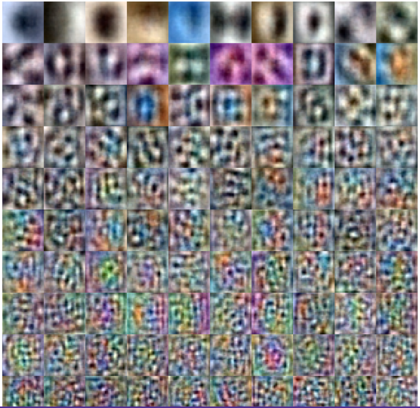
$$\frac{1}{N}\sum_{n=1}^{N}(e_1^T x_n - e_1^T \bar{x})^2 = e_1^T \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T e_1$$

Leading to the problem: maximize $e_1^T \Sigma e_1$ s.t $e_1^T e_1 = 1$ where Σ is the covariance matrix.





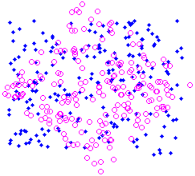
Global Variations are Not Everything



A Little Closer to Independence: Whitening

- Note that the covariance matrix of the data in PCA space is diagonal, i.e. the data is completely uncorrelated
- Whitening the data is equalizing the variance of the uncorrelated data
- The whitening transform is given by $V = WD^{-1/2}W^T$
- The whiteness property of the data is invariant to orthogonal transforms. Let $E[zz^T] = \mathbf{I}$, and let $y = Uz$ where U is an orthogonal transform.
- Then $E[yy^T] = E[Uzz^T U^T] = UE[zz^T]U^T = \mathbf{I}$. Whitening gives the linear independent , modulo an orthogonal transform

For radially symmetric distributions (e.g. Gaussian) we are done!

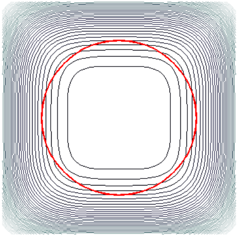


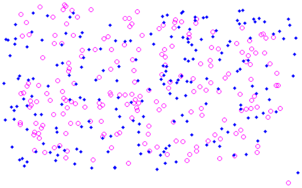
Independent Component Analysis

- Generative model $x = As$, where neither the A nor the s are known. Focus on picking out one of the s components at a time, i.e. $y = b^T As (= q_1 s_1 + q_2 s_2$ for 2D)
- Assume that the s mixture components are i.i.d and non-Gaussian, by the central limit theorem any mixture of the variables is more "Gaussian" than the individual distributions
- A good measure of non-Gaussianity is $k(y) = E[y^4] - 3(E[y^2])^2$. For whitened data, $k(y) = E[y^4] - 3$
- Since the data has been whitened, we constrain $E[y^2] = q_1^2 var(s_1) + q_2^2 var(s_2) + cov(s_1, s_2) = q_1^2 + q_2^2 = 1$

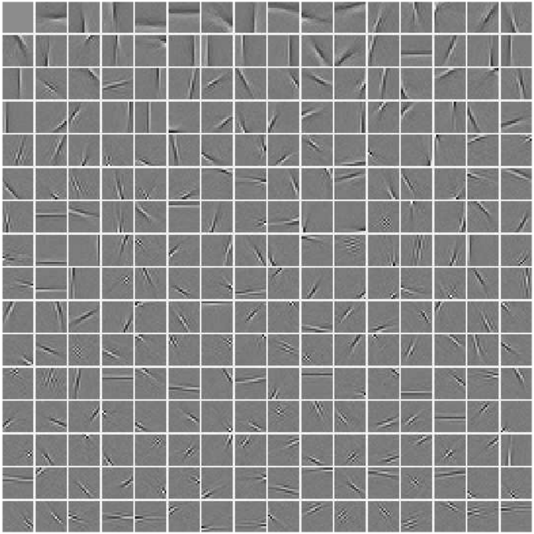
Independent Component Analysis

- max $|kurt(y)| = |q_1^4 kurt(s_1) + q_2^4 kurt(s_2)|$ s.t. $q_1^2 + q_2^2 = 1$
- Assuming that s_1 and s_2 are i.i.d then $kurt(s_1) = kurt(s_2)$.
- However we don't directly have access to the q variables (they are mixed by A) making it expensive to enforce the constraint $\|q\|^2 = 1$. If the data is whitened however, and we seek the linear combination $w^T z$ that maximizes non-Gaussianity then it can be show that $\|q\| = \|w\|$

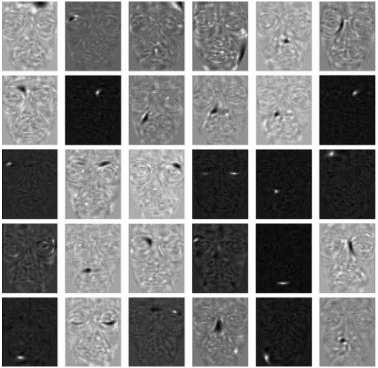




ICA Results



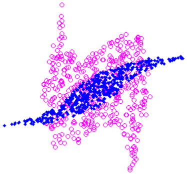
ICA Results



- Note that vectors found by ICA are more localized, i.e. individual activations are sparsely distributed (super-Gaussian) over the data

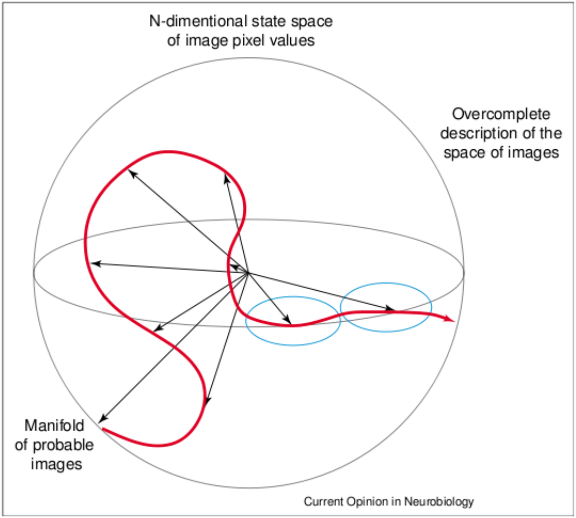
Limit of Classical ICA

The $x = As$ model is invalid for nonlinear data manifolds... in low dimensional space



- One way to approximate a nonlinear independent direction is to use a local linear approximation, which requires an over-complete basis
- Heuristic requirements for independence: (1) Basis vectors should be sparsely activated, (2) Basis vectors should be "quasiorthogonal"
- In 100 dimensional space, it is possible to arrange 400 basis vectors with more than 80 degrees between any two

Sparse Coding



Sparse Coding

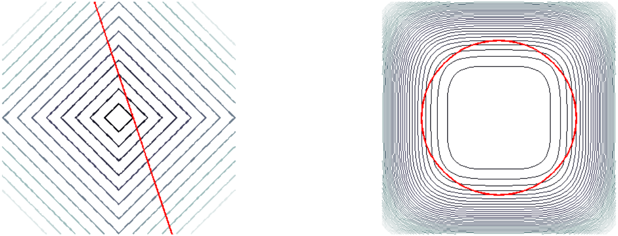
- Find a sparsely activated over-complete basis which describes the data
- Direct measure of sparsity, the L_0 norm, results in a combinatorial optimization problem
- Note that another popular measure of sparsity is kurtosis, which links directly back to ICA
- Modern formulations of sparse coding use the L_1 norm as a sparsity measure

$$min \|z\|_1 \text{ s.t } Wz = s \text{ (BP)}$$

$$\min_z \frac{1}{2}\|s - Wz\|_2^2 + \lambda\|z\|_1 \text{ (BPDN)}$$

$$\min_{z,W} \frac{1}{2}\|s - Wz\|_2^2 + \lambda\|z\|_1 \text{ (SC)}$$

Why L_1?



Left: Minimize L_1, goal: sparsity
Right: Maximize kurotsis, goal: independence

Implementation

$$\min_{z,W} \frac{1}{2}\|s - Wz\|_2^2 + \lambda\|z\|_1$$

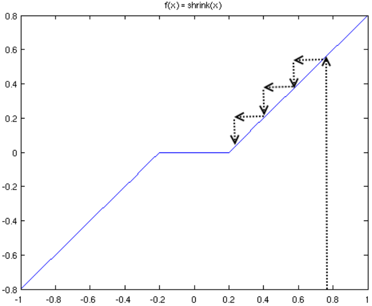
- Alternately optimize z (inference) and W (basis update)
- It is easy to reduce $\|z\|_1$ and increase the norm of the columns of W, thus the columns of W must be normalized to unity

ISTA Inference

- Given a fixed basis W there exists a fixed point point algorithm for finding the optimal coefficients z^* (i.e. inference)

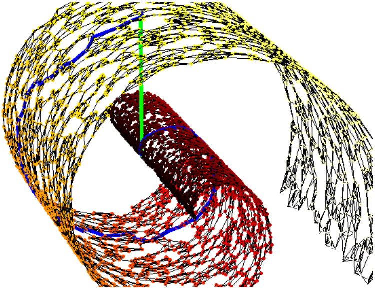
$$z_{k+1} = shrink(z_k - \eta_1 \nabla_{z_k} \frac{1}{2} \|s - Wz_k\|_2^2)$$

- Application of the $shrink()$ function corresponds to a gradient step in L_1



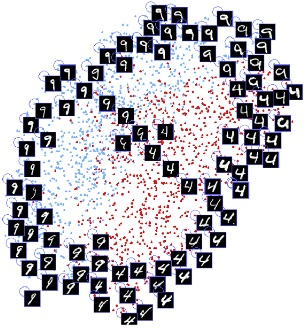
Metric Learning

- Geodesic v.s. Euclidean distance
- Another way to pose the problem is to find a distance preserving mapping to lower dimensional space
- For densely sampled manifolds this makes sense, but for realistic data we are satisfied with preserving some distance metric of interest, possibly mangling others



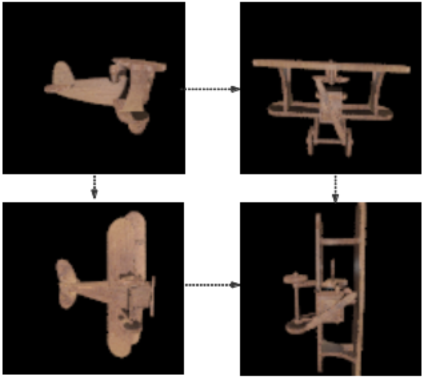
DrLIM

- We wish to find a mapping $G_W(X_i) : \mathbb{R}^D \to \mathbb{R}^d$, where $D > d$ which translates labeled similarity relationships in the input space to Euclidean distances in the output space
- If (X_1, X_2) are similar then $Y = 0$, otherwise $Y = 1$
- Let $D_W(X_1, X_2) = \|G_W(X_1), G_W(X_2)\|_2$
- $L(W, Y, X_1, X_2) = (1 - Y)\frac{1}{2}D_W^2 + Y\frac{1}{2}\{max(0, m - D_W)\}^2$

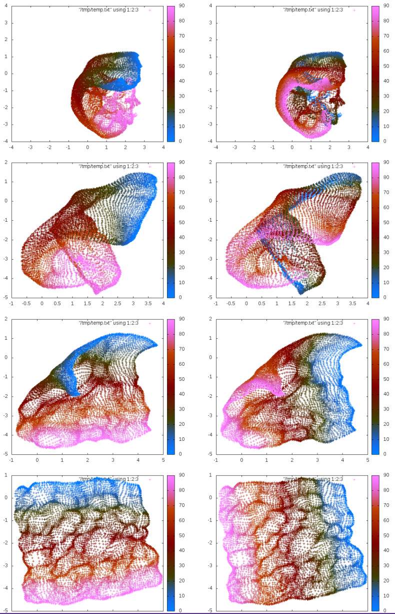


Speaking of Manifolds: A "Toy" Example

- 2-dimensional manifold living in a $\approx 10,000$-dimensional space (96x96 images)
- Similarity relationships can be naturally assigned via adjacent frames in a video



Speaking of Manifolds: A "Toy" Example



Thank You