

CSCI 4360/6360 Data Science II

Dictionary Learning

Embeddings

- Principal Components Analysis (PCA)
 - **Sparse & Kernel PCA (*last Thursday!*)**
 - Independent Components Analysis (ICA)
 - Non-negative Matrix Factorization (NMF)
 - Locally-linear Embeddings (LLE)
-
- **Dictionary Learning (*today!*)**

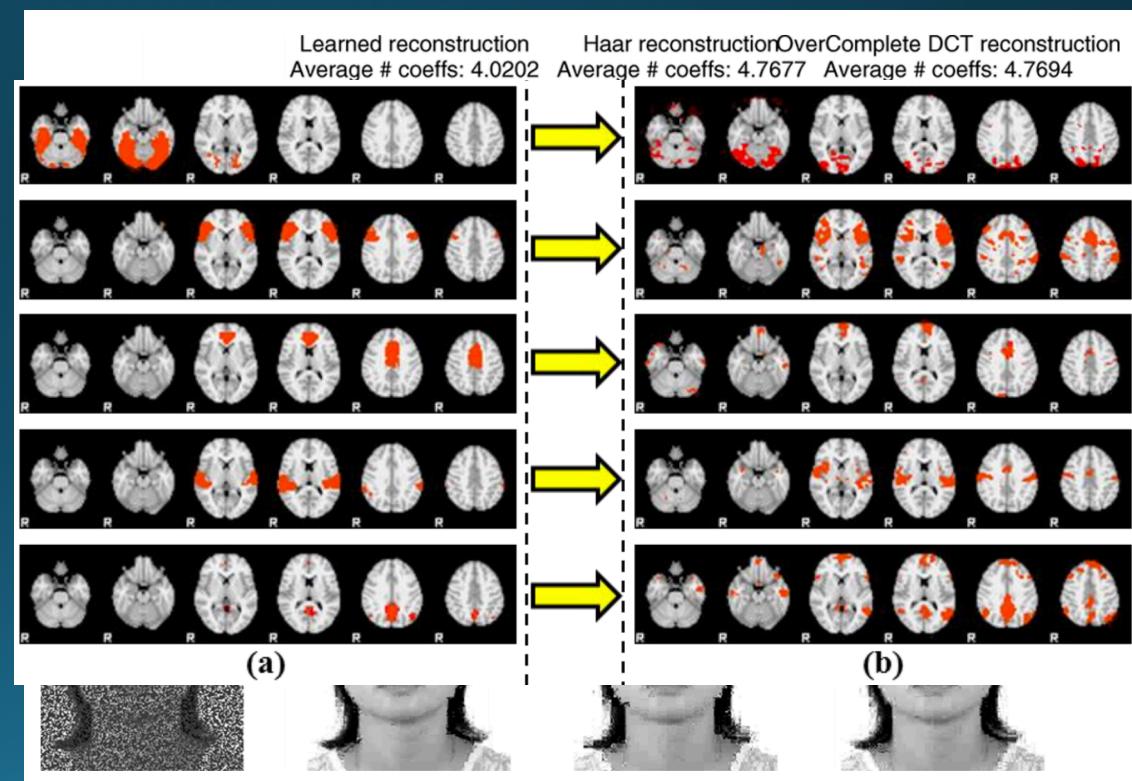


Dictionary Learning

- “*Given a set of signals belonging to a certain class, one wishes to extract the relevant information by identifying the generating causes; that is, recovering the elementary signals (atoms) that efficiently represent the data.*”
 - *Regularization, Optimization, Kernels, and Support Vector Machines*, Ch. 2
- Every embedding strategy ever?

Dictionary Learning

- Sparse coding
- ℓ^p sparsity
- Hierarchical sparse coding
- K-SVD
- Elastic net



Motivations

- Dictionary learning is ideally formulated for image denoising (and is indeed a major application of dictionary learning)



Measurements
(image)

$$y = x_{orig} + w$$

Original
image

Noise

Motivations

$$y = x_{orig} + w$$

- Easily converted to an energy minimization problem

$$E(\vec{x}) = \|\vec{y} - \vec{x}\|_2^2 + Pr(\vec{x})$$

- Some classical priors

- Smoothness
- Total variation
- Wavelet sparsity
- Lasso
- ...

$$\begin{aligned} & \lambda \|\mathcal{L}\vec{x}\|_2^2 \\ & \lambda \|\nabla \vec{x}\|_1^2 \\ & \lambda \|W\vec{x}\|_1 \\ & \lambda \|\vec{x}\|_1 \end{aligned}$$

Energy
minimization
becomes a MAP
estimation!

Dictionary Learning

- We have our data X
- and wish to represent it using some small number k atoms ($k \ll n$)
- When combined with coefficients, the linear combinations with the atoms should yield a nearly complete representation of X

$$X = [\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n]^T \in \mathbb{R}^{n \times m}$$

$$\vec{x}_i \cong \sum_{j=1}^k \theta_{ji} \vec{b}_j, \forall i = 1, \dots, n$$

$$B = [\vec{b}_1, \vec{b}_2, \dots, \vec{b}_k]^T \in \mathbb{R}^{k \times m}$$

$$\Theta = [\vec{\theta}_1, \vec{\theta}_2, \dots, \vec{\theta}_n]^T \in \mathbb{R}^{n \times k}$$

Dictionary Learning

- This gives the minimization

$$\min_{B, \Theta} \sum_{i=1}^n \left(\|\vec{x}_i - B\vec{\theta}_i\|_q^q + h(\vec{\theta}_i) \right)$$

where h promotes sparsity in the coefficients, and B is chosen from a constraint set

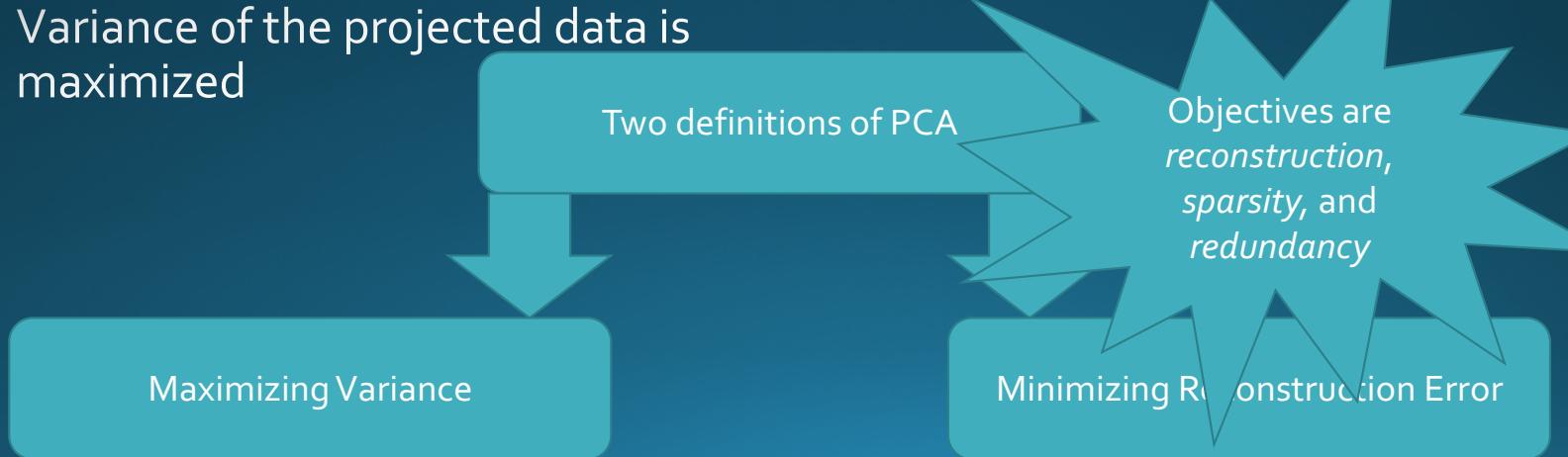
- The general dictionary learning problem then follows

$$\phi(\Theta, B) = \frac{1}{2} \|X - B\Theta\|_F^2 + h(\Theta) + g(B)$$

where specific choices of h and g are what differentiate the different kinds of dictionary learning (e.g. hierarchical, K-SVD, etc)

Dictionary Learning vs PCA

- Remember the operational definition of PCA?
 1. Orthogonal projection of data
 2. Lower-dimensional linear space known as the *principal subspace*
 3. Variance of the projected data is maximized



- Dictionary Learning (**sparse** coding)
 1. Minimize reconstruction error
 2. Linear combination of *atoms*
 3. Sparse, overcomplete basis

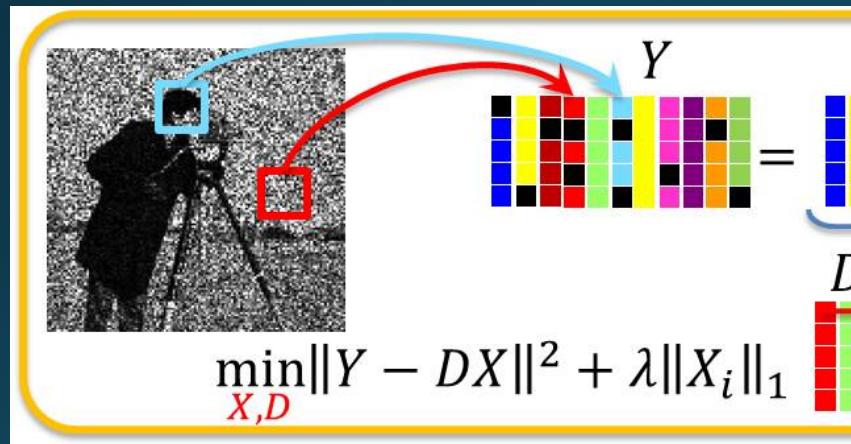
Dictionary Learning

$$Pr(\mathbf{x}) = \lambda \|\boldsymbol{\alpha}\|_0 \text{ for } \mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}$$

$$\underbrace{\begin{pmatrix} \mathbf{x} \\ \vdots \end{pmatrix}}_{\mathbf{x} \in \mathbb{R}^m} = \underbrace{\left(\mathbf{d}_1 \mid \mathbf{d}_2 \mid \cdots \mid \mathbf{d}_p \right)}_{\mathbf{D} \in \mathbb{R}^{m \times p}} \underbrace{\begin{pmatrix} \boldsymbol{\alpha}[1] \\ \boldsymbol{\alpha}[2] \\ \vdots \\ \boldsymbol{\alpha}[p] \end{pmatrix}}_{\boldsymbol{\alpha} \in \mathbb{R}^p, \text{sparse}}$$

Applications

- Image denoising
 - Sparse basis forces out noise



- object categorization
- Image restoration & inpainting



Dictionary Learning

B is typically implicitly constrained to fall within a *convex set C* of the $k \times m$ reals, to make optimization tractable

- General formulation

$$\phi(\Theta, B) = \frac{1}{2} \|X - B\Theta\|^2 + h(\Theta) + g(B)$$

- More common: set g to identity*, and h to L_1 norm

$$\phi(\Theta, B) = \min_{B \in \mathcal{C}} \frac{1}{2} \sum_{i=1}^n \|\vec{x}_i - B\vec{\theta}_i\|_2^2 + \lambda \|\vec{\theta}_i\|_1$$

Optimization

- Problems with the objective function?

$$\phi(\Theta, B) = \min_{B \in \mathcal{C}} \frac{1}{2} \sum_{i=1}^n \|\vec{x}_i - B\vec{\theta}_i\|_2^2 + \lambda \|\vec{\theta}_i\|_1$$

- Squared loss is convex
- Regularization is convex
- Squared loss + regularization is **not convex**
- Even worse, often **non-smooth**

Optimization

- Alternating minimization algorithm
 - Two-block Gauss-Seidel
- Streaming online learning
 - At iteration (or minibatch) t , signal \vec{x}_t and sparse code $\vec{\theta}_t$ are computing using the current dictionary

$$\begin{aligned} A\vec{x} &= \vec{b} & A &= L_* + U \\ L_*\vec{x}^{(k+1)} &= \vec{b} - U\vec{x}^{(k)} \end{aligned}$$

$$\vec{\theta}_t = \arg \min_{\vec{\theta}} \frac{1}{2} \|\vec{x}_t - B_{t-1} \vec{\theta}\|_2^2 + \lambda \|\vec{\theta}\|_1$$

- Which can then be used to update the dictionary

$$g_t(B) = \frac{1}{t} \sum_{i=1}^t \frac{1}{2} \|\vec{x}_i - B \vec{\theta}_i\|_2^2 + \lambda \|\vec{\theta}_i\|_1$$

- g can be efficiently solved using block coordinate descent on columns of B

Rank-1 Dictionary Learning (R1DL)

- KDD 2016

**Scalable Fast Rank-1 Dictionary Learning
for fMRI Big Data Analysis**

- "Scalable fast"

R1DL

- Reformulates dictionary learning as an alternating least-squares problem
 - (embraces the optimization procedure)
- Uses ℓ_0 -“norm” instead of ℓ_1
 - Given rank-1 formulation, this is an inexpensive way of guaranteeing sparsity
- Iteratively learns rank-1 dictionary atoms until k have been found
 - “Deflates” data matrix on each iteration

R1DL

- Energy function L
- Data matrix S , vectors u and v
 - $\|u\| = 1$
 - $\|v\|_0 \leq r$, where r is the sparsity constraint (literally, # of nonzero elements in v)
- Iterate until convergence of u (atoms) and v (sparse codes)

$$\vec{v} = \arg \min_{\vec{v}} \|S - \vec{u}\vec{v}^T\|_F \quad \vec{u} = \arg \min_{\vec{u}} \|S - \vec{u}\vec{v}^T\|_F = \frac{S\vec{v}}{\|S\vec{v}\|}$$
$$\|\vec{u}^{(j+1)} - \vec{u}^{(j)}\| < \epsilon$$

- “Deflate” data matrix $S^{(t+1)} = S^{(t)} - \vec{u}\vec{v}^T$
- Repeat until k atoms & sparse codes are learned

Summary

- Dictionary learning is focused on developing a basis of *atoms* and *coefficients*
 - Coefficients are *sparse*
 - Atoms form an *overcomplete* representation of the data
 - Chosen to minimize *reconstruction error*
- Explicitly factorizes out noise
 - Can be customized in the form of a prior
- Optimization is often non-convex and non-smooth, requiring alternating minimization strategies or online learning
- R1DL focuses on leveraging optimization strategies to iteratively learn the basis, one atom at a time
- Other variants include K-SVD, Hierarchical DL, and Elastic Net

Questions?

Course Details

- Feedback on proposals is coming
 - Expect this weekend
- Assignment 5 is out today!
 - The final assignment!
 - Due Thursday, November 2
- Assignment 4 is due **tonight!**
- Next week:
 - Conclude dimensionality reduction on Tuesday
 - Begin neural networks on Thursday
 - November is **all NNs, all the time**

References

- “Sparse coding with an overcomplete basis set: A strategy employed by V1?”,
<http://www.sciencedirect.com/science/article/pii/S0042698997001697>
- “Learning image representations from the pixel level via hierarchical sparse coding”,
<https://pdfs.semanticscholar.org/96be/a45d9b962c8159883a0d07493a2ea42784e4.pdf>
- “K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation”, <http://www.cs.technion.ac.il/~freddy/papers/120.pdf>
- “Scalable Fast Rank-1 Dictionary Learning for fMRI Big Data Analysis”,
<http://www.kdd.org/kdd2016/papers/files/adpo864-li.pdf>
- “Dictionary learning for sparse coding: Algorithms and convergence analysis”,
<https://pdfs.semanticscholar.org/284d/9c3702b8820e2bd89b1c96532f12b3afb336.pdf>
- “Sparse Coding and Dictionary Learning for Image Analysis”,
https://lear.inrialpes.fr/people/mairal/tutorial_iccv09/tuto_part2.pdf