A-OPTIMAL NON-NEGATIVE PROJECTION FOR IMAGE REPRESENTATION

Authors: Haifeng Liu, Zheng Yang, zhaohui Wu and Xuelong Li

Zhejiang University

Presenter: Wang Yuxiang

Outline of Presentation

I. Non-negative Matrix Factorization
II. Optimal design of experiment
III. The algorithm of this paper
IV. Convergence: a incomplete summary
V. Experiments
VI. Critiques and questions

Non-negative Matrix Factorization

- Nature 1999: Lee and Seung use NMF to learn parts model. (cited 3417 times)
- NIPS'01 paper: Lee and Seung, Algorithms for NMF (cited 2429 times)
  - Solve

$$\min_{U,V>0} \|X - UV\|_F^2$$

  By iterating

$$U_{ij} \leftarrow U_{ij} \frac{(XV^T)_{ij}}{(UVV^T)_{ij}}$$

$$V_{ij} \leftarrow V_{ij} \frac{(U^T X)_{ij}}{(U^T UV)_{ij}}$$

  - Convergence proof using the auxiliary function.

Interpretation of NMF

- NMF (Lee&Seung, 1999)
  - U: meaningful parts,
  - V: coefficients of linear combination
  - They must be positive
- Sparse NMF (Hoyer, 2004)
  - V is sparse (each objects contains only a small number of parts.)
- Other variants
  - Add various other constraints on U and V

Application of NMF

- Represent object by parts (as in many papers discussed in class)
  - Clustering of image objects
  - Clustering of text/documents (Wu, 03)
- Dimension Reduction/Compact Representation
  - This paper concerns linear regression on the reduced representation
  - But apply to object clustering…

Why regression?

- Jiaming and Shahzor asked this question.
- Short answer: this is their assumption.
  - Not justified in the paper. Used mainly as a heuristic regularization term. Likely it's better than nothing…
- Long answer:
  - NMF is used to obtain a low-rank representation (capturing both data and noise).
  - The final goal is to apply regression to the compact representation (so that # of data > dimension).

Optimal design in the "Design of Experiments"

- Design experiments to estimate statistical models efficiently. (Learning)
  - Less # of samples, less money and labor
  - Better estimation when data are fixed.
- Optimal design:
  - Unbiased
  - Minimum-variance
- Easy for single variable estimator (e.g., mean)
  - Least square estimator minimizes variance.

Optimality criteria

- When there are multiple parameters…
  - A-optimality (trace of inv(X'X) average variance)
  - C-optimality (predefined linear combination)
  - D-optimality (determinant of inv(X'X))
  - E-optimality (min eigenvalue of Info Mat)
  - G-optimality (minimax variance)
  - and many more…

Remarks on different criteria

- It is hard to say which metric is better.
- All are better than not doing optimal design at all. (John Cornell, 02)
- A-Optimality is the most basic criterion and easiest for analysis.
  - Maybe the reason why the authors choose this.
  - Practical concerns of image representation: is A-Optimality good enough?

Fisher information matrix

- Inverse of covariance matrix
  - Minimize variance is equivalent of maximize information.
  - The current state-of-the-art features in object recognition is based on Fisher Information.

Back to the paper here

A-Optimal Non-Negative Projection (ANP)

- Regression model

$$y_i = w^T v_i + \epsilon$$

- The regularized least square solution of

$$\min_w \sum_{i=1}^{n} (y_i + w^T v_i)^2 + \lambda ||w||^2$$

- in closed-form is

$$\hat{w} = (VV^T + \lambda I)^{-1} Vy$$

A-Optimal Non-Negative Projection (ANP)

- Covariance matrix is given in closed form

$$Cov(\hat{w}) \propto (VV^T + \lambda I)^{-1}$$

- So A-optimal solution of the regression model together with NMF

$$\min_{U,V>0} ||X - UV||^2 + \lambda_1 Tr((VV^T + \lambda_2 I)^{-1})$$

- Inverse is troublesome, so

$$\min_{U,V>0,W} ||X - UV||^2 + \lambda_1 (\left||I - VW|\right|^2 + \lambda_2 ||W||^2)$$

Alternating algorithm

1. Fix U,V, solve

$$W = (V^T V + \lambda_2 I)^{-1} V^T$$

2. Fix W, solve

$$u_{ij} \leftarrow u_{ij} \frac{(XV^T)_{ij}}{(UVV^T)_{ij}}$$

$$v_{ij} \leftarrow v_{ij} \frac{(U^T X + \lambda_1 W t^+)_{ij}}{(U^T UV + \lambda_1 W t^- + \lambda_1 VWW^T)_{ij}}$$

3. Repeat until converge.

Convergence: an incomplete summary

- What does it mean
  - when we say a sequence 'converges'?
  - when an author claim their algorithm has convergence guarantee?
- It depends
  - Trickier than expected.
  - Bear with me if you find it trivial

Levels of convergence guarantee

- $f(x_k)$ converges
  - lim $f(x_k)$ -> C, when k-> inf
  - No cues what the limit is, e.g., some Block Coordinate Descent (BCD) convergence.
  - Simple proof: 1. bounded from below (e.g., all norm objectives), 2. monotonic decreasing.

Levels of convergence guarantee

- $f(x_k)$ converges to the a local/global minimum/stationary point
  - C has the described property
  - For proof: This is much harder than the brainless previous level. It needs to show ALL points invariant to the updates obey KKT condition.
  - No cues when it converges.
- Example: EM, Proximal Alternating Alg…

Levels of convergence guarantee

- $f(x_k)$ converges to the a stationary point with a rate of convergence $O(1/k)$
  - $||x_k - x^*|| < O(1/k)$
  - For proof: Usually requires in-depth understanding of the iterates.
  - E.G., gradient descent is $O(1/k)$ for convex obj and Lipchitz continuous. With Nesterov Acceleration Point it becomes $O(1/k^2)$.
  - Newton's method is locally $O(1/k^2)$.

Local and global convergence

- Global convergence
  - Converge with any initialization.
  - Does NOT imply it converges to global solution!
  - E.G., BCD convergence.
- Local convergence
  - Converge only when initial variable is close to a solution.
  - E.G., Newton method.

Levels of convergence guarantee

- Good algorithms usually have at least convergence guarantee to some stationary points.
- Great algorithms usually have strongest guarantee of convergence.
  - Rate of convergence.
  - Time and space complexity of each iterate.
- Guarantees are usually weak for non-convex/non-differentiable optimization.

Here in this paper

- How strong is the guarantee here?
- The method is guaranteed to converge to a stationary point.
  - but may take infinity number of iterations.
- Empirically, such alternating methods
  - Fast at beginning
  - Slow to approach solution

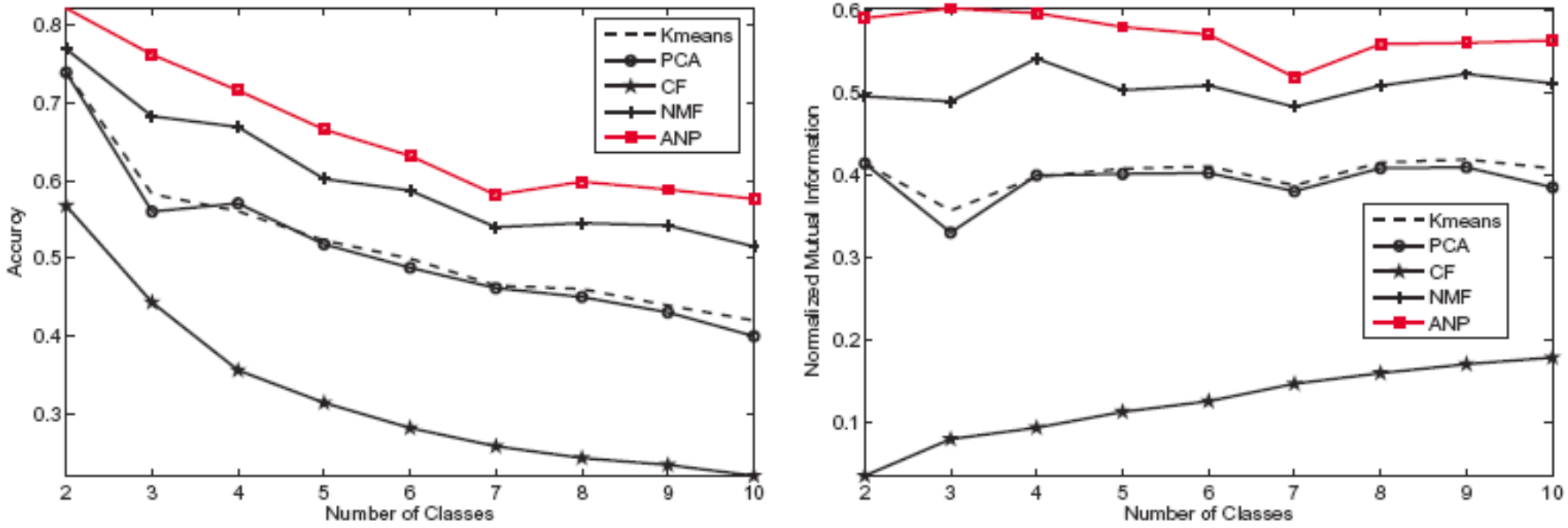
Great names in optimization

- Yurii Nesterov, Arkadi Nemirovski
- The famous Stephen Boyd: ADMM
- Here at NUS:
  - Prof. Toh Kim Chuan: APG
- Closely connected to our group
  - Prof. Zhouchen Lin from PKU

Experiments on object clustering

- Caltech-101 object: SIFT+freq hist
- Cambridge ORL face: raw pixel
- Method:
  - Original representation
  - NMF to get a compact representation
  - K-means for clustering

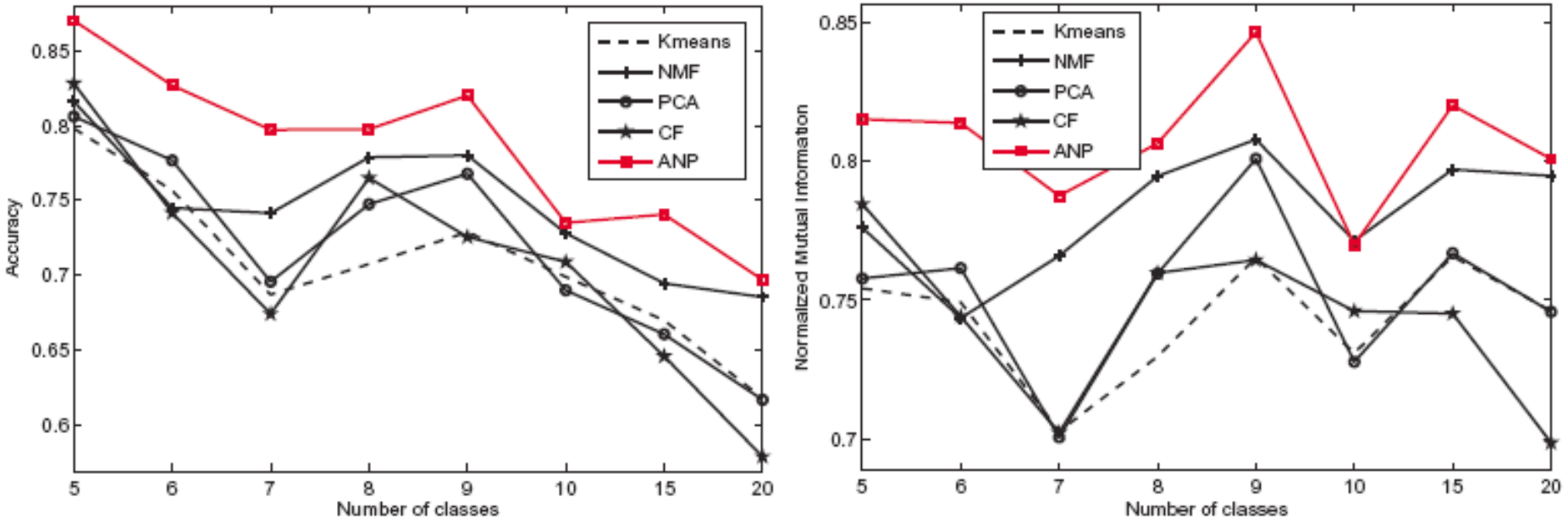
Results on Caltech101



(a) Accuracy (b) Normalized mutual information

Figure 3. The clustering result on the Caltech-101 database

Results on ORL



(a) Accuracy (b) Normalized mutual information

Figure 2. The clustering result on the ORL database

Critiques and questions

- Hunfuko: Uniqueness of NMF and ANP
  - The author does not discuss uniqueness of NMF. It does certainly return an approximate solution with positive U and V.
  - For an analysis of NMF's uniquenss, check out Donoho's paper: “When Does Non-Negative Matrix Factorization Give a Correct Decomposition into Parts?”

Critiques and questions

- Shahzor: There are many local optima. Does heuristics based initialization help the quality of the solution obtained (over that from random initialization)?

- Shahzor: what is the connection between regularization of a regression model and NMF? Why can it be simply carried through?

Critiques and questions

- Jiaming: Why the minimization of the size of the parameter covariance can be used to minimize the expected prediction error?
  - It can't.

$$\operatorname{MSPE}(L) = \sum_{i=1}^{n} \left(\operatorname{E}\left[\widehat{g}(x_i)\right] - g(x_i)\right)^2 + \sum_{i=1}^{n} \operatorname{var}\left[\widehat{g}(x_i)\right].$$

  - It can minimize the variance term, but only at a cost of increasing bias term.

Critiques and questions

- Jiaming: Why to use the linear regression model? What is the relationship between the w, y in (6) and the U, X in (4)?
- If they are the same, why is there only one w in (6) while U in (4) contains m bases?
  - y is label of data point v. w is coefficients.
  - They assume the data can be represented by a simple regression model. Using NMF can capture the noise too.

Critiques and questions

- Jiaming: In Fig. 2(a), why the accuracy is not monotonically decreasing according to the number of class?
  - I think it depends on the inner dimension of U, V. They increase r together with K.

Summary

- The paper introduces the optimal design of experiments to vision community.
- The improvement is good comparing to baseline algorithms, but not good enough for a 2012 paper. The state-of-the-art is much better?
- The use of linear regression model is unjustified.
- The convergence result is weak.