NNLS and NMF

Christian Knoll,
Jakob Mangelberger

Signal Processing and Speech Communication Laboratory

January 30, 2012

Outline

Outline

LSI

- Least squares: $min||Ex - f||$
  $E \in \mathbb{R}^{m \times n}$
  $x \ \in \mathbb{R}^{n \times 1}$
  $f \ \in \mathbb{R}^{m \times 1}$

LSI

- Least squares: $min||Ex - f||$
  $E \in \mathbb{R}^{m \times n}$
  $x \in \mathbb{R}^{n \times 1}$
  $f \in \mathbb{R}^{m \times 1}$
- many applications where additional information is needed
  - introduction of inequality constraints
  - linear constraints only

LSI

- Least squares: $min||Ex - f||$
  $E \in \mathbb{R}^{m \times n}$
  $x \ \in \mathbb{R}^{n \times 1}$
  $f \ \in \mathbb{R}^{m \times 1}$
- many applications where additional information is needed
  - introduction of inequality constraints
  - linear constraints only
- LSI: $min||Ex - f||$
  $s.t. \ \ Gx \geq h$

NNLS

- these constraints allow e.g.
  - upper and lower bound to each variable
  - sum of all variables is bounded
  - fitted curve is monotone or convex
  - nonnegative solution vector

NNLS

- these constraints allow e.g.
  - upper and lower bound to each variable
  - sum of all variables is bounded
  - fitted curve is monotone or convex
  - nonnegative solution vector
- NNLS: $min||Ex - f||$

  $s.t.\ \ x \geq 0$

Karush Kuhn Tucker Conditions

- Conditions for optimal solution

Karush Kuhn Tucker Conditions

- Conditions for optimal solution
- Generalization of Lagrange multipliers which hold only for equality constraints

Karush Kuhn Tucker Conditions

- Conditions for optimal solution
- Generalization of Lagrange multipliers which hold only for equality constraints
- 2 sets and 2 solution vectors
  $t = Gx - h$
  $y = E^T(Ex - f)$ (gradient vector)

Karush Kuhn Tucker Conditions

- Conditions for optimal solution
- Generalization of Lagrange multipliers which hold only for equality constraints
- 2 sets and 2 solution vectors
  $t = Gx - h$
  $y = E^T(Ex - f)$ (gradient vector)
- $t = 0 \; for \; i \in A$; $t > 0 \; for \; i \in B$
  $y \geq 0 \; for \; i \in A$; $y = 0 \; for \; i \in B$

Active Set

- Optimization Problem with set of constraints $g_i(x) \geq 0$

Active Set

- Optimization Problem with set of constraints $g_i(x) \geq 0$
- active at x if $g_i(x) = 0$
  nonactive otherwise

Active Set

- Optimization Problem with set of constraints $g_i(x) \geq 0$
- active at x if $g_i(x) = 0$
  nonactive otherwise
- active set holds the constraints which influence the solution

Solving NNLS

- empty non active set A

Solving NNLS

- empty non active set A
- active set B with all entries
  variables in A can take arbitrary values

Solving NNLS

- empty non active set A
- active set B with all entries
  variables in A can take arbitrary values
- compute the dual vector w
  based on the gradient of the current solution vector x

Solving NNLS

- empty non active set A
- active set B with all entries
  variables in A can take arbitrary values
- compute the dual vector w
  based on the gradient of the current solution vector x
- take the greatest w_i and move it to A

Solving NNLS con.

- calculate the solution vector x

Solving NNLS con.

- calculate the solution vector x
- iterate until the active set is empty

Solving NNLS con.

- calculate the solution vector x
- iterate until the active set is empty
- x and the dual vector w satisfy KKT-condition

Solving NNLS con.

- calculate the solution vector x
- iterate until the active set is empty
- x and the dual vector w satisfy KKT-condition
- x is a solution vector for the nonnegative least squares problem $Ex \approx f$

Outline

Nonnegative Least Squares Problem
- Types of Least Squares Problems
- Additional Definitions
- Solving NNLS

NMF
- Problem Formulation
- Solving NMF
- Algorithms

Example
- Comparison to other Matrix Factorizations

Problem Formulation

- Matrix Factorization(lower rank approximation): $A \approx WH$
  $A \in \mathbb{R}^{m \times n}$
  $W \in \mathbb{R}^{m \times k}$
  $H \in \mathbb{R}^{k \times n}$
  $k << m, n$

Problem Formulation

- Matrix Factorization(lower rank approximation): $A \approx WH$
  $A \in \mathbb{R}^{m \times n}$
  $W \in \mathbb{R}^{m \times k}$
  $H \in \mathbb{R}^{k \times n}$
  $k << m, n$
- applied to data analysis problems

Problem Formulation

- Matrix Factorization(lower rank approximation): $A \approx WH$
  $A \in \mathbb{R}^{m \times n}$
  $W \in \mathbb{R}^{m \times k}$
  $H \in \mathbb{R}^{k \times n}$
  $k << m, n$
- applied to data analysis problems
- nonnegative basis vectors

Problem Formulation cont.

- $min||A - WH||^2 s.t. \ W, H \geq 0$

Problem Formulation cont.

- $min||A - WH||^2 s.t. \; W, H \geq 0$
- W: basis Matrix
  H: coefficient Matrix
  A: each column corresponds to a data point

Problem Formulation cont.

- $min||A - WH||^2 s.t.\ W, H \geq 0$
- W: basis Matrix
  H: coefficient Matrix
  A: each column corresponds to a data point
- NMF gives a simple interpretation using non- negative basis vectors.

Problem Formulation cont.

- $min||A - WH||^2 s.t.\ W, H \geq 0$
- W: basis Matrix
  H: coefficient Matrix
  A: each column corresponds to a data point
- NMF gives a simple interpretation using non- negative basis vectors.
  - text data mining
  - cancer class discovery
  - image processing

Optimization

- Optimization problem is nonlinear and not convex
- Only local optimum might be found
- Can be generalized to the k-means clustering problem (NP-Complete)
- Solution is not unique:

$$\|A - WH\|_F = \|A - WXX^{-1}H\|_F$$

with $WX \geq 0$ and $X^{-1}H \geq 0$

Two Block Coordinate Descent

- Reformulate the non-convex optimization problem
- Two-block coordinate descent problem:

$$\min_{W \geq 0} \|H^T W^T - A^T\|_F^2$$

$$\min_{H \geq 0} \|WH - A\|_F^2$$

- Use known $H, W \geq 0$
- Subproblems can be solved using:
  - quasi-Newton optimization
  - NNLS
  - Projected gradient descent

Two Block Coordinate Descent

- Iterate the subproblems:
  - Initialize H with a non-negative matrix
  - Repeat until a stopping criterion is satisfied
    - $W^{t+1} = \arg\min_W f(W, H^{(t)}) \qquad s.t. W \geq 0$
    - $H^{t+1} = \arg\min_H f(W^{t-1}, H) \qquad s.t. H \geq 0$
  - $t \leftarrow t + 1$
- Initialization with either W or H

Convergence of the Two Block Coordinate Descent

- According to KKT optimality conditions, (W, H) is a stationary point iff

$$W \geq 0$$
$$\nabla_W f(W, H) = WHH^T - AH^T \geq 0$$
$$W. * \nabla_W f(W, H) = 0$$

- Block coordinate descent algorithm:
  - Limit point $\rightarrow$ Stationary point
  - For unique solution of the subproblem
- Two block coordinate descent:
  - Any limit point $\rightarrow$ Stationary point
  - Uniqueness not required
  - Sequence (W,H) based on Optimal solutions

NMF / NUR

- Commonly utilized NMF algorithm
- Lee and Seung (1999)
- Norm-based multiplicative update rules
  - Squared Euclidean distance
  - Divergence based update rule
- Easy to implement
- Computationally cheap

NMF / NUR

- Squared Euclidean distance:

$$W_{iq} \leftarrow W_{iq} \frac{\left(AH^T\right)_{iq}}{\left(W\left(HH^T\right)\right)_{iq}}$$

$$H_{qj} \leftarrow H_{qj} \frac{\left(W^T A\right)_{qj}}{\left(\left(W^T W\right) H\right)_{qj}}$$

- Variation of the gradient descent method
- Step size $\rightarrow$ Rescaling

- Division by 0 must be prevented
- Gradient descent might converge slowly

NMF / NUR

- Gradient of squared distance:

$$\min \nabla_H f(W, H) = (W^T A)_{qj} - (W^T W H)_{qj}$$

- Gradient descent:

$$H_{qj} \leftarrow H_{qj} + \eta_{qj} \left[(W^T A)_{qj} - (W^T W H)_{qj} \right]$$

- Diagonally rescale the variables using the step size:

$$\eta_{qj} = \frac{H_{qj}}{(W^T W H)_{qj}}$$

NMF / ANLS

- Two block coordinate descent as alternating non-negative least squares
- NNLS with multiple right hand sides

$$\min_{G \geq 0} \|BG - Y\|_F^2$$

- $B \in \mathbb{R}^{p \times q}$ and $Y \in \mathbb{R}^{p \times l}$
- Decouple into l independent NLS problems with single right hand side:

$$\min_{G \geq 0} \|BG - Y\|_F^2 \rightarrow \min_{\mathbf{g_1} \geq 0} \|B\mathbf{g_1} - \mathbf{y_1}\|_2^2, \ldots, \min_{\mathbf{g_l} \geq 0} \|B\mathbf{g_l} - \mathbf{y_l}\|_2^2$$

NMF / ANLS

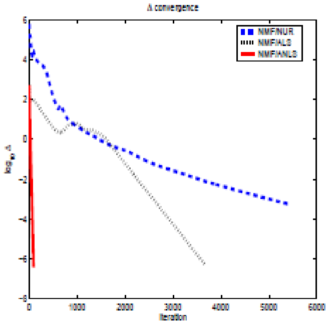
- Each NNLS problem

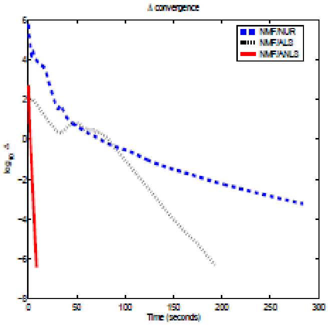
$$\min_{\mathbf{g_j} \geq 0} \|B\mathbf{g_j} - \mathbf{y_j}\|_2$$

  can be solved using the active set method.
- MATLAB function *lsqnonneg*
- Fast algorithms necessary
  - precompute cross-product terms
  - e.g. Pseudo-inverse for unconstrained LS
  - For NLS precompute $B^T B$ and $B^T Y$ and extract passive set components
  - Column parallel implementation $\rightarrow$ compute all passive sets

Performance

Algorithms	NMF/NUR	NMF/ALS	NMF/ANLS
$\#(W = 0)$ (%)	2.71%*	2.83%	2.71%
$\#(H = 0)$ (%)	18.42%*	16.67%	18.42%
$\|A - WH\|_F / \|A\|_F$	0.5027	0.5032	0.5027
# of iterations	5385	3670	90
Computing time	284.0 sec.	192.8 sec.	8.3 sec.





RNMF / ANLS

- Base matrix is often assumed to be full rank
- Regularized NMF:

$$\min_{W,H} \left(\|A - WH\|_F^2 + \alpha \|W\|_F^2 + \beta \|H\|_F^2 \right), s.t. W, H \geq 0$$

- For ANLS the iterations become:

$$\min_{W \geq 0} \left\| \begin{bmatrix} H^T \\ \sqrt{\alpha} I_k \end{bmatrix} W^T - \begin{bmatrix} A^T \\ 0_{k \times m} \end{bmatrix} \right\|_F^2$$

$$\min_{H \geq 0} \left\| \begin{bmatrix} W \\ \sqrt{\beta} I_k \end{bmatrix} H - \begin{bmatrix} A \\ 0_{k \times n} \end{bmatrix} \right\|_F^2$$

- $\alpha \geq 0$ and $\beta \geq 0$
- Imposes strong convexity on the subproblems
- Utilized in fast NNLS algorithms

SNMF / ANLS

- NMF often generates sparse solutions
- Sparseness cannot generally be assumed
- Impose L_1-norm based constraint

$$\min_{W,H} \left(\|A - WH\|_F^2 + \eta \|W\|_F^2 + \beta \sum_{j=1}^{n} \|\mathbf{h_j}\|_1^2 \right), s.t. W, H \geq 0$$

$\eta \geq 0$ suppresses growth of W
$\beta \geq 0$ balances trade-off between accuracy and sparseness

SNMF / ANLS

- For ANLS the iterations become:

$$\min_{W \geq 0} \left\| \begin{bmatrix} H^T \\ \sqrt{\eta} I_k \end{bmatrix} W^T - \begin{bmatrix} A^T \\ 0_{k \times m} \end{bmatrix} \right\|_F^2$$

$$\min_{H \geq 0} \left\| \begin{bmatrix} W \\ \sqrt{\beta} \mathbf{e_{1 \times k}} \end{bmatrix} H - \begin{bmatrix} A \\ 0_{1 \times n} \end{bmatrix} \right\|_F^2$$

- Does not necessarily improve the solution or interpretation

Outline

NMF / ANLS Implementation

- Implementation in MATLAB using *lsqnonneg*
- CBCL face image database
  - 2429 face images
  - 19 $\times$ 19 pixels
- Convergence criterium: $\|A - WH\|_F^2 / \|A\|_F^2 \leq \epsilon$

Result

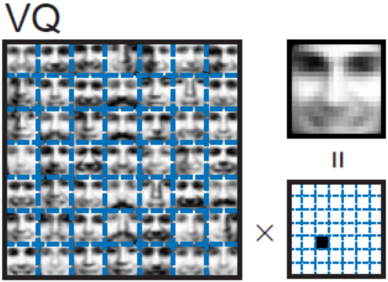
- Column vectors of the base matrix

Comparison to other Matrix Factorizations

- Daniel D. Lee and H. Sebastian Seung, "Learning the Parts of Objects by Nonnegative Matrix Factorization", 1999.
- Comparison of PCA, VQ and NMF
- NMF using norm-based multiplicative update rule

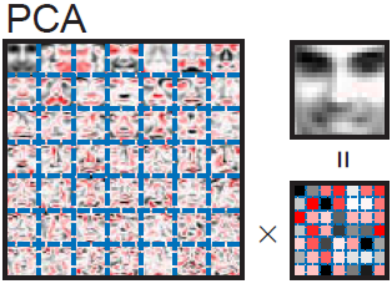
Factorization with VQ

- H constrained to be a unary vector
- Learns prototypical faces



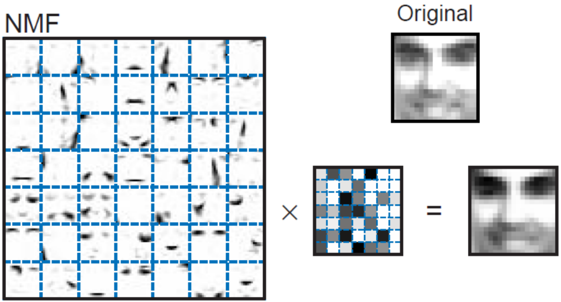
Factorization with PCA

- H constrained to have orthogonal rows
- Allows linear combination of "eigenfaces"



Factorization with NMF

- $H, W \geq 0$
- Allows only additive combinations of basis images



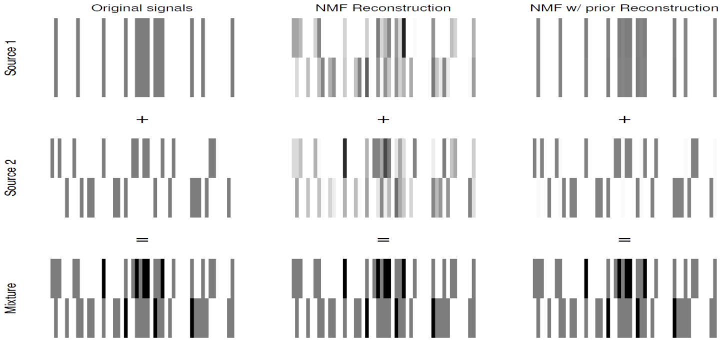
Result

- Column vectors of the base matrix

Non-Stationary Speech Denoising

- Remove non-stationary noise from speech using NMF
- Factorize the STFT using Divergence error function
- Base matrix represents "spectral building blocks"
- Training stage: build NMF for clean speech and clean noise
- Application stage: Estimate NMF on Training stage initialization
- Use "spectral building blocks" of former clean speech

Non-Stationary Speech Denoising



References

Daniel D. Lee and H. Sebastian Seung,
"Learning the Parts of Objects by Nonnegative Matrix Factorization,"
Nature, vol. 401, pp. 788-791, 1999.

Daniel D. Lee and H. Sebastian Seung,
"Algorithms for Non-negative Matrix Factorization,"
NIPS, 2001.

Lawson, C. and Hanson, R.,
"Solving Least Squares Problems,"
Prentice-Hall, 1974.

Kim, H. and Park, H.,
"Non-negative matrix Factorization Based on Alternating Non-negativity Constrained Least Squares and Active Set Method,"
SIAM Journal on Matrix Analysis and Applications, vol. 30, pp. 713-73, 2008.

Wilson, K.W.; Ray, B.; Smaragdis, P.; Divakaran A.,
"Speech denoising using nonnegative matrix factorization with priors,"
Acoustics, Speech and Signal Processing, 2008, pp. 4029-4032, 2008.