Gradient Projection Approaches for Nonnegative Matrix Factorization

Silvia Bonettini

University of Ferrara

The NMF problem

- Given
  - a data matrix $V \in \mathbb{R}^{n \times m}$
  - a positive integer $r < m$

- find $W \in \mathbb{R}^{n \times r}, H \in \mathbb{R}^{r \times m}$ such that

$$\min_{W \geq 0, H \geq 0} \frac{1}{2} \|WH - V\|_F^2$$

Applications

- Find a new basis for data representation
  - Data compression;
  - Classification/regression problems;
  - Blind source separation;
  - Text mining;
  - …
- Nonnegativity constraints yield a decomposition of data by parts [Lee&Seung, 1999][Donoho&Stodden 2004]

Finding parts of objects

- ORL face database (400 images, 112x92 pixels)

…

NMF with r=25

Other NMF formulations

- Kullback-Leibler divergence [Lee&Seung, 1999]

$$\min_{W_{ij}\geq 0, H_{hk}\geq 0} \sum_{ij} V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} + (WH)_{ij} - V_{ij}$$

- Tikhonov regularization [Pauca,Piper,Plemmons, 2006]

$$\min_{W\geq 0, H\geq 0} \frac{1}{2}\|WH - V\|_F^2 + \frac{\beta_W}{2}\|W\|_F^2 + \frac{\beta_H}{2}\|H\|_F^2$$

- Sparsity constraints [Hoyer, 2004]

- Normalization – flux conservation constraints [Lanteri *et al.*, 2010]

Features of the problem

- Nonlinear nonconvex in (W, H)

$$\min_{W \geq 0, H \geq 0} \frac{1}{2}\|WH - V\|_F^2 \equiv f(W, H)$$

- Convex quadratic(but non strictly convex) restriction to W or H

- Suggests a "natural" block decomposition

NMF algorithms: Multiplicative algorithms

[Lee&Seung, 1999]

$$W^{(k+1)} = W^{(k)} \frac{V H^{(k)^T}}{W^{(k)} H^{(k)} H^{(k)^T}}$$

One ISRA step to $f(\cdot, H^{(k)})$

$$H^{(k+1)} = H^{(k)} \frac{W^{(k+1)^T} V}{W^{(k+1)^T} W^{(k+1)} H^{(k)}}$$

One ISRA step to $f(W^{(k+1)}, \cdot)$

- Convergence: [Lin, 2007] (with hypotheses on V)
- Quite slow in practice.

Alternating Least Squares algorithms

$$W^{(k+1)} = \arg\min_{W \geq 0} f(W, H^{(k)})$$

$$H^{(k+1)} = \arg\min_{H \geq 0} f(W^{(k+1)}, H)$$

- Convergence: [Grippo&Sciandrone, 2000] (also for non strictly convex problems)
- Practical implementation: solve the two minimum problems by applying any iterative method. [Lin, 2007],[Zdunek, Cichocki, 2008], [Berry *et al.*, 2007]

Inexact ALS methods

- The practical ALS algorithms compute only approximate minima;
- The result of Grippo& Sciandrone holds for exact solutions;
- Does the ALS method with approximate minima still converge?

Cyclic Block Coordinate Gradient Projection method

[Bonettini, 2011]

$$W^{(k+1)} = W^{(k)} + \sum_{\ell=0}^{N_W} \lambda_W^{(k,\ell)} d_W^{(k,\ell)}$$

N_W Gradient Projection steps to $f(\cdot, H^{(k)})$

$$H^{(k+1)} = H^{(k)} + \sum_{\ell=0}^{N_H} \lambda_H^{(k,\ell)} d_H^{(k,\ell)}$$

N_H Gradient Projection steps to $f(W^{(k+1)}, \cdot)$

- Convergence: for any finite number of GP iterations
- Holds also for general nonlinear problems and for any ciclic block decomposition.

Gradient projection direction

$$W^{(k,\ell+1)} = W^{(k,\ell)} + \lambda_W^{(k,\ell)} d_W^{(k,\ell)}$$

Armijo (sufficient decrease of the objective function)

$$d_W^{(k,\ell)} = [W^{(k,\ell)} - \alpha^{(k,\ell)} D^{(k,\ell)} \nabla_W f(W^{(k,\ell)}, H^{(k)})]^+ - W^{(k,\ell)}$$

Steplength parameter (e.g. Barzilai-Borwein)

Scaling matrix (e.g. Split Gradient)

Evaluation of the results

- Two indices
  - Objective function value

$$f(W^{(k)}, H^{(k)})$$

  - Projected gradient norm

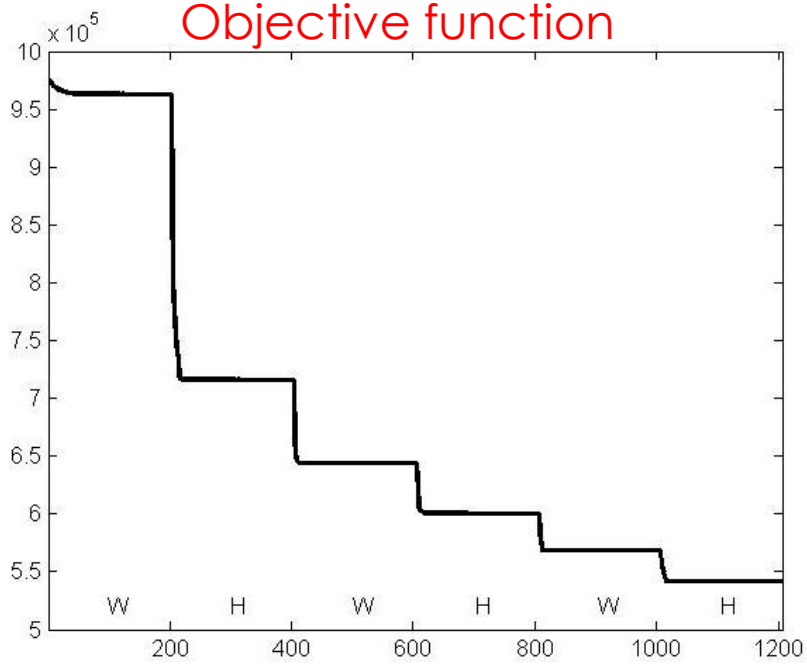
$$\nabla_W^P f(W^{(k)}, H^{(k)}) = [W^{(k)} - \nabla_W f(W^{(k)}, H^{(k)})]^+ - W^{(k)}$$

$$\nabla_H^P f(W^{(k)}, H^{(k)}) = [H^{(k)} - \nabla_H f(W^{(k)}, H^{(k)})]^+ - H^{(k)}$$

Numerical results

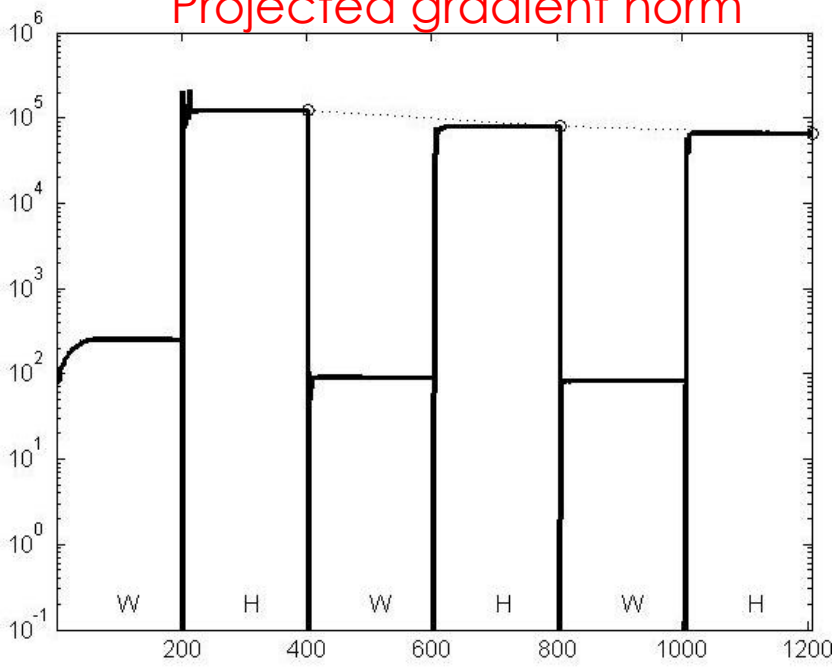
- After the first few GP iterations, no significant improvements in both objective function value and projected gradient norm.

Objective function



Inner iterations

Projected gradient norm



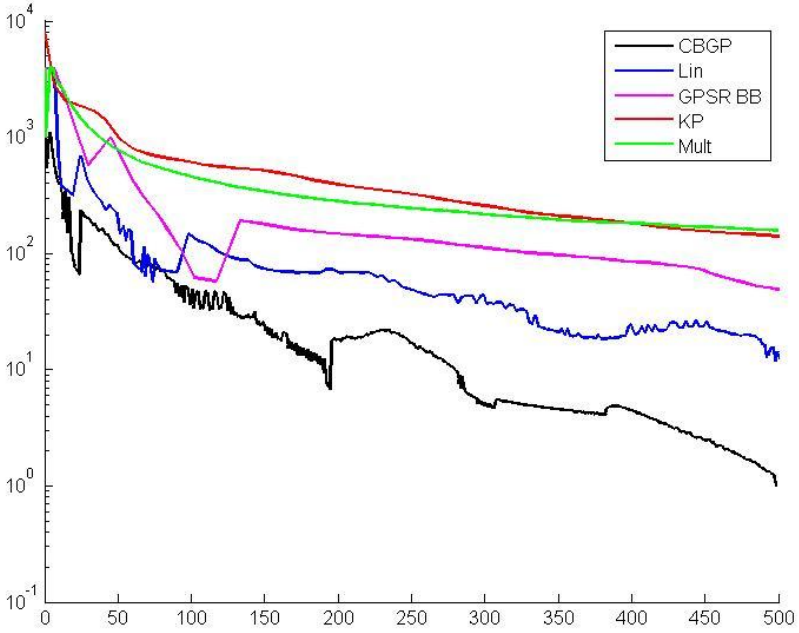
Inner iterations

CBGP algorithm

- Steplength choice by means of the Barzilai-Borwein rules
- Scaling matrix = Identity matrix
- Adaptive stopping criterion for the inner iterations

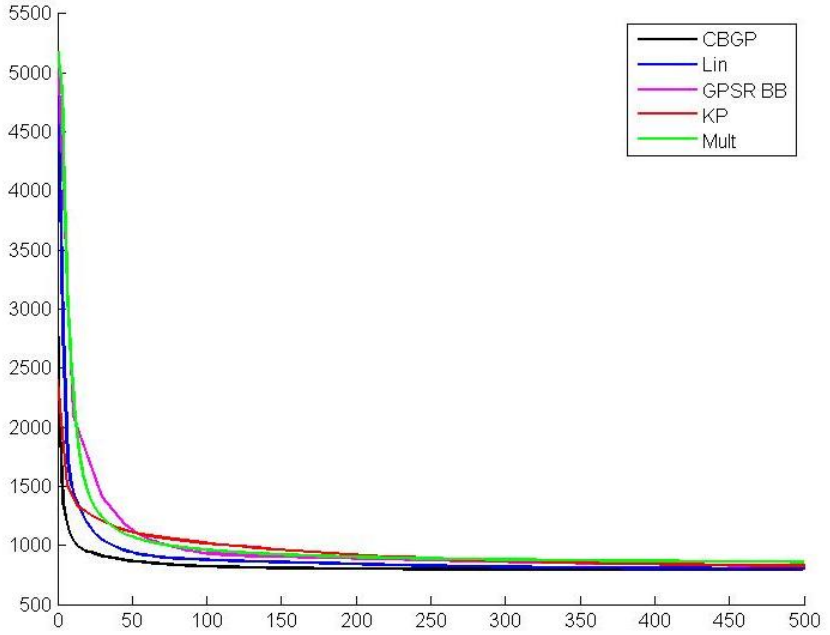
Comparison with other NMF algorithms

Projected gradient norm



Time (sec.)

Objective function value

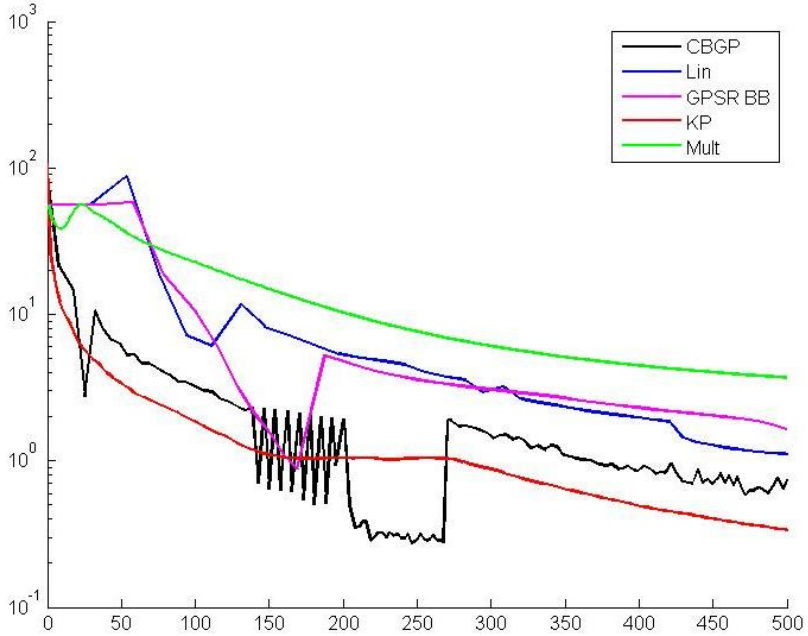


Time (sec.)

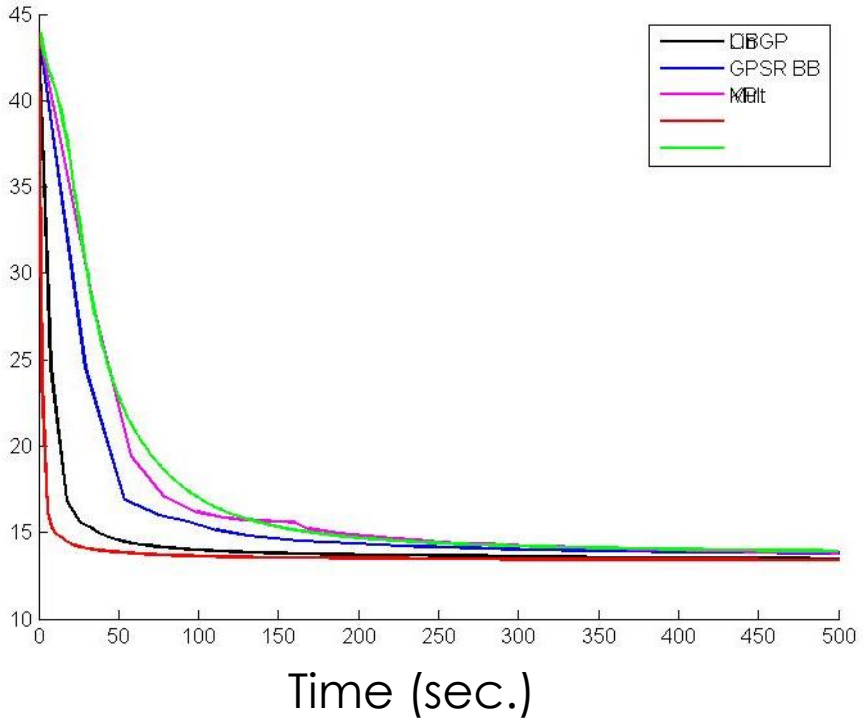
CBCL face database (n=361 m=2429 r=49)

Comparison with other NMF algorithms

Projected gradient norm



Objective function value



ORL face database (n=10304 m=400 r=25)

Future work

- On NMF
  - Consider other NMF formulations;
  - Investigate how inexact solutions affect the quality of the computed solution;
  - More study on the effect of scaling.
- Apply CBGP to signal processing problems whose formulation is similar to NMF
  - Blind deconvolution?