Penalty Hyperparameters Optimization in Non-negative Matrix Factorization problems

Laura Selicato
laura.selicato@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO
MIδAS Research group

Introduction

- Hyperparameters (HPs) direct impact on the performance of any algorithms and its reproducibility, especially in the context of unsupervised learning.
- How to automatically choose optimal penalization HPs in Non-negative Matrix Factorization (NMF)?
Bi-level approach: the selection of HPs is incorporated directly into the algorithm as part of the updating process.
- Proposal: a new algorithm **AltBi** for tuning penalization HPs in NMF problems.

AltBi

Algorithm 1: Alternate Bi-level

Data: $\mathbf{X} \in \mathbb{R}_+^{n\times m}$, factorization rank r.
Result: $\mathbf{W} \in \mathbb{R}_+^{n\times r}$, $\mathbf{H} \in \mathbb{R}_+^{r\times m}$, $\overline{\lambda^*}$.
Initial $\mathbf{W}^{(0)} \in \mathbb{R}_+^{n\times r}$, $\mathbf{H}^{(0)} \in \mathbb{R}_+^{m\times r}$,
$\mathbf{L}^{(0)} = diag(\overline{\lambda^{(0)}})$, T length of bunch ;
while *(err > tol) & (iter < MaxIter)* **do**
  $\mathbf{H} = \text{update}(\mathbf{X}, \mathbf{W}, \mathbf{H})$;
  for *t in T* **do**
    $(\mathbf{w}^{(t)}, \nabla_{\overline{\lambda^t}} f) =$
     $\text{bi-level}(\mathbf{X}, \nabla_{\overline{\lambda^{t-1}}} f, \mathbf{w}(\lambda)^{(t-1)}, \mathbf{H})$;
  end
  $\overline{\lambda^{\text{iter}}} = \text{update}(\overline{\lambda^{\text{iter}-1}}, \nabla_{\overline{\lambda^{\text{iter}-1}}} f)$;
  iter+ = 1;
end

Datasets

Model :

$$\mathbf{X} \approx \mathbf{Y} = \mathbf{WH} \quad \text{with} \quad \mathbf{W} \in \mathbb{R}_+^{n\times r}, \mathbf{H} \in \mathbb{R}_+^{r\times m}.$$

- Artificial datasets:
  1) factor $\mathbf{W}$ and $\mathbf{H}$ were generated randomly as full rank uniform distributed matrices.
  2) each column in $\mathbf{W}$ is expressed as real sinusoidal wave signal. $\mathbf{H}$ was generated as full rank sparse matrix, sparseness level α_H.
- Source signals taken from the file AC10_art_spectr_noi of MATLAB toolbox NMFLAB for Signal Processing.
- Real reflectance signals taken from the U.S. Geological Survey (USGS) database.

Conclusion

Novelty of our HPO proposal is including the minimization of the penalization HP into the optimization problem in a bi-level fashion. Results on existence and convergence of solution to the considered tasks are also demonstrated; numerical experiments and comparisons are also promising.

All the experiments confirmed the expected behaviour of AltBi in term of an identification problem. The SIR statistics for estimating the spectral signatures in the matrix $\mathbf{W}$ and the abundance maps in matrix $\mathbf{H}$, obtained with AltBi are significantly better than those obtained with MU and MU-P. Also in terms of Response and Loss function performance and in terms of added sparsity.

Mathematics of our proposal method

Problem :

$$\min_{\mathbf{H}\geq 0, \mathbf{W}\geq 0} D_\beta(\mathbf{X}, \mathbf{WH}) + \mathscr{R}(\mathbf{LW}). \tag{1}$$

$\mathbf{L} = \text{diag}(\overline{\lambda}) \in \mathbb{R}^{n\times n}$ is diagonal matrix of HPs associated with each row $\mathbf{w} \in \mathbb{R}^r$ of $\mathbf{W}$;
$\mathscr{R} : \mathbb{R}^{n\times r} \to \mathbb{R}$ is the penalization functions.

- $\mathbf{W}$ is fixed to estimate $\mathbf{H}$;
- $\mathbf{H}$ is fixed to estimate $\mathbf{W}$, incorporating the choice of HPs into the updating process.

Bi-level task on $\mathbf{w} \in \mathbb{R}^r$
$\min\{f(\lambda) : \lambda \in \Lambda\}$
$f(\lambda) = \inf\{\mathscr{E}(\mathbf{w}_{(\lambda)}, \lambda) : \mathbf{w}_{(\lambda)} \in \underset{\mathbf{u}\in\mathbb{R}^r}{\text{argmin}}\, \mathscr{L}_\lambda(\mathbf{u})\}$ is the *Response function*.

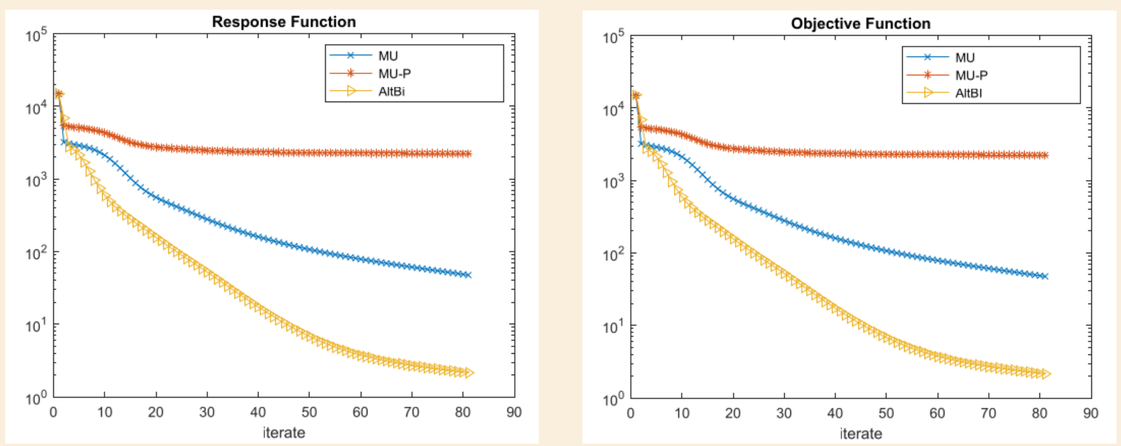
The *Error and Loss functions* are $\mathscr{E} : \mathbb{R}^r \times \Lambda \to \mathbb{R} : (\mathbf{w}, \lambda) \mapsto \sum_{j=1}^{m} d_\beta(X_{j:}, \sum_{k=1}^{r} w_k(\lambda) H_{kj})$ and

$\mathscr{L}_\lambda : \mathbb{R}^r \to \mathbb{R} : \mathbf{w} \mapsto \sum_{j=1}^{m} d_\beta(X_{j:}, \sum_{k=1}^{r} w_k H_{kj}) + \lambda \varkappa(\mathbf{w})$, with $\varkappa : \mathbb{R}^r \to \mathbb{R}$ s. t. $\sum_{i=1}^{n} \lambda \varkappa(\mathbf{w}) = \mathscr{R}(\mathbf{LW})$.

- The bi-level problem verifies the existence and convergence theorems under certain assumptions.
- The optimization for $\overline{\lambda}$ comes from the estimation of the gradient $\nabla_{\overline{\lambda}} f$, called Hypergradient

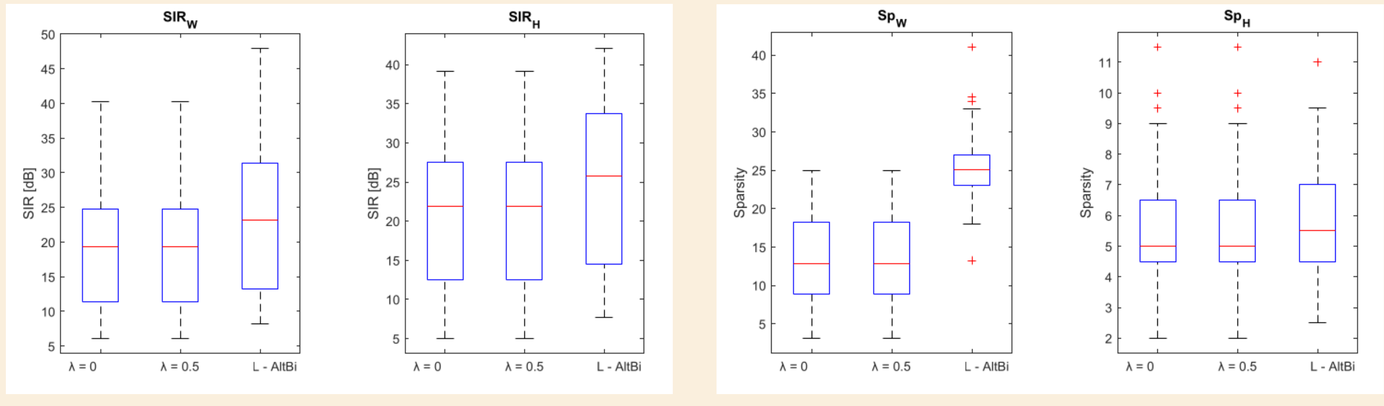
Results on real dataset

- All tests performed show similar behaviors compared with the standard unpenalized Multiplicative Update (MU) and the standard penalized one (P-MU); no noisy perturbations were used:
  1) Benchmark A was used with $n = 1000$, $m = 50$, $r = 4$.
  2) Benchmark B was used $n = 1000$, $m = 50$, $r = 4$, $\alpha_H = 0.1$.
  3) Benchmark C was used with $n = 1000$, $m = 50$, $r = 5$, $\alpha_H = 0.1$.
  4) Benchmark D was used with $n = 224$, $m = 3025$, $r = 5$.



(a) **(b)**

Figure 1: Evolution of Response (*left*) and Objective (*right*) functions w.r.t. iterations



(a) SIR statistics estimating $W_{i,:}$ and $H_{:,j}$ **(b)** Statistics of the sparseness measure

Acknowledgements

Work was supported in part by the GNCS-INDAM (Gruppo Nazionale per il Calcolo Scientifico of Istituto Nazionale di Alta Matematica) Francesco Severi.