SPAMS: a SPArse Modeling Software, v2.2

Julien Mairal
julien.mairal@m4x.org

March 23, 2012

1 Introduction

SPAMS (SPArse Modeling Software) is an open-source optimization toolbox under licence GPLv3. It implements algorithms for solving various machine learning and signal processing problems involving sparse regularizations.

The library is coded in C++, is compatible with Linux, Mac, and Windows 32bits and 64bits Operating Systems. It is interfaced with Matlab, but can be called from any C++ application. A R and Python interface has been developed by Jean-Paul Chieze.

It requires an implementation of BLAS and LAPACK for performing efficient linear algebra operations such as the one provided by matlab/R, atlas, netlib, or the one provided by Intel (Math Kernel Library). It also exploits multi-core CPUs when this feature is supported by the compiler, through OpenMP.

The current licence is GPLv3 available at `http://www.gnu.org/licenses/gpl.html`, which limits its usage. For other usages (such as the use in proprietary softwares), please contact the author.

Version 2.2 of SPAMS is divided into three main "toolboxes" and has a few additional miscellaneous functions:

- The **Dictionary learning and matrix factorization toolbox** contains the online learning technique of [18, 19] and its variants for solving various matrix factorization problems:
  - Dictionary Learning for sparse coding.
  - Sparse principal component analysis.
  - Non-negative matrix factorization.
  - Non-negative sparse coding.
- The **Sparse decomposition toolbox** contains efficient implementations of
  - Orthogonal Matching Pursuit, (or Forward Selection) [28, 21].
  - The LARS/homotopy algorithm [8] (variants for solving Lasso and Elastic-Net problems).
  - A weighted version of LARS.
  - OMP and LARS when data come with a binary mask.
  - A coordinate-descent algorithm for ℓ_1-decomposition problems [10, 9, 29].
  - A greedy solver for simultaneous signal approximation as defined in [27, 26] (SOMP).

  - A solver for simulatneous signal approximation with ℓ_1/ℓ_2-regularization based on block-coordinate descent.
  - A homotopy method for the Fused-Lasso Signal Approximation as defined in [9] with the homotopy method presented in the appendix of [19].
  - A tool for projecting efficiently onto a few convex sets inducing sparsity such as the ℓ_1-ball using the method of [3, 16, 7], and Elastic-Net or Fused Lasso constraint sets as proposed in the appendix of [19].

- The **Proximal toolbox**: An implementation of proximal methods (ISTA and FISTA [1]) for solving a large class of sparse approximation problems with different combinations of loss and regularizations. One of the main features of this toolbox is to provide a robust stopping criterion based on *duality gaps* to control the quality of the optimization, whenever possible. It also handles sparse feature matrices for large-scale problems. The following regularizations are implemented:

  - Tikhonov regularization (squared ℓ_2-norm).
  - ℓ_1-norm, ℓ_2, ℓ_∞-norms.
  - Elastic-Net [31].
  - Fused Lasso [25].
  - Tree-structured sum of ℓ_2-norms (see [13, 14]).
  - Tree-structured sum of ℓ_∞-norms (see [13, 14]).
  - General sum of ℓ_∞-norms (see [20]).
  - Mixed ℓ_1/ℓ_2-norms on matrices [30, 23].
  - Mixed ℓ_1/ℓ_∞-norms on matrices [30, 23].
  - Mixed ℓ_1/ℓ_2-norms on matrices plus ℓ_1 [24].
  - Mixed ℓ_1/ℓ_∞-norms on matrices plus ℓ_1.
  - group-lasso with ℓ_2 or ℓ_∞-norms.
  - group-lasso+ℓ_1.
  - multi-task tree-structured sum of ℓ_∞-norms (see [20]).
  - trace norm.
  - ℓ_0 pseudo-norm (only with ISTA)
  - Tree-structured ℓ_0 (only with ISTA)
  - rank regularization for matrices (only with ISTA)

  All of these regularization functions can be used with the following losses

  - square loss.
  - square loss with missing observations.
  - logistic loss, weighted logistic loss.
  - multi-class logistic

  This toolbox can also enforce positivity constraints and handles sparse matrices.

- A few tools for performing linear algebra operations such as a conjugate gradient algorithm, and manipulating sparse matrices.

The toolbox was written by Julien Mairal when he was at INRIA, with the collaboration of Francis Bach (INRIA), Jean Ponce (Ecole Normale Supérieure), Guillermo Sapiro (University of Minnesota) and Rodolphe Jenatton (INRIA).

A R and Python interface has been written by Jean-Paul Chieze (INRIA), and a few contributors have helped us making compilation scripts for various platforms.

2 Installation

The toolbox used to come with pre-compiled binaries for various plateforms. Now the sources are available and you will have to compile it yourself. Pre-compiled binaries will still be available for Linux 64 bits plateforms.

The user has the choice of the BLAS library, but the Intel MKL is recommended for the best performance. Note that the builtin blas library of Matlab is a version of the Intel MKL (not the most recent one though).

The folder `doc` contains the documentation in pdf and html. `build/` contains the binary files, including the help for each command. `test_release` contains various matlab scripts which call the different functions of this toolbox.

The software package comes also with a script bash that has to be used to launch matlab for Linux and/or Mac OS versions. The install procedure is described in a file called `HOW_TO_INSTALL`.

3 Dictionary Learning and Matrix Factorization Toolbox

This is the section for dictionary learning and matrix factorization, corresponding to [18, 19].

3.1 Function mexTrainDL

This is the main function of the toolbox, implementing the learning algorithms of [19]. Given a training set $\mathbf{x}^1, \ldots,$. It aims at solving

$$\min_{\mathbf{D}\in\mathcal{C}} \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}^i} \Big(\frac{1}{2}\|\mathbf{x}^i - \mathbf{D}\boldsymbol{\alpha}^i\|_2^2 + \psi(\boldsymbol{\alpha}^i)\Big). \tag{1}$$

ψ is a sparsity-inducing regularizer and $\mathcal{C}$ is a constraint set for the dictionary. As shown in [19] and in the help file below, various combinations can be used for ψ and $\mathcal{C}$ for solving different matrix factorization problems. What is more, positivity constraints can be added to $\boldsymbol{\alpha}$ as well. The function admits several modes for choosing the optimization parameters, using the parameter-free strategy proposed in [18], or using the parameters t_0 and ρ presented in [19]. **Note that for problems of a reasonable size, and when ψ is the ℓ_1-norm, the function mexTrainDL_Memory can be faster but uses more memory.**

```
%
% Usage:   [D [model]]=mexTrainDL(X,param[,model]);
%          model is optional
%
% Name: mexTrainDL
%
% Description: mexTrainDL is an efficient implementation of the
%     dictionary learning technique presented in
%
%     "Online Learning for Matrix Factorization and Sparse Coding"
%     by Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro
```

```
%     arXiv:0908.0050
%
%     "Online Dictionary Learning for Sparse Coding"
%     by Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro
%     ICML 2009.
%
%     Note that if you use param.mode=1 or 2, if the training set has a
%     reasonable size and you have enough memory on your computer, you
%     should use mexTrainDL_Memory instead.
%
%
%     It addresses the dictionary learning problems
%        1) if param.mode=0
%     min_{D in C} (1/n) sum_{i=1}^n (1/2)||x_i-Dalpha_i||_2^2  s.t. ...
%                                                  ||alpha_i||_1 <= lambda
%        2) if param.mode=1
%     min_{D in C} (1/n) sum_{i=1}^n  ||alpha_i||_1  s.t.  ...
%                                          ||x_i-Dalpha_i||_2^2 <= lambda
%        3) if param.mode=2
%     min_{D in C} (1/n) sum_{i=1}^n (1/2)||x_i-Dalpha_i||_2^2 + ...
%                                 lambda||alpha_i||_1 + lambda_2||alpha_i||_2^2
%        4) if param.mode=3, the sparse coding is done with OMP
%     min_{D in C} (1/n) sum_{i=1}^n (1/2)||x_i-Dalpha_i||_2^2  s.t. ...
%                                                  ||alpha_i||_0 <= lambda
%        5) if param.mode=4, the sparse coding is done with OMP
%     min_{D in C} (1/n) sum_{i=1}^n  ||alpha_i||_0  s.t.  ...
%                                          ||x_i-Dalpha_i||_2^2 <= lambda
%
%%     C is a convex set verifying
%        1) if param.modeD=0
%           C={  D in Real^{m x p}  s.t.  forall j,  ||d_j||_2^2 <= 1 }
%        2) if param.modeD=1
%           C={  D in Real^{m x p}  s.t.  forall j,  ||d_j||_2^2 + ...
%                                                  gamma1||d_j||_1 <= 1 }
%        3) if param.modeD=2
%           C={  D in Real^{m x p}  s.t.  forall j,  ||d_j||_2^2 + ...
%                                 gamma1||d_j||_1 + gamma2 FL(d_j) <= 1 }
%        4) if param.modeD=3
%           C={  D in Real^{m x p}  s.t.  forall j,  (1-gamma1)||d_j||_2^2 + ...
%                                 gamma1||d_j||_1 <= 1 }

%
%     Potentially, n can be very large with this algorithm.
%
% Inputs: X:  double m x n matrix   (input signals)
%               m is the signal size
%               n is the number of signals to decompose
%         param: struct
%            param.D: (optional) double m x p matrix   (dictionary)
%              p is the number of elements in the dictionary
%              When D is not provided, the dictionary is initialized
%              with random elements from the training set.
%           param.lambda  (parameter)
%           param.lambda2  (optional, by default 0)
%           param.iter (number of iterations).  If a negative number is
%              provided it will perform the computation during the
%              corresponding number of seconds. For instance param.iter=-5
%              learns the dictionary during 5 seconds.
%            param.mode (optional, see above, by default 2)
%            param.posAlpha (optional, adds positivity constraints on the
```

```
%              coefficients, false by default, not compatible with
%              param.mode =3,4)
%            param.modeD (optional, see above, by default 0)
%            param.posD (optional, adds positivity constraints on the
%              dictionary, false by default, not compatible with
%              param.modeD=2)
%            param.gamma1 (optional parameter for param.modeD >= 1)
%            param.gamma2 (optional parameter for param.modeD = 2)
%            param.batchsize (optional, size of the minibatch, by default
%              512)
%            param.modeParam (optimization mode).
%              1) if param.modeParam=0, the optimization uses the
%                 parameter free strategy of the ICML paper
%              2) if param.modeParam=1, the optimization uses the
%                 parameters rho as in arXiv:0908.0050
%              3) if param.modeParam=2, the optimization uses exponential
%                 decay weights with updates of the form
%                 A_{t} <- rho A_{t-1} + alpha_t alpha_t^T
%            param.rho (optional) tuning parameter (see paper arXiv:0908.0050)
%            param.clean (optional, true by default. prunes
%              automatically the dictionary from unused elements).
%            param.numThreads (optional, number of threads for exploiting
%              multi-core / multi-cpus. By default, it takes the value -1,
%              which automatically selects all the available CPUs/cores).
%
% Output:
%         param.D: double m x p matrix   (dictionary)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%         - single precision setting
%
% Author: Julien Mairal, 2009
```

3.2 Function mexTrainDL_Memory

Memory-consuming version of mexTrainDL. This function is well adapted to small/medium-size problems: It requires storing all the coefficients $\boldsymbol{\alpha}$ and is therefore impractical for very large datasets. However, in many situations, one can afford this memory cost and it is better to use this method, which is faster than mexTrainDL. Note that unlike mexTrainDL this function does not allow warm-restart. _Memory.m

3.3 Function nmf

This function is an example on how to use the function mexTrainDL for the problem of non-negative matrix factorization formulated in [15]. Note that mexTrainDL can be replaced by mexTrainDL_Memory in this function for small or medium datasets.

```
%
% Usage:   [U [,V]]=nmf(X,param);
%
% Name: nmf
%
% Description: mexTrainDL is an efficient implementation of the
%     non-negative matrix factorization technique presented in
%
```

```
%     "Online Learning for Matrix Factorization and Sparse Coding"
%     by Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro
%     arXiv:0908.0050
%
%     "Online Dictionary Learning for Sparse Coding"
%     by Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro
%     ICML 2009.
%
%     Potentially, n can be very large with this algorithm.
%
% Inputs: X:  double m x n matrix   (input signals)
%               m is the signal size
%               n is the number of signals to decompose
%         param: struct
%            param.K (number of required factors)
%            param.iter (number of iterations).  If a negative number
%              is provided it will perform the computation during the
%              corresponding number of seconds. For instance param.iter=-5
%              learns the dictionary during 5 seconds.
%            param.batchsize (optional, size of the minibatch, by default
%               512)
%            param.modeParam (optimization mode).
%               1) if param.modeParam=0, the optimization uses the
%                  parameter free strategy of the ICML paper
%               2) if param.modeParam=1, the optimization uses the
%                  parameters rho as in arXiv:0908.0050
%               3) if param.modeParam=2, the optimization uses exponential
%                  decay weights with updates of the form
%                  A_{t} <- rho A_{t-1} + alpha_t alpha_t^T
%            param.rho (optional) tuning parameter (see paper
%              arXiv:0908.0050)
%            param.t0 (optional) tuning parameter (see paper
%              arXiv:0908.0050)
%            param.clean (optional, true by default. prunes automatically
%              the dictionary from unused elements).
%            param.batch (optional, false by default, use batch learning
%              instead of online learning)
%            param.numThreads (optional, number of threads for exploiting
%                 multi-core / multi-cpus. By default, it takes the value -1,
%                 which automatically selects all the available CPUs/cores).
%         model: struct (optional) learned model for "retraining" the data.
%
% Output:
%         U: double m x p matrix
%         V: double p x n matrix   (optional)
%         model: struct (optional) learned model to be used for
%           "retraining" the data.
%
% Author: Julien Mairal, 2009
function [U V] = nmf(X,param)

param.lambda=0;
param.mode=2;
param.posAlpha=1;
param.posD=1;
param.whiten=0;
U=mexTrainDL(X,param);
param.pos=1;
if nargout == 2
   if issparse(X) % todo allow sparse matrices X for mexLasso
```

```
      maxbatch=ceil(10000000/size(X,1));
      for jj = 1:maxbatch:size(X,2)
         indbatch=jj:min((jj+maxbatch-1),size(X,2));
         Xb=full(X(:,indbatch));
         V(:,indbatch)=mexLasso(Xb,U,param);
      end
   else
      V=mexLasso(X,U,param);
   end
end
```

3.4 Function nnsc

This function is an example on how to use the function mexTrainDL for the problem of non-negative sparse coding as defined in [12]. Note that mexTrainDL can be replaced by mexTrainDL_Memory in this function for small or medium datasets.

```
%
% Usage:   [U [,V]]=nnsc(X,param);
%
% Name: nmf
%
% Description: mexTrainDL is an efficient implementation of the
%     non-negative sparse coding technique presented in
%
%     "Online Learning for Matrix Factorization and Sparse Coding"
%     by Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro
%     arXiv:0908.0050
%
%     "Online Dictionary Learning for Sparse Coding"
%     by Julien Mairal, Francis Bach, Jean Ponce and Guillermo Sapiro
%     ICML 2009.
%
%     Potentially, n can be very large with this algorithm.
%
% Inputs: X:  double m x n matrix   (input signals)
%               m is the signal size
%               n is the number of signals to decompose
%         param: struct
%           param.K (number of required factors)
%           param.lambda (parameter)
%           param.iter (number of iterations).  If a negative number
%              is provided it will perform the computation during the
%              corresponding number of seconds. For instance param.iter=-5
%              learns the dictionary during 5 seconds.
%           param.batchsize (optional, size of the minibatch, by default
%              512)
%           param.modeParam (optimization mode).
%              1) if param.modeParam=0, the optimization uses the
%                 parameter free strategy of the ICML paper
%              2) if param.modeParam=1, the optimization uses the
%                 parameters rho as in arXiv:0908.0050
%              3) if param.modeParam=2, the optimization uses exponential
%                 decay weights with updates of the form
%                 A_{t} <- rho A_{t-1} + alpha_t alpha_t^T
%           param.rho (optional) tuning parameter (see paper
%              arXiv:0908.0050)
%           param.t0 (optional) tuning parameter (see paper
```

```
%              arXiv:0908.0050)
%           param.clean (optional, true by default. prunes automatically
%              the dictionary from unused elements).
%           param.batch (optional, false by default, use batch learning
%              instead of online learning)
%           param.numThreads (optional, number of threads for exploiting
%              multi-core / multi-cpus. By default, it takes the value -1,
%              which automatically selects all the available CPUs/cores).
%         model: struct (optional) learned model for "retraining" the data.
%
% Output:
%         U: double m x p matrix
%         V: double p x n matrix   (optional)
%         model: struct (optional) learned model to be used for
%            "retraining" the data.
%
% Author: Julien Mairal, 2009

[U V] = function nnsc(X,param)

param.mode=2;
param.posAlpha=1;
param.posD=1;
param.whiten=0;
U=mexTrainDL(X,param);
param.pos=1;
if nargout == 2
   V=mexLasso(X,U,param);
end
```

4 Sparse Decomposition Toolbox

This toolbox implements several algorithms for solving signal reconstruction problems. It is mostly adapted for solving a large number of small/medium scale problems, but can be also efficient sometimes with large scale ones.

4.1 Function mexOMP

This is a fast implementation of the Orthogonal Matching Pursuit algorithm (or forward selection) [21, 28]. Given a matrix of signals $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$ in $\mathbb{R}^{m \times n}$ and a dictionary $\mathbf{D} = [\mathbf{d}^1, \ldots, \mathbf{d}^p]$ in $\mathbb{R}^{m \times p}$, the algorithm computes a matrix $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n]$ in $\mathbb{R}^{p \times n}$, where for each column $\mathbf{x}$ of $\mathbf{X}$, it returns a coefficient vector $\boldsymbol{\alpha}$ which is an approximate solution of the following NP-hard problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_0 \leq L. \tag{2}$$

or

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_0 \quad \text{s.t.} \quad \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \varepsilon. \tag{3}$$

For efficienty reasons, the method first computes the covariance matrix $\mathbf{D}^T\mathbf{D}$, then for each signal, it computes $\mathbf{D}^T\mathbf{x}$ and performs the decomposition with a Cholesky-based algorithm (see [6] for instance).

Note that mexOMP can return the "greedy" regularization path if needed (see below):

```
%
% Usage:   A=mexOMP(X,D,param);
% or       [A path]=mexOMP(X,D,param);
%
% Name: mexOMP
%
% Description: mexOMP is an efficient implementation of the
%     Orthogonal Matching Pursuit algorithm. It is optimized
%     for solving a large number of small or medium-sized
%     decomposition problem (and not for a single large one).
%     It first computes the Gram matrix D'D and then perform
%     a Cholesky-based OMP of the input signals in parallel.
%     X=[x^1,...,x^n] is a matrix of signals, and it returns
%     a matrix A=[alpha^1,...,alpha^n] of coefficients.
%
%     it addresses for all columns x of X,
%         min_{alpha} ||alpha||_0  s.t  ||x-Dalpha||_2^2 <= eps
%         or
%         min_{alpha} ||x-Dalpha||_2^2  s.t. ||alpha||_0 <= L
%
%
% Inputs: X:  double m x n matrix   (input signals)
%            m is the signal size
%            n is the number of signals to decompose
%         D:  double m x p matrix   (dictionary)
%            p is the number of elements in the dictionary
%            All the columns of D should have unit-norm !
%         param: struct
%            param.L (maximum number of elements in each decomposition)
%            param.eps (threshold on the squared l2-norm of the residual
%            param.numThreads (optional, number of threads for exploiting
%            multi-core / multi-cpus. By default, it takes the value -1,
%            which automatically selects all the available CPUs/cores).
%
% Output: A: double sparse p x n matrix (output coefficients)
%         path (optional): double dense p x L matrix (regularization path of the first signal)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%      - single precision setting (even though the output alpha is double
%        precision)
%      - Passing an int32 vector of length n to param.L allows to provide
%        a different parameter L for each input signal x_i
%      - Passing a double vector of length n to param.eps allows to provide
%        a different parameter eps for each input signal x_i
%
% Author: Julien Mairal, 2009
```

4.2 Function mexOMPMask

This is a variant of mexOMP with the possibility of handling a binary mask. Given a binary mask $\mathbf{B} = [\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^n]$ in $\{0,1\}^{m \times n}$, it returns a matrix $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n]$ such that for every column $\mathbf{x}$ of $\mathbf{X}$, $\boldsymbol{\beta}$ of $\mathbf{B}$, it computes a column $\boldsymbol{\alpha}$ of $\mathbf{A}$ by addressing

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \| \operatorname{diag}(\boldsymbol{\beta})(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}) \|_2^2 \;\; \text{s.t.} \;\; \|\boldsymbol{\alpha}\|_0 \leq L, \tag{4}$$

or

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} \|\boldsymbol{\alpha}\|_0 \ \text{ s.t. } \ \|\operatorname{diag}(\boldsymbol{\beta})(\mathbf{x}-\mathbf{D}\boldsymbol{\alpha})\|_2^2 \leq \varepsilon\frac{\|\boldsymbol{\beta}\|_0}{m}, \tag{5}$$

where $\operatorname{diag}(\boldsymbol{\beta})$ is a diagonal matrix with the entries of $\boldsymbol{\beta}$ on the diagonal.

```
%
% Usage:   A=mexOMPMask(X,D,B,param);
% or       [A path]=mexOMPMask(X,D,B,param);
%
% Name: mexOMPMask
%
% Description: mexOMPMask is a variant of mexOMP that allow using
%     a binary mask B
%
%     for all columns x of X, and columns beta of B, it computes a column
%         alpha of A by addressing
%         min_{alpha} ||alpha||_0  s.t  ||diag(beta)*(x-Dalpha)||_2^2
%                                                          <= eps*||beta||_0/m
%         or
%         min_{alpha} ||diag(beta)*(x-Dalpha)||_2^2  s.t. ||alpha||_0 <= L
%
%
% Inputs: X:  double m x n matrix   (input signals)
%            m is the signal size
%            n is the number of signals to decompose
%         D:  double m x p matrix   (dictionary)
%            p is the number of elements in the dictionary
%            All the columns of D should have unit-norm !
%         B:  boolean m x n matrix   (mask)
%               p is the number of elements in the dictionary
%         param: struct
%            param.L (maximum number of elements in each decomposition)
%            param.eps (threshold on the squared l2-norm of the residual
%            param.numThreads (optional, number of threads for exploiting
%            multi-core / multi-cpus. By default, it takes the value -1,
%            which automatically selects all the available CPUs/cores).
%
% Output: A: double sparse p x n matrix (output coefficients)
%         path (optional): double dense p x L matrix
%                                    (regularization path of the first signal)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%      - single precision setting (even though the output alpha is double
%        precision)
%      - Passing an int32 vector of length n to param.L allows to provide
%        a different parameter L for each input signal x_i
%      - Passing a double vector of length n to param.eps allows to provide
%        a different parameter eps for each input signal x_i
%
% Author: Julien Mairal, 2010
```

4.3 Function mexLasso

This is a fast implementation of the LARS algorithm [8] (variant for solving the Lasso) for solving the Lasso or Elastic-Net. Given a matrix of signals $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$ in $\mathbb{R}^{m\times n}$ and a dictionary $\mathbf{D}$ in $\mathbb{R}^{m\times p}$, depending on the input parameters, the algorithm returns

a matrix of coefficients $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n]$ in $\mathbb{R}^{p \times n}$ such that for every column $\mathbf{x}$ of $\mathbf{X}$, the corresponding column $\boldsymbol{\alpha}$ of $\mathbf{A}$ is the solution of

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\alpha}\|_1 \leq \lambda, \tag{6}$$

or

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \lambda, \tag{7}$$

or

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\alpha}\|_2^2. \tag{8}$$

For efficiency reasons, the method first compute the covariance matrix $\mathbf{D}^T\mathbf{D}$, then for each signal, it computes $\mathbf{D}^T\mathbf{x}$ and performs the decomposition with a Cholesky-based algorithm (see [8] for instance). The implementation has also an option to add **positivity constraints** on the solutions $\boldsymbol{\alpha}$. When the solution is very sparse and the problem size is reasonable, this approach can be very efficient. Moreover, it gives the solution with an exact precision, and its performance does not depend on the correlation of the dictionary elements, except when the solution is not unique (the algorithm breaks in this case).

Note that mexLasso can return the whole regularization path of the first signal $\mathbf{x}_1$ and can handle implicitely the matrix $\mathbf{D}$ if the quantities $\mathbf{D}^T\mathbf{D}$ and $\mathbf{D}^T\mathbf{x}$ are passed as an argument, see below:

```
%
% Usage:  [A [path]]=mexLasso(X,D,param);
%  or:    [A [path]]=mexLasso(X,Q,q,param);
%
% Name: mexLasso
%
% Description: mexLasso is an efficient implementation of the
%     homotopy-LARS algorithm for solving the Lasso.
%
%     if the function is called this way [A [path]]=mexLasso(X,D,param),
%     it aims at addressing the following problems
%     for all columns x of X, it computes one column alpha of A
%     that solves
%       1) when param.mode=0
%         min_{alpha} ||x-Dalpha||_2^2 s.t. ||alpha||_1 <= lambda
%       2) when param.mode=1
%         min_{alpha} ||alpha||_1 s.t. ||x-Dalpha||_2^2 <= lambda
%       3) when param.mode=2
%         min_{alpha} 0.5||x-Dalpha||_2^2 + lambda||alpha||_1 +0.5 lambda2||alpha||_2^2
%
%     if the function is called this way [A [path]]=mexLasso(X,Q,q,param),
%     it solves the above optimisation problem, when Q=D'D and q=D'x.
%
%     Possibly, when param.pos=true, it solves the previous problems
%     with positivity constraints on the vectors alpha
%
% Inputs: X:  double m x n matrix   (input signals)
%               m is the signal size
%               n is the number of signals to decompose
%         D:  double m x p matrix   (dictionary)
%               p is the number of elements in the dictionary
%         param: struct
%               param.lambda  (parameter)
%               param.lambda2  (optional parameter for solving the Elastic-Net)
%                              for mode=0 and mode=1, it adds a ridge on the Gram Matrix
```

```
%               param.L (optional), maximum number of steps of the homotopy algorithm (can
%                        be used as a stopping criterion)
%               param.pos (optional, adds non-negativity constraints on the
%                 coefficients, false by default)
%               param.mode (see above, by default: 2)
%               param.numThreads (optional, number of threads for exploiting
%                 multi-core / multi-cpus. By default, it takes the value -1,
%                 which automatically selects all the available CPUs/cores).
%               param.cholesky (optional, default false),  choose between Cholesky
%                 implementation or one based on the matrix inversion Lemma
%               param.ols (optional, default false), perform an orthogonal projection
%                 before returning the solution.
%               param.max_length_path (optional) maximum length of the path.
%
% Output: A: double sparse p x n matrix (output coefficients)
%         path: optional,  returns the regularisation path for the first signal
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%         - single precision setting (even though the output alpha is double
%           precision)
%
% Author: Julien Mairal, 2009
```

4.4 Function mexLassoWeighted

This is a fast implementation of a weighted version of LARS [8]. Given a matrix of signals $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$ in $\mathbb{R}^{m \times n}$, a matrix of weights $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^n] \in \mathbb{R}^{p \times n}$, and a dictionary $\mathbf{D}$ in $\mathbb{R}^{m \times p}$, depending on the input parameters, the algorithm returns a matrix of coefficients $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n]$ in $\mathbb{R}^{p \times n}$, such that for every column $\mathbf{x}$ of $\mathbf{X}$, $\mathbf{w}$ of $\mathbf{W}$, it computes a column $\boldsymbol{\alpha}$ of $\mathbf{A}$, which is the solution of

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad \text{s.t.} \quad \|\operatorname{diag}(\mathbf{w})\boldsymbol{\alpha}\|_1 \leq \lambda, \tag{9}$$

or

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\operatorname{diag}(\mathbf{w})\boldsymbol{\alpha}\|_1 \quad \text{s.t.} \quad \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 \leq \lambda, \tag{10}$$

or

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda \|\operatorname{diag}(\mathbf{w})\boldsymbol{\alpha}\|_1. \tag{11}$$

The implementation has also an option to add **positivity constraints** on the solutions $\boldsymbol{\alpha}$. This function is potentially useful for implementing efficiently the randomized Lasso of [22], or reweighted-ℓ_1 schemes [4].

```
%
% Usage:  A=mexLassoWeighted(X,D,W,param);
%
% Name: mexLassoWeighted.
%
% WARNING: This function has not been tested intensively
%
% Description: mexLassoWeighted is an efficient implementation of the
%     LARS algorithm for solving the weighted Lasso. It is optimized
%     for solving a large number of small or medium-sized
%     decomposition problem (and not for a single large one).
%     It first computes the Gram matrix D'D and then perform
```

```
%     a Cholesky-based OMP of the input signals in parallel.
%     For all columns x of X, and w of W, it computes one column alpha of A
%     which is the solution of
%       1) when param.mode=0
%         min_{alpha} ||x-Dalpha||_2^2   s.t.
%                                    ||diag(w)alpha||_1 <= lambda
%       2) when param.mode=1
%         min_{alpha} ||diag(w)alpha||_1  s.t.
%                                      ||x-Dalpha||_2^2 <= lambda
%       3) when param.mode=2
%         min_{alpha} 0.5||x-Dalpha||_2^2  +
%                                       lambda||diag(w)alpha||_1
%     Possibly, when param.pos=true, it solves the previous problems
%     with positivity constraints on the vectors alpha
%
% Inputs: X:  double m x n matrix   (input signals)
%               m is the signal size
%               n is the number of signals to decompose
%         D:  double m x p matrix   (dictionary)
%               p is the number of elements in the dictionary
%         W:  double p x n matrix   (weights)
%         param: struct
%            param.lambda  (parameter)
%            param.L (optional, maximum number of elements of each
%            decomposition)
%            param.pos (optional, adds positivity constraints on the
%            coefficients, false by default)
%            param.mode (see above, by default: 2)
%            param.numThreads (optional, number of threads for exploiting
%            multi-core / multi-cpus. By default, it takes the value -1,
%            which automatically selects all the available CPUs/cores).
%
% Output:  A: double sparse p x n matrix (output coefficients)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%         - single precision setting (even though the output alpha is double
%           precision)
%
% Author: Julien Mairal, 2009
```

4.5 Function mexLassoMask

This is a variant of mexLasso with the possibility of adding a mask $\mathbf{B} = [\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^n]$, as in mexOMPMask. For every column $\mathbf{x}$ of $\mathbf{X}$, $\boldsymbol{\beta}$ of $\mathbf{B}$, it computes a column $\boldsymbol{\alpha}$ of $\mathbf{A}$, which is the solution of

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \| \operatorname{diag}(\boldsymbol{\beta})(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha})\|_2^2 \;\; \text{s.t.} \;\; \|\boldsymbol{\alpha}\|_1 \leq \lambda, \tag{12}$$

or

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \|\boldsymbol{\alpha}\|_1 \;\; \text{s.t.} \;\; \| \operatorname{diag}(\boldsymbol{\beta})(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha})\|_2^2 \leq \lambda \frac{\|\boldsymbol{\beta}\|_0}{m}, \tag{13}$$

or

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2} \| \operatorname{diag}(\boldsymbol{\beta})(\mathbf{x} - \mathbf{D}\boldsymbol{\alpha})\|_2^2 + \lambda \frac{\|\boldsymbol{\beta}\|_0}{m} \|\boldsymbol{\alpha}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\alpha}\|_2^2. \tag{14}$$

```
%
% Usage:   A=mexLassoMask(X,D,B,param);
```

```
%
% Name: mexLassoMask
%
% Description: mexLasso is a variant of mexLasso that handles
%     binary masks. It aims at addressing the following problems
%     for all columns x of X, and beta of B, it computes one column alpha of A
%     that solves
%       1) when param.mode=0
%         min_{alpha} ||diag(beta)(x-Dalpha)||_2^2 s.t. ||alpha||_1 <= lambda
%       2) when param.mode=1
%         min_{alpha} ||alpha||_1 s.t. ||diag(beta)(x-Dalpha)||_2^2
%                                                          <= lambda*||beta||_0/m
%       3) when param.mode=2
%         min_{alpha} 0.5||diag(beta)(x-Dalpha)||_2^2 +
%                                              lambda*(||beta||_0/m)*||alpha||_1
%     Possibly, when param.pos=true, it solves the previous problems
%     with positivity constraints on the vectors alpha
%
% Inputs: X:  double m x n matrix   (input signals)
%               m is the signal size
%               n is the number of signals to decompose
%         D:  double m x p matrix   (dictionary)
%               p is the number of elements in the dictionary
%         B:  boolean m x n matrix   (mask)
%               p is the number of elements in the dictionary
%         param: struct
%               param.lambda  (parameter)
%               param.L (optional, maximum number of elements of each
%                 decomposition)
%               param.pos (optional, adds positivity constraints on the
%                 coefficients, false by default)
%               param.mode (see above, by default: 2)
%               param.numThreads (optional, number of threads for exploiting
%                 multi-core / multi-cpus. By default, it takes the value -1,
%                 which automatically selects all the available CPUs/cores).
%
% Output: A: double sparse p x n matrix (output coefficients)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%         - single precision setting (even though the output alpha is double
%           precision)
%
% Author: Julien Mairal, 2010
```

4.6 Function mexCD

Coordinate-descent approach for solving Eq. (8) and Eq. (7). Note that unlike mexLasso, it is not implemented to solve the Elastic-Net formulation. To solve Eq. (7), the algorithm solves a sequence of problems of the form (8) using simple heuristics. Coordinate descent is very simple and in practice very powerful. It performs better when the correlation between the dictionary elements is small.

```
%
% Usage:   A=mexCD(X,D,A0,param);
%
% Name: mexCD
```

```
%
% Description: mexCD addresses l1-decomposition problem with a
%     coordinate descent type of approach.
%     It is optimized for solving a large number of small or medium-sized
%     decomposition problem (and not for a single large one).
%     It first computes the Gram matrix D'D.
%     This method is particularly well adapted when there is low
%     correlation between the dictionary elements and when one can benefit
%     from a warm restart.
%     It aims at addressing the two following problems
%     for all columns x of X, it computes a column alpha of A such that
%       2) when param.mode=1
%         min_{alpha} ||alpha||_1 s.t. ||x-Dalpha||_2^2 <= lambda
%         For this constraint setting, the method solves a sequence of
%         penalized problems (corresponding to param.mode=2) and looks
%         for the corresponding Lagrange multplier with a simple but
%         efficient heuristic.
%       3) when param.mode=2
%         min_{alpha} 0.5||x-Dalpha||_2^2 + lambda||alpha||_1
%
% Inputs: X:  double m x n matrix   (input signals)
%                m is the signal size
%                n is the number of signals to decompose
%         D:  double m x p matrix   (dictionary)
%                p is the number of elements in the dictionary
%                All the columns of D should have unit-norm !
%         A0:  double sparse p x n matrix   (initial guess)
%         param: struct
%            param.lambda  (parameter)
%            param.mode (optional, see above, by default 2)
%            param.itermax (maximum number of iterations)
%            param.tol (tolerance parameter)
%            param.numThreads (optional, number of threads for exploiting
%            multi-core / multi-cpus. By default, it takes the value -1,
%            which automatically selects all the available CPUs/cores).
%
% Output: A: double sparse p x n matrix (output coefficients)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%         - single precision setting (even though the output alpha
%           is double precision)
%
% Author: Julien Mairal, 2009
```

4.7 Function mexSOMP

This is a fast implementation of the Simultaneous Orthogonal Matching Pursuit algorithm. Given a set of matrices $\mathbf{X} = [\mathbf{X}^1, \ldots, \mathbf{X}^n]$ in $\mathbb{R}^{m \times N}$, where the $\mathbf{X}^i$'s are in $\mathbb{R}^{m \times n_i}$, and a dictionary $\mathbf{D}$ in $\mathbb{R}^{m \times p}$, the algorithm returns a matrix of coefficients $\mathbf{A} = [\mathbf{A}^1, \ldots, \mathbf{A}^n]$ in $\mathbb{R}^{p \times N}$ which is an approximate solution of the following NP-hard problem

$$\forall i \quad \min_{\mathbf{A}^i \in \mathbb{R}^{p \times n_i}} \|\mathbf{X}^i - \mathbf{D}\mathbf{A}^i\|_F^2 \ \text{ s.t. } \ \|\mathbf{A}^i\|_{0,\infty} \leq L. \tag{15}$$

or

$$\forall i \quad \min_{\mathbf{A}^i \in \mathbb{R}^{p \times n_i}} \|\mathbf{A}^i\|_{0,\infty} \ \text{ s.t. } \ \|\mathbf{X}^i - \mathbf{D}\mathbf{A}^i\|_F^2 \leq \varepsilon n_i. \tag{16}$$

To be efficient, the method first compute the covariance matrix $\mathbf{D}^T\mathbf{D}$, then for each signal, it computes $\mathbf{D}^T\mathbf{X}^i$ and performs the decomposition with a Cholesky-based algorithm.

```
%
% Usage:   alpha=mexSOMP(X,D,list_groups,param);
%
% Name: mexSOMP
%     (this function has not been intensively tested).
%
% Description: mexSOMP is an efficient implementation of a
%     Simultaneous Orthogonal Matching Pursuit algorithm. It is optimized
%     for solving a large number of small or medium-sized
%     decomposition problem (and not for a single large one).
%     It first computes the Gram matrix D'D and then perform
%     a Cholesky-based OMP of the input signals in parallel.
%     It aims at addressing the following NP-hard problem
%
%     X is a matrix structured in groups of signals, which we denote
%     by X=[X_1,...,X_n]
%
%     for all matrices X_i of X,
%         min_{A_i} ||A_i||_{0,infty}  s.t  ||X_i-D A_i||_2^2 <= eps*n_i
%         where n_i is the number of columns of X_i
%
%         or
%
%         min_{A_i} ||X_i-D A_i||_2^2  s.t. ||A_i||_{0,infty} <= L
%
%
% Inputs: X:  double m x N matrix   (input signals)
%            m is the signal size
%            N is the total number of signals
%         D:  double m x p matrix   (dictionary)
%            p is the number of elements in the dictionary
%            All the columns of D should have unit-norm !
%         list_groups : int32 vector containing the indices (starting at 0)
%            of the first elements of each groups.
%         param: struct
%            param.L (maximum number of elements in each decomposition)
%            param.eps (threshold on the squared l2-norm of the residual
%            param.numThreads (optional, number of threads for exploiting
%            multi-core / multi-cpus. By default, it takes the value -1,
%            which automatically selects all the available CPUs/cores).
%
% Output: alpha: double sparse p x N matrix (output coefficients)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%      - single precision setting (even though the output alpha is double
%        precision)
%
% Author: Julien Mairal, 2010
```

4.8 Function mexL1L2BCD

This is a fast implementation of a simultaneous signal decomposition formulation. Given a set of matrices $\mathbf{X} = [\mathbf{X}^1, \ldots, \mathbf{X}^n]$ in $\mathbb{R}^{m \times N}$, where the $\mathbf{X}^i$'s are in $\mathbb{R}^{m \times n_i}$, and a dictionary

$\mathbf{D}$ in $\mathbb{R}^{m\times p}$, the algorithm returns a matrix of coefficients $\mathbf{A} = [\mathbf{A}^1, \ldots, \mathbf{A}^n]$ in $\mathbb{R}^{p\times N}$ which is an approximate solution of the following NP-hard problem

$$\forall i \quad \min_{\mathbf{A}^i \in \mathbb{R}^{p\times n_i}} \|\mathbf{X}^i - \mathbf{D}\mathbf{A}^i\|_F^2 \;\text{ s.t. }\; \|\mathbf{A}^i\|_{1,2} \leq \frac{\lambda}{n_i}. \tag{17}$$

or

$$\forall i \quad \min_{\mathbf{A}^i \in \mathbb{R}^{p\times n_i}} \|\mathbf{A}^i\|_{1,2} \;\text{ s.t. }\; \|\mathbf{X}^i - \mathbf{D}\mathbf{A}^i\|_F^2 \leq \lambda n_i. \tag{18}$$

To be efficient, the method first compute the covariance matrix $\mathbf{D}^T\mathbf{D}$, then for each signal, it computes $\mathbf{D}^T\mathbf{X}^i$ and performs the decomposition with a Cholesky-based algorithm.

```
%
% Usage:   alpha=mexL1L2BCD(X,D,alpha0,list_groups,param);
%
% Name: mexL1L2BCD
%     (this function has not been intensively tested).
%
% Description: mexL1L2BCD is a solver for a
%     Simultaneous signal decomposition formulation based on block
%     coordinate descent.
%
%     X is a matrix structured in groups of signals, which we denote
%     by X=[X_1,...,X_n]
%
%     if param.mode=2, it solves
%         for all matrices X_i of X,
%         min_{A_i} 0.5||X_i-D A_i||_2^2 + lambda/sqrt(n_i)||A_i||_{1,2}
%         where n_i is the number of columns of X_i
%     if param.mode=1, it solves
%         min_{A_i} ||A_i||_{1,2} s.t. ||X_i-D A_i||_2^2  <= n_i lambda
%
%
% Inputs: X:  double m x N matrix   (input signals)
%            m is the signal size
%            N is the total number of signals
%         D:  double m x p matrix   (dictionary)
%            p is the number of elements in the dictionary
%         alpha0: double dense p x N matrix (initial solution)
%         list_groups : int32 vector containing the indices (starting at 0)
%            of the first elements of each groups.
%         param: struct
%            param.lambda (regularization parameter)
%            param.mode (see above, by default 2)
%            param.itermax (maximum number of iterations, by default 100)
%            param.tol (tolerance parameter, by default 0.001)
%            param.numThreads (optional, number of threads for exploiting
%            multi-core / multi-cpus. By default, it takes the value -1,
%            which automatically selects all the available CPUs/cores).
%
% Output: alpha: double sparse p x N matrix (output coefficients)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%      - single precision setting (even though the output alpha is double
%        precision)
%
% Author: Julien Mairal, 2010
```

4.9 Function mexSparseProject

This is a multi-purpose function, implementing fast algorithms for projecting on convex sets, but it also solves the fused lasso signal approximation problem. The proposed method is detailed in [19]. The main problems addressed by this function are the following: Given a matrix $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ in $\mathbb{R}^{m \times n}$, it finds a matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n]$ in $\mathbb{R}^{m \times n}$ so that for all column $\mathbf{u}$ of $\mathbf{U}$, it computes a column $\mathbf{v}$ of $\mathbf{V}$ solving

$$\min_{\mathbf{v} \in \mathbb{R}^m} \|\mathbf{u} - \mathbf{v}\|_2^2 \;\; \text{s.t.} \;\; \|\mathbf{v}\|_1 \leq \tau, \tag{19}$$

or

$$\min_{\mathbf{v} \in \mathbb{R}^m} \|\mathbf{u} - \mathbf{v}\|_2^2 \;\; \text{s.t.} \;\; \lambda_1 \|\mathbf{v}\|_1 + \lambda_2 \|\mathbf{v}\|_2^2 \leq \tau, \tag{20}$$

or

$$\min_{\mathbf{v} \in \mathbb{R}^m} \|\mathbf{u} - \mathbf{v}\|_2^2 \;\; \text{s.t.} \;\; \lambda_1 \|\mathbf{v}\|_1 + \lambda_2 \|\mathbf{v}\|_2^2 + \lambda_3 FL(\mathbf{v}) \leq \tau, \tag{21}$$

or

$$\min_{\mathbf{v} \in \mathbb{R}^m} \frac{1}{2} \|\mathbf{u} - \mathbf{v}\|_2^2 + \lambda_1 \|\mathbf{v}\|_1 + \lambda_2 \|\mathbf{v}\|_2^2 + \lambda_3 FL(\mathbf{v}). \tag{22}$$

Note that for the two last cases, the method performs a small approximation. The method follows the regularization path, goes from one kink to another, and stop whenever the constraint is not satisfied anymore. The solution returned by the algorithm is the one obtained at the last kink of the regularization path, which is in practice close, but not exactly the same as the solution. This will be corrected in a future release of the toolbox.

```
%
% Usage:  V=mexSparseProject(U,param);
%
% Name: mexSparseProject
%
% Description: mexSparseProject solves various optimization
%     problems, including projections on a few convex sets.
%     It aims at addressing the following problems
%     for all columns u of U in parallel
%       1) when param.mode=1 (projection on the l1-ball)
%           min_v ||u-v||_2^2  s.t.  ||v||_1 <= thrs
%       2) when param.mode=2
%           min_v ||u-v||_2^2  s.t. ||v||_2^2 + lamuda1||v||_1 <= thrs
%       3) when param.mode=3
%           min_v ||u-v||_2^2  s.t  ||v||_1 + 0.5lamuda1||v||_2^2 <= thrs
%       4) when param.mode=4
%           min_v 0.5||u-v||_2^2 + lamuda1||v||_1  s.t  ||v||_2^2 <= thrs
%       5) when param.mode=5
%           min_v 0.5||u-v||_2^2 + lamuda1||v||_1 +lamuda2 FL(v) + ...
%                                                 0.5lamuda_3 ||v||_2^2
%          where FL denotes a "fused lasso" regularization term.
%       6) when param.mode=6
%          min_v ||u-v||_2^2 s.t lamuda1||v||_1 +lamuda2 FL(v) + ...
%                                           0.5lamuda3||v||_2^2 <= thrs
%
%        When param.pos=true and param.mode <= 4,
%        it solves the previous problems with positivity constraints
%
% Inputs: U:  double m x n matrix   (input signals)
%               m is the signal size
%               n is the number of signals to project
%         param: struct
%           param.thrs (parameter)
```

```
%           param.lambda1 (parameter)
%           param.lambda2 (parameter)
%           param.lambda3 (parameter)
%           param.mode (see above)
%           param.pos (optional, false by default)
%           param.numThreads (optional, number of threads for exploiting
%             multi-core / multi-cpus. By default, it takes the value -1,
%             which automatically selects all the available CPUs/cores).
%
% Output: V: double m x n matrix (output matrix)
%
% Note: this function admits a few experimental usages, which have not
%     been extensively tested:
%         - single precision setting
%
% Author: Julien Mairal, 2009
```

5 Proximal Toolbox

The previous toolbox we have presented is well adapted for solving a large number of small and medium-scale sparse decomposition problems with the square loss, which is typical from the classical dictionary learning framework. We now present a new software package that is adapted for solving a wide range of possibly large-scale learning problems, with several combinations of losses and regularization terms. The method implements the proximal methods of [1], and includes the proximal solvers for the tree-structured regularization of [13], and the solver of [20] for general structured sparse regularization. The solver for structured sparse regularization norms includes a C++ max-flow implementation of the push-relabel algorithm of [11], with heuristics proposed by [5].

This implementation also provides robust stopping criteria based on *duality gaps*, which are presented in Appendix A. It can handle intercepts (unregularized variables). The general formulation that our software can solve take the form

$$\min_{\mathbf{w}\in\mathbb{R}^p} [g(\mathbf{w}) \triangleq f(\mathbf{w}) + \lambda\psi(\mathbf{w})],$$

where f is a smooth loss function and ψ is a regularization function. When one optimizes a matrix $\mathbf{W}$ in $\mathbb{R}^{p\times r}$ instead of a vector $\mathbf{w}$ in $\mathbb{R}^p$, we will write

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}} [g(\mathbf{W}) \triangleq f(\mathbf{W}) + \lambda\psi(\mathbf{W})].$$

Note that the software can possibly handle nonnegativity constraints.

We start by presenting the type of regularization implemented in the software

5.1 Regularization Functions

Our software can handle the following regularization functions ψ for vectors $\mathbf{w}$ in $\mathbb{R}^p$:

- **The Tikhonov regularization**: $\psi(\mathbf{w}) \triangleq \frac{1}{2}\|\mathbf{w}\|_2^2$.
- **The ℓ_1-norm**: $\psi(\mathbf{w}) \triangleq \|\mathbf{w}\|_1$.
- **The Elastic-Net**: $\psi(\mathbf{w}) \triangleq \|\mathbf{w}\|_1 + \gamma\|\mathbf{w}\|_2^2$.

- **The Fused-Lasso**: $\psi(\mathbf{w}) \stackrel{\Delta}{=} \|\mathbf{w}\|_1 + \gamma\|\mathbf{w}\|_2^2 + \gamma_2 \sum_{i=1}^{p-1} |\mathbf{w}_{i+1} - \mathbf{w}_i|$.
- **The group Lasso**: $\psi(\mathbf{w}) \stackrel{\Delta}{=} \sum_{g\in\mathcal{G}} \eta_g \|\mathbf{w}_g\|_2$, where $\mathcal{G}$ are groups of variables of the same size.
- **The group Lasso with ℓ_∞-norm**: $\psi(\mathbf{w}) \stackrel{\Delta}{=} \sum_{g\in\mathcal{G}} \eta_g \|\mathbf{w}_g\|_\infty$, where $\mathcal{G}$ are groups of variables of the same size.
- **The sparse group Lasso**: same as above but with an additional ℓ_1 term.
- **The tree-structured sum of ℓ_2-norms**: $\psi(\mathbf{w}) \stackrel{\Delta}{=} \sum_{g\in\mathcal{G}} \eta_g \|\mathbf{w}_g\|_2$, where $\mathcal{G}$ is a tree-structured set of groups [13], and the η_g are positive weights.
- **The tree-structured sum of ℓ_∞-norms**: $\psi(\mathbf{w}) \stackrel{\Delta}{=} \sum_{g\in\mathcal{G}} \eta_g \|\mathbf{w}_g\|_\infty$. See [13]
- **General sum of ℓ_∞-norms**: $\psi(\mathbf{w}) \stackrel{\Delta}{=} \sum_{g\in\mathcal{G}} \eta_g \|\mathbf{w}_g\|_\infty$, where no assumption are made on the groups $\mathcal{G}$.

Our software also handles regularization functions ψ on matrices $\mathbf{W}$ in $\mathbb{R}^{p\times r}$ (note that $\mathbf{W}$ can be transposed in these formulations). In particular,

- **The ℓ_1/ℓ_2-norm**: $\psi(\mathbf{W}) \stackrel{\Delta}{=} \sum_{i=1}^p \|\mathbf{W}_i\|_2$, where $\mathbf{W}_i$ denotes the i-th row of $\mathbf{W}$.
- **The ℓ_1/ℓ_∞-norm**: $\psi(\mathbf{W}) \stackrel{\Delta}{=} \sum_{i=1}^p \|\mathbf{W}_i\|_\infty$,
- **The ℓ_1/ℓ_2+ℓ_1-norm**: $\psi(\mathbf{W}) \stackrel{\Delta}{=} \sum_{i=1}^p \|\mathbf{W}_i\|_2 + \lambda_2 \sum_{i,j} |\mathbf{W}_{ij}|$.
- **The ℓ_1/ℓ_∞+ℓ_1-norm**: $\psi(\mathbf{W}) \stackrel{\Delta}{=} \sum_{i=1}^p \|\mathbf{W}_i\|_\infty + \lambda_2 \sum_{i,j} |\mathbf{W}_{ij}|$,
- **The ℓ_1/ℓ_∞-norm on rows and columns**: $\psi(\mathbf{W}) \stackrel{\Delta}{=} \sum_{i=1}^p \|\mathbf{W}_i\|_\infty + \lambda_2 \sum_{j=1}^r \|\mathbf{W}^j\|_\infty$, where $\mathbf{W}^j$ denotes the j-th column of $\mathbf{W}$.
- **The multi-task tree-structured sum of ℓ_∞-norms**:

$$\psi(\mathbf{W}) \stackrel{\Delta}{=} \sum_{i=1}^r \sum_{g\in\mathcal{G}} \eta_g \|\mathbf{w}_g^i\|_\infty + \gamma \sum_{g\in\mathcal{G}} \eta_g \max_{j\in g} \|\mathbf{W}_j\|_\infty, \tag{23}$$

  where the first double sums is in fact a sum of independent structured norms on the columns $\mathbf{w}^i$ of $\mathbf{W}$, and the right term is a tree-structured regularization norm applied to the ℓ_∞-norm of the rows of $\mathbf{W}$, thereby inducing the tree-structured regularization at the row level. $\mathcal{G}$ is here a tree-structured set of groups.
- **The multi-task general sum of ℓ_∞-norms** is the same as Eq. (23) except that the groups $\mathcal{G}$ are general overlapping groups.
- **The trace norm**: $\psi(\mathbf{W}) \stackrel{\Delta}{=} \|\mathbf{W}\|_*$.

Non-convex regularizations are also implemented with the ISTA algorithm (no duality gaps are of course provided in these cases):

- **The ℓ_0-pseudo-norm**: $\psi(\mathbf{w}) \stackrel{\Delta}{=} \|\mathbf{w}\|_0$.
- **The rank**: $\psi(\mathbf{W}) \stackrel{\Delta}{=} \text{randk}(\mathbf{W})$.

- **The tree-structured ℓ_0-pseudo-norm**: $\psi(\mathbf{w}) \triangleq \sum_{g\in\mathcal{G}} \delta_{\mathbf{w}_g\neq 0}$.

All of these regularization terms for vectors or matrices can be coupled with nonnegativity constraints. It is also possible to add an intercept, which one wishes not to regularize, and we will include this possibility in the next sections. There are also a few hidden undocumented options which are available in the source code.

We now present 3 functions for computing proximal operators associated to the previous regularization functions.

5.2 Function mexProximalFlat

This function computes the proximal operators associated to many regularization functions, for input signals $\mathbf{U} = [\mathbf{u}^1, \ldots, \mathbf{u}^n]$ in $\mathbb{R}^{p\times n}$, it finds a matrix $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^n]$ in $\mathbb{R}^{p\times n}$ such that:

- If one chooses a regularization function on vectors, for every column $\mathbf{u}$ of $\mathbf{U}$, it computes one column $\mathbf{v}$ of $\mathbf{V}$ solving

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{v}\|_0, \tag{24}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{v}\|_1, \tag{25}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{v}\|_2^2, \tag{26}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\|\mathbf{v}\|_1 + \lambda_2\|\mathbf{v}\|_2^2, \tag{27}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\sum_{j=1}^{p-1}|\mathbf{v}_{j+1}^i - \mathbf{v}_j^i| + \lambda_2\|\mathbf{v}\|_1 + \lambda_3\|\mathbf{v}\|_2^2, \tag{28}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\sum_{g\in\mathcal{T}}\delta^g(\mathbf{v}), \tag{29}$$

where $\mathcal{T}$ is a tree-structured set of groups (see [14]), and $\delta^g(\mathbf{v}) = 0$ if $\mathbf{v}_g = 0$ and 1 otherwise. It can also solve

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\sum_{g\in\mathcal{T}}\eta^g\|\mathbf{v}_g\|_2, \tag{30}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\sum_{g\in\mathcal{T}}\eta^g\|\mathbf{v}_g\|_\infty, \tag{31}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda\sum_{g\in\mathcal{G}}\eta^g\|\mathbf{v}_g\|_\infty, \tag{32}$$

where $\mathcal{G}$ is any kind of set of groups.

This function can also solve the following proximal operators on matrices

$$\min_{\mathbf{V}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{U}-\mathbf{V}\|_F^2 + \lambda\sum_{i=1}^{p}\|\mathbf{V}_i\|_2, \tag{33}$$

where $\mathbf{V}_i$ is the i-th row of $\mathbf{V}$, or

$$\min_{\mathbf{V}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{U}-\mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{V}_i\|_\infty, \tag{34}$$

or

$$\min_{\mathbf{V}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{U}-\mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{V}_i\|_2 + \lambda_2 \sum_{i=1}^{p}\sum_{j=1}^{n} |\mathbf{V}_{ij}|, \tag{35}$$

or

$$\min_{\mathbf{V}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{U}-\mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{V}_i\|_\infty + \lambda_2 \sum_{i=1}^{p}\sum_{j=1}^{n} |\mathbf{V}_{ij}|, \tag{36}$$

or

$$\min_{\mathbf{V}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{U}-\mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{V}_i\|_\infty + \lambda_2 \sum_{j=1}^{n} \|\mathbf{V}^j\|_\infty. \tag{37}$$

where $\mathbf{V}^j$ is the j-th column of $\mathbf{V}$.

See usage details below:

```
%
% Usage:   alpha=mexProximalFlat(U,param);
%
% Name: mexProximalFlat
%
% Description: mexProximalFlat computes proximal operators. Depending
%         on the value of param.regul, it computes
%
%         Given an input matrix U=[u^1,\ldots,u^n], it computes a matrix
%         V=[v^1,\ldots,v^n] such that
%         if one chooses a regularization functions on vectors, it computes
%         for each column u of U, a column v of V solving
%         if param.regul='l0'
%             argmin 0.5||u-v||_2^2 + lambda||v||_0
%         if param.regul='l1'
%             argmin 0.5||u-v||_2^2 + lambda||v||_1
%         if param.regul='l2'
%             argmin 0.5||u-v||_2^2 + lambda||v||_2^2
%         if param.regul='elastic-net'
%             argmin 0.5||u-v||_2^2 + lambda||v||_1 + lambda_2||v||_2^2
%         if param.regul='fused-lasso'
%             argmin 0.5||u-v||_2^2 + lambda FL(v) + ...
%                                ...  lambda_2||v||_1 + lambda_3||v||_2^2
%         if param.regul='linf'
%             argmin 0.5||u-v||_2^2 + lambda||v||_inf
%         if param.regul='l2-not-squared'
%             argmin 0.5||u-v||_2^2 + lambda||v||_2
%         if param.regul='group-lasso-l2'
%             argmin 0.5||u-v||_2^2 + lambda sum_g ||v_g||_2
%             where the groups are consecutive entries of v of size param.group_size
%         if param.regul='group-lasso-linf'
%             argmin 0.5||u-v||_2^2 + lambda sum_g ||v_g||_inf
%         if param.regul='sparse-group-lasso-l2'
%             argmin 0.5||u-v||_2^2 + lambda sum_g ||v_g||_2 + lambda_2 ||v||_1
%             where the groups are consecutive entries of v of size param.group_size
%         if param.regul='sparse-group-lasso-linf'
%             argmin 0.5||u-v||_2^2 + lambda sum_g ||v_g||_inf + lambda_2 ||v||_1
%         if param.regul='trace-norm-vec'
%             argmin 0.5||u-v||_2^2 + lambda ||mat(v)||_*
```

```
%            where mat(v) has param.group_size rows
%
%         if one chooses a regularization function on matrices
%         if param.regul='l1l2',  V=
%             argmin 0.5||U-V||_F^2 + lambda||V||_{1/2}
%         if param.regul='l1linf',  V=
%             argmin 0.5||U-V||_F^2 + lambda||V||_{1/inf}
%         if param.regul='l1l2+l1',  V=
%             argmin 0.5||U-V||_F^2 + lambda||V||_{1/2} + lambda_2||V||_{1/1}
%         if param.regul='l1linf+l1',  V=
%             argmin 0.5||U-V||_F^2 + lambda||V||_{1/inf} + lambda_2||V||_{1/1}
%         if param.regul='l1linf+row-column',  V=
%             argmin 0.5||U-V||_F^2 + lambda||V||_{1/inf} + lambda_2||V'||_{1/inf}
%         if param.regul='trace-norm',  V=
%             argmin 0.5||U-V||_F^2 + lambda||V||_*
%         if param.regul='rank',  V=
%             argmin 0.5||U-V||_F^2 + lambda rank(V)
%         if param.regul='none',  V=
%             argmin 0.5||U-V||_F^2
%
%         for all these regularizations, it is possible to enforce non-negativity constraints
%         with the option param.pos, and to prevent the last row of U to be regularized, with
%         the option param.intercept
%
% Inputs: U:  double m x n matrix   (input signals)
%               m is the signal size
%         param: struct
%               param.lambda  (regularization parameter)
%               param.regul (choice of regularization, see above)
%               param.lambda2  (optional, regularization parameter)
%               param.lambda3  (optional, regularization parameter)
%               param.verbose (optional, verbosity level, false by default)
%               param.intercept (optional, last row of U is not regularized,
%                 false by default)
%               param.transpose (optional, transpose the matrix in the regularizaiton function)
%               param.group_size (optional, for regularization functions assuming a group
%                 structure)
%               param.pos (optional, adds positivity constraints on the
%                 coefficients, false by default)
%               param.numThreads (optional, number of threads for exploiting
%                 multi-core / multi-cpus. By default, it takes the value -1,
%                 which automatically selects all the available CPUs/cores).
%
% Output: alpha: double m x n matrix (output coefficients)
%
% Author: Julien Mairal, 2010
```

5.3 Function mexProximalTree

This function computes the proximal operators associated to tree-structured regularization functions, for input signals $\mathbf{U} = [\mathbf{u}^1, \ldots, \mathbf{u}^n]$ in $\mathbb{R}^{p \times n}$, and a tree-structured set of groups [13], it computes a matrix $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^n]$ in $\mathbb{R}^{p \times n}$. When one uses a regularization function on vectors, it computes a column $\mathbf{v}$ of $\mathbf{V}$ for every column $\mathbf{u}$ of $\mathbf{U}$:

$$\min_{\mathbf{v} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 + \lambda \sum_{g \in \mathcal{T}} \eta^g \|\mathbf{v}_g\|_2, \tag{38}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda \sum_{g\in\mathcal{T}} \eta^g \|\mathbf{v}_g\|_\infty, \tag{39}$$

or

$$\min_{\mathbf{v}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|_2^2 + \lambda \sum_{g\in\mathcal{T}} \delta^g(\mathbf{v}), \tag{40}$$

where $\delta^g(\mathbf{v}) = 0$ if $\mathbf{v}_g = 0$ and 1 otherwise (see appendix of [14]).

When the multi-task tree-structured regularization function is used, it solves

$$\min_{\mathbf{V}\in\mathbb{R}^{p\times n}} \frac{1}{2}\|\mathbf{U}-\mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{n} \sum_{g\in\mathcal{T}} \eta^g \|\mathbf{v}_g^i\|_\infty + \lambda_2 \sum_{g\in\mathcal{T}} \max_{j\in g} \|\mathbf{v}_g^j\|_\infty, \tag{41}$$

which is a formulation presented in [20].

This function can also be used for computing the proximal operators addressed by mexProximalFlat (it will just not take into account the tree structure). The way the tree is incoded is presented below, (and examples are given in the file test_ProximalTree.m, with more usage details:

```
%
% Usage:   alpha=mexProximalTree(U,tree,param);
%
% Name: mexProximalTree
%
% Description: mexProximalTree computes a proximal operator. Depending
%         on the value of param.regul, it computes
%
%         Given an input matrix U=[u^1,\ldots,u^n], and a tree-structured set of groups T,
%         it returns a matrix V=[v^1,\ldots,v^n]:
%
%         when the regularization function is for vectors,
%         for every column u of U, it compute a column v of V solving
%         if param.regul='tree-l0'
%             argmin 0.5||u-v||_2^2 + lambda \sum_{g \in T} \delta^g(v)
%         if param.regul='tree-l2'
%           for all i, v^i =
%             argmin 0.5||u-v||_2^2 + lambda\sum_{g \in T} \eta_g||v_g||_2
%         if param.regul='tree-linf'
%           for all i, v^i =
%             argmin 0.5||u-v||_2^2 + lambda\sum_{g \in T} \eta_g||v_g||_inf
%
%         when the regularization function is for matrices:
%         if param.regul='multi-task-tree'
%            V=argmin 0.5||U-V||_F^2 + lambda \sum_{i=1}^n\sum_{g \in T} \eta_g||v^i_g||_inf + ...
%                                                lambda_2 \sum_{g \in T} \eta_g max_{j in g}||V_j||_{inf}
%
%         it can also be used with any non-tree-structured regularization addressed by mexProximalFlat
%
%         for all these regularizations, it is possible to enforce non-negativity constraints
%         with the option param.pos, and to prevent the last row of U to be regularized, with
%         the option param.intercept
%
% Inputs: U:  double m x n matrix   (input signals)
%               m is the signal size
%         tree: struct
%               with four fields, eta_g, groups, own_variables and N_own_variables.
%
%               The tree structure requires a particular organization of groups and variables
%                  * Let us denote by N = |T|, the number of groups.
```

```
%                   the groups should be ordered T={g1,g2,\ldots,gN} such that if gi is included
%                   in gj, then j <= i. g1 should be the group at the root of the tree
%                   and contains every variable.
%                 * Every group is a set of  contiguous indices for instance
%                   gi={3,4,5} or gi={4,5,6,7} or gi={4}, but not {3,5};
%                 * We define root(gi) as the indices of the variables that are in gi,
%                   but not in its descendants. For instance for
%                   T={ g1={1,2,3,4},g2={2,3},g3={4} }, then, root(g1)={1},
%                   root(g2)={2,3}, root(g3)={4},
%                   We assume that for all i, root(gi) is a set of contigous variables
%                 * We assume that the smallest of root(gi) is also the smallest index of gi.
%
%                 For instance,
%                   T={ g1={1,2,3,4},g2={2,3},g3={4} }, is a valid set of groups.
%                   but we can not have
%                   T={ g1={1,2,3,4},g2={1,2},g3={3} }, since root(g1)={4} and 4 is not the
%                   smallest element in g1.
%
%               We do not lose generality with these assumptions since they can be fullfilled for any
%               tree-structured set of groups after a permutation of variables and a correct ordering of the
%               groups.
%               see more examples in test_ProximalTree.m of valid tree-structured sets of groups.
%
%               The first fields sets the weights for every group
%                  tree.eta_g            double N vector
%
%               The next field sets inclusion relations between groups
%               (but not between groups and variables):
%                  tree.groups           sparse (double or boolean) N x N matrix
%                  the (i,j) entry is non-zero if and only if i is different than j and
%                  gi is included in gj.
%                  the first column corresponds to the group at the root of the tree.
%
%               The next field define the smallest index of each group gi,
%               which is also the smallest index of root(gi)
%               tree.own_variables    int32 N vector
%
%               The next field define for each group gi, the size of root(gi)
%               tree.N_own_variables  int32 N vector
%
%               examples are given in test_ProximalTree.m
%
%         param: struct
%               param.lambda  (regularization parameter)
%               param.regul (choice of regularization, see above)
%               param.lambda2  (optional, regularization parameter)
%               param.lambda3  (optional, regularization parameter)
%               param.verbose (optional, verbosity level, false by default)
%               param.intercept (optional, last row of U is not regularized,
%                 false by default)
%               param.pos (optional, adds positivity constraints on the
%                 coefficients, false by default)
%               param.numThreads (optional, number of threads for exploiting
%                 multi-core / multi-cpus. By default, it takes the value -1,
%                 which automatically selects all the available CPUs/cores).
%
% Output: alpha: double m x n matrix (output coefficients)
%
% Author: Julien Mairal, 2010
```

5.4 Function mexProximalGraph

This function computes the proximal operators associated to structured sparse regularization, for input signals $\mathbf{U} = [\mathbf{u}^1, \ldots, \mathbf{u}^n]$ in $\mathbb{R}^{p \times n}$, and a set of groups [20], it returns a matrix $\mathbf{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^n]$ in $\mathbb{R}^{p \times n}$. When one uses a regularization function on vectors, it computes a column $\mathbf{v}$ of $\mathbf{V}$ for every column $\mathbf{u}$ of $\mathbf{U}$:

$$\min_{\mathbf{v} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_2^2 + \lambda \sum_{g \in \mathcal{G}} \eta^g \|\mathbf{v}_g\|_\infty, \tag{42}$$

or with a regularization function on matrices, it computes $\mathbf{V}$ solving

$$\min_{\mathbf{V} \in \mathbb{R}^{p \times n}} \frac{1}{2}\|\mathbf{U} - \mathbf{V}\|_F^2 + \lambda \sum_{i=1}^{n} \sum_{g \in \mathcal{G}} \eta^g \|\mathbf{v}_g^i\|_\infty + \lambda_2 \sum_{g \in \mathcal{G}} \max_{j \in g} \|\mathbf{v}_g^j\|_\infty, \tag{43}$$

This function can also be used for computing the proximal operators addressed by mexProximalFlat. The way the graph is incoded is presented below (and also in the example file test_ProximalGraph.m, with more usage details:

```
%
% Usage:   alpha=mexProximalGraph(U,graph,param);
%
% Name: mexProximalGraph
%
% Description: mexProximalGraph computes a proximal operator. Depending
%         on the value of param.regul, it computes
%
%         Given an input matrix U=[u^1,\ldots,u^n], and a set of groups G,
%         it computes a matrix V=[v^1,\ldots,v^n] such that
%
%         if param.regul='graph'
%         for every column u of U, it computes a column v of V solving
%             argmin 0.5||u-v||_2^2 + lambda\sum_{g \in G} \eta_g||v_g||_inf
%
%         if param.regul='graph+ridge'
%         for every column u of U, it computes a column v of V solving
%             argmin 0.5||u-v||_2^2 + lambda\sum_{g \in G} \eta_g||v_g||_inf + lambda_2||v||_2^2
%
%
%         if param.regul='multi-task-graph'
%            V=argmin 0.5||U-V||_F^2 + lambda \sum_{i=1}^n\sum_{g \in G} \eta_g||v^i_g||_inf + ...
%                                                lambda_2 \sum_{g \in G} \eta_g max_{j in g}||V_j||_{inf}
%
%         it can also be used with any regularization addressed by mexProximalFlat
%
%         for all these regularizations, it is possible to enforce non-negativity constraints
%         with the option param.pos, and to prevent the last row of U to be regularized, with
%         the option param.intercept
%
% Inputs: U:  double p x n matrix   (input signals)
%               m is the signal size
%         graph: struct
%               with three fields, eta_g, groups, and groups_var
%
%               The first fields sets the weights for every group
```

```
%                   graph.eta_g            double N vector
%
%               The next field sets inclusion relations between groups
%               (but not between groups and variables):
%                   graph.groups           sparse (double or boolean) N x N matrix
%                   the (i,j) entry is non-zero if and only if i is different than j and
%                   gi is included in gj.
%
%               The next field sets inclusion relations between groups and variables
%                   graph.groups_var       sparse (double or boolean) p x N matrix
%                   the (i,j) entry is non-zero if and only if the variable i is included
%                   in gj, but not in any children of gj.
%
%               examples are given in test_ProximalGraph.m
%
%         param: struct
%               param.lambda  (regularization parameter)
%               param.regul (choice of regularization, see above)
%               param.lambda2  (optional, regularization parameter)
%               param.lambda3  (optional, regularization parameter)
%               param.verbose (optional, verbosity level, false by default)
%               param.intercept (optional, last row of U is not regularized,
%                 false by default)
%               param.pos (optional, adds positivity constraints on the
%                 coefficients, false by default)
%               param.numThreads (optional, number of threads for exploiting
%                 multi-core / multi-cpus. By default, it takes the value -1,
%                 which automatically selects all the available CPUs/cores).
%
% Output: alpha: double p x n matrix (output coefficients)
%
% Author: Julien Mairal, 2010
```

After having presented the regularization terms which our software can handle, we present the various formulations that we address

5.5 Problems Addressed

We present here regression or classification formulations and their multi-task variants.

5.5.1 Regression Problems with the Square Loss

Given a training set $\{\mathbf{x}^i, y_i\}_{i=1}^n$, with $\mathbf{x}^i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for all i in $[1; n]$, we address

$$\min_{\mathbf{w}\in\mathbb{R}^p, b\in\mathbb{R}} \sum_{i=1}^{n} \frac{1}{2}(y_i - \mathbf{w}^\top \mathbf{x}^i - b)^2 + \lambda\psi(\mathbf{w}),$$

where b is an optional variable acting as an "intercept", which is not regularized, and ψ can be any of the regularization functions presented above. Let us consider the vector $\mathbf{y}$ in $\mathbb{R}^n$ that carries the entries y_i. The problem without the intercept takes the following form, which we have already encountered in the previous toolbox, but with different notations:

$$\min_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\psi(\mathbf{w}),$$

where the $\mathbf{X} = [\mathbf{x}^i, \ldots, \mathbf{x}^n]^T$ (the $\mathbf{x}^i$'s are here the rows of $\mathbf{X}$).

5.5.2 Classification Problems with the Logistic Loss

The next formulation that our software can solve is the regularized logistic regression formulation. We are again given a training set $\{\mathbf{x}^i, y_i\}_{i=1}^n$, with $\mathbf{x}^i \in \mathbb{R}^p$, but the variables y_i are now in $\{-1, +1\}$ for all i in $[1; n]$. The optimization problem we address is

$$\min_{\mathbf{w}\in\mathbb{R}^p, b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i(\mathbf{w}^\top \mathbf{x}^i + b)} + \lambda\psi(\mathbf{w}),$$

with again ψ taken to be one of the regularization function presented above, and b is an optional intercept.

5.5.3 Multi-class Classification Problems with the Softmax Loss

We have also implemented a multi-class logistic classifier (or softmax). For a classification problem with r classes, we are given a training set $\{\mathbf{x}^i, y_i\}_{i=1}^n$, where the variables $\mathbf{x}^i$ are still vectors in $\mathbb{R}^p$, but the y_i's have integer values in $\{1, 2, \ldots, r\}$. The formulation we address is the following multi-class learning problem

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}, \mathbf{b}\in\mathbb{R}^r} \frac{1}{n}\sum_{i=1}^{n} \log\Big(\sum_{j=1}^{r} e^{(\mathbf{w}^j - \mathbf{w}^{\mathbf{y}_i})^\top \mathbf{x}^i + \mathbf{b}_j - \mathbf{b}_{\mathbf{y}_i}}\Big) + \lambda\sum_{j=1}^{r}\psi(\mathbf{w}^j), \tag{44}$$

where $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^r]$ and the optional vector $\mathbf{b}$ in $\mathbb{R}^r$ carries intercepts for each class.

5.5.4 Multi-task Regression Problems with the Square Loss

We are now considering a problem with r tasks, and a training set $\{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^n$, where the variables $\mathbf{x}^i$ are still vectors in $\mathbb{R}^p$, and $\mathbf{y}^i$ is a vector in $\mathbb{R}^r$. We are looking for r regression vectors $\mathbf{w}^j$, for $j \in [1; r]$, or equivalently for a matrix $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^r]$ in $\mathbb{R}^{p\times r}$. The formulation we address is the following multi-task regression problem

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}, \mathbf{b}\in\mathbb{R}^r} \sum_{j=1}^{r}\sum_{i=1}^{n} \frac{1}{2}(\mathbf{y}_j^i - \mathbf{w}^\top \mathbf{x}^i - \mathbf{b}_j)^2 + \lambda\psi(\mathbf{W}),$$

where ψ is any of the regularization function on matrices we have presented in the previous section. Note that by introducing the appropriate variables $\mathbf{Y}$, the problem without intercept could be equivalently rewritten

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_{\mathrm{F}}^2 + \lambda\psi(\mathbf{W}).$$

5.5.5 Multi-task Classification Problems with the Logistic Loss

The multi-task version of the logistic regression follows the same principle. We consider r tasks, and a training set $\{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^n$, with the $\mathbf{x}^i$'s in $\mathbb{R}^p$, and the $\mathbf{y}^i$'s are vectors in $\{-1, +1\}^r$. We look for a matrix $\mathbf{W} = [\mathbf{w}^1, \ldots, \mathbf{w}^r]$ in $\mathbb{R}^{p\times r}$. The formulation is the following multi-task regression problem

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r}, \mathbf{b}\in\mathbb{R}^r} \sum_{j=1}^{r}\frac{1}{n}\sum_{i=1}^{n} \log\left(1 + e^{-\mathbf{y}_j^i(\mathbf{w}^\top \mathbf{x}^i + \mathbf{b}_j)}\right) + \lambda\psi(\mathbf{W}).$$

5.5.6 Multi-task and Multi-class Classification Problems with the Softmax Loss

The multi-task/multi-class version directly follows from the formulation of Eq. (44), but associates with each class a task, and as a consequence, regularizes the matrix $\mathbf{W}$ in a particular way:

$$\min_{\mathbf{W}\in\mathbb{R}^{p\times r},\mathbf{b}\in\mathbb{R}^r} \frac{1}{n}\sum_{i=1}^{n} \log\Big(\sum_{j=1}^{r} e^{(\mathbf{w}^j-\mathbf{w}^{\mathbf{y}_i})^\top \mathbf{x}^i+\mathbf{b}_j-\mathbf{b}_{\mathbf{y}_i}}\Big) + \lambda\psi(\mathbf{W}).$$

How duality gaps are computed for any of these formulations is presented in Appendix A. We now present the main functions for solving these problems

5.6 Function mexFistaFlat

Given a matrix $\mathbf{X} = [\mathbf{x}^1,\ldots,\mathbf{x}^p]^T$ in $\mathbb{R}^{m\times p}$, and a matrix $\mathbf{Y} = [\mathbf{y}^1,\ldots,\mathbf{y}^n]$, it solves the optimization problems presented in the previous section, with the same regularization functions as mexProximalFlat. see usage details below:

```
%
% Usage: W=mexFistaFlat(Y,X,W0,param);
%
% Name: mexFistaFlat
%
% Description: mexFistaFlat solves sparse regularized problems.
%         X is a design matrix of size m x p
%         X=[x^1,...,x^n]', where the x_i's are the rows of X
%         Y=[y^1,...,y^n] is a matrix of size m x n
%         It implements the algorithms FISTA, ISTA and subgradient descent.
%
%           - if param.loss='square' and param.regul is a regularization function for vectors,
%             the entries of Y are real-valued,  W = [w^1,...,w^n] is a matrix of size p x n
%             For all column y of Y, it computes a column w of W such that
%               w = argmin 0.5||y- X w||_2^2 + lambda psi(w)
%
%           - if param.loss='square' and param.regul is a regularization function for matrices
%             the entries of Y are real-valued,  W is a matrix of size p x n.
%             It computes the matrix W such that
%               W = argmin 0.5||Y- X W||_F^2 + lambda psi(W)
%
%           - param.loss='square-missing' : same as param.loss='square', but handles missing data
%             represented by NaN (not a number) in the matrix Y
%
%           - if param.loss='logistic' and param.regul is a regularization function for vectors,
%             the entries of Y are either -1 or +1, W = [w^1,...,w^n] is a matrix of size p x n
%             For all column y of Y, it computes a column w of W such that
%               w = argmin (1/m)sum_{j=1}^m log(1+e^(-y_j x^j' w)) + lambda psi(w),
%             where x^j is the j-th row of X.
%
%           - if param.loss='logistic' and param.regul is a regularization function for matrices
%             the entries of Y are either -1 or +1, W is a matrix of size p x n
%               W = argmin sum_{i=1}^n(1/m)sum_{j=1}^m log(1+e^(-y^i_j x^j' w^i)) + lambda psi(W)
%
%           - if param.loss='multi-logistic' and param.regul is a regularization function for vectors,
%             the entries of Y are in {0,1,...,N} where N is the total number of classes
%             W = [W^1,...,W^n] is a matrix of size p x Nn, each submatrix W^i is of size p x N
%             for all submatrix WW of W, and column y of Y, it computes
%               WW = argmin (1/m)sum_{j=1}^m log(sum_{j=1}^r e^(x^j'(ww^j-ww^{y_j}))) + lambda sum_{j=1}^N ps
```

```
%             where ww^j is the j-th column of WW.
%
%           - if param.loss='multi-logistic' and param.regul is a regularization function for matrices,
%             the entries of Y are in {0,1,...,N} where N is the total number of classes
%             W is a matrix of size p x N, it computes
%               W = argmin (1/m)sum_{j=1}^m log(sum_{j=1}^r e^(x^j'(w^j-w^{y_j}))) + lambda psi(W)
%             where ww^j is the j-th column of WW.
%
%           - param.loss='cur' : useful to perform sparse CUR matrix decompositions,
%               W = argmin 0.5||Y-X*W*X||_F^2 + lambda psi(W)
%
%
%         The function psi are those used by mexProximalFlat (see documentation)
%
%         This function can also handle intercepts (last row of W is not regularized),
%         and/or non-negativity constraints on W, and sparse matrices for X
%
% Inputs: Y:  double dense m x n matrix
%         X:  double dense or sparse m x p matrix
%         W0:  double dense p x n matrix or p x Nn matrix (for multi-logistic loss)
%              initial guess
%         param: struct
%            param.loss (choice of loss, see above)
%            param.regul (choice of regularization, see function mexProximalFlat)
%            param.lambda (regularization parameter)
%            param.lambda2 (optional, regularization parameter, 0 by default)
%            param.lambda3 (optional, regularization parameter, 0 by default)
%            param.verbose (optional, verbosity level, false by default)
%            param.intercept (optional, last row of U is not regularized,
%                false by default)
%            param.pos (optional, adds positivity constraints on the
%                coefficients, false by default)
%            param.numThreads (optional, number of threads for exploiting
%                multi-core / multi-cpus. By default, it takes the value -1,
%                which automatically selects all the available CPUs/cores).
%            param.max_it (optional, maximum number of iterations, 100 by default)
%            param.tol (optional, tolerance for stopping criteration, which is a relative duality gap
%                if it is available, or a relative change of parameters).
%            param.gamma (optional, multiplier for increasing the parameter L in fista, 1.5 by default)
%            param.L0 (optional, initial parameter L in fista, 0.1 by default, should be small enough)
%            param.fixed_step (deactive the line search for L in fista and use param.L0 instead)
%            param.compute_gram (optional, pre-compute X^TX, false by default).
%            param.intercept (optional, do not regularize last row of W, false by default).
%            param.ista (optional, use ista instead of fista, false by default).
%            param.subgrad (optional, if not param.ista, use subradient descent instead of fista, false by de
%            param.a, param.b (optional, if param.subgrad, the gradient step is a/(t+b)
%            also similar options as mexProximalFlat
%
%            the function also implements the ADMM algorithm via an option param.admm=true. It is not documen
%            and you need to look at the source code to use it.
%
% Output:  W:  double dense p x n matrix or p x Nn matrix (for multi-logistic loss)
%
% Author: Julien Mairal, 2010
```

5.7 Function mexFistaTree

Given a matrix $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^p]^T$ in $\mathbb{R}^{m \times p}$, and a matrix $\mathbf{Y} = [\mathbf{y}^1, \ldots, \mathbf{y}^n]$, it solves the optimization problems presented in the previous section, with the same regularization functions as mexProximalTree. see usage details below:

```
%
% Usage: W=mexFistaTree(Y,X,W0,tree,param);
%
% Name: mexFistaTree
%
% Description: mexFistaTree solves sparse regularized problems.
%         X is a design matrix of size m x p
%         X=[x^1,...,x^n]', where the x_i's are the rows of X
%         Y=[y^1,...,y^n] is a matrix of size m x n
%         It implements the algorithms FISTA, ISTA and subgradient descent for solving
%
%           min_W  loss(W) + lambda psi(W)
%
%         The function psi are those used by mexProximalTree (see documentation)
%         for the loss functions, see the documentation of mexFistaFlat
%
%         This function can also handle intercepts (last row of W is not regularized),
%         and/or non-negativity constraints on W and sparse matrices X
%
% Inputs: Y:  double dense m x n matrix
%         X:  double dense or sparse m x p matrix
%         W0:  double dense p x n matrix or p x Nn matrix (for multi-logistic loss)
%              initial guess
%         tree: struct (see documentation of mexProximalTree)
%         param: struct
%            param.loss (choice of loss, see above)
%            param.regul (choice of regularization, see function mexProximalFlat)
%            param.lambda (regularization parameter)
%            param.lambda2 (optional, regularization parameter, 0 by default)
%            param.lambda3 (optional, regularization parameter, 0 by default)
%            param.verbose (optional, verbosity level, false by default)
%            param.intercept (optional, last row of U is not regularized,
%                false by default)
%            param.pos (optional, adds positivity constraints on the
%                coefficients, false by default)
%            param.numThreads (optional, number of threads for exploiting
%                multi-core / multi-cpus. By default, it takes the value -1,
%                which automatically selects all the available CPUs/cores).
%            param.max_it (optional, maximum number of iterations, 100 by default)
%            param.tol (optional, tolerance for stopping criteration, which is a relative duality gap
%                if it is available, or a relative change of parameters).
%            param.gamma (optional, multiplier for increasing the parameter L in fista, 1.5 by default)
%            param.L0 (optional, initial parameter L in fista, 0.1 by default, should be small enough)
%            param.fixed_step (deactive the line search for L in fista and use param.L0 instead)
%            param.compute_gram (optional, pre-compute X^TX, false by default).
%            param.intercept (optional, do not regularize last row of W, false by default).
%            param.ista (optional, use ista instead of fista, false by default).
%            param.subgrad (optional, if not param.ista, use subradient descent instead of fista, false by de
%            param.a, param.b (optional, if param.subgrad, the gradient step is a/(t+b)
%            also similar options as mexProximalTree
%
%            the function also implements the ADMM algorithm via an option param.admm=true. It is not documen
%            and you need to look at the source code to use it.
%
```

```
% Output:  W:  double dense p x n matrix or p x Nn matrix (for multi-logistic loss)
%
% Author: Julien Mairal, 2010
```

5.8 Function mexFistaGraph

Given a matrix $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^p]^T$ in $\mathbb{R}^{m \times p}$, and a matrix $\mathbf{Y} = [\mathbf{y}^1, \ldots, \mathbf{y}^n]$, it solves the optimization problems presented in the previous section, with the same regularization functions as mexProximalGraph. see usage details below:

```
%
% Usage: W=mexFistaGraph(Y,X,W0,graph,param);
%
% Name: mexFistaGraph
%
% Description: mexFistaGraph solves sparse regularized problems.
%         X is a design matrix of size m x p
%         X=[x^1,...,x^n]', where the x_i's are the rows of X
%         Y=[y^1,...,y^n] is a matrix of size m x n
%         It implements the algorithms FISTA, ISTA and subgradient descent.
%
%         It implements the algorithms FISTA, ISTA and subgradient descent for solving
%
%           min_W  loss(W) + lambda psi(W)
%
%         The function psi are those used by mexProximalGraph (see documentation)
%         for the loss functions, see the documentation of mexFistaFlat
%
%         This function can also handle intercepts (last row of W is not regularized),
%         and/or non-negativity constraints on W.
%
% Inputs: Y:  double dense m x n matrix
%         X:  double dense or sparse m x p matrix
%         W0:  double dense p x n matrix or p x Nn matrix (for multi-logistic loss)
%              initial guess
%         graph: struct (see documentation of mexProximalGraph)
%         param: struct
%            param.loss (choice of loss, see above)
%            param.regul (choice of regularization, see function mexProximalFlat)
%            param.lambda (regularization parameter)
%            param.lambda2 (optional, regularization parameter, 0 by default)
%            param.lambda3 (optional, regularization parameter, 0 by default)
%            param.verbose (optional, verbosity level, false by default)
%            param.intercept (optional, last row of U is not regularized,
%                false by default)
%            param.pos (optional, adds positivity constraints on the
%                coefficients, false by default)
%            param.numThreads (optional, number of threads for exploiting
%                multi-core / multi-cpus. By default, it takes the value -1,
%                which automatically selects all the available CPUs/cores).
%            param.max_it (optional, maximum number of iterations, 100 by default)
%            param.tol (optional, tolerance for stopping criteration, which is a relative duality gap
%                if it is available, or a relative change of parameters).
%            param.gamma (optional, multiplier for increasing the parameter L in fista, 1.5 by default)
%            param.L0 (optional, initial parameter L in fista, 0.1 by default, should be small enough)
%            param.fixed_step (deactive the line search for L in fista and use param.L0 instead)
%            param.compute_gram (optional, pre-compute X^TX, false by default).
```

```
%            param.intercept (optional, do not regularize last row of W, false by default).
%            param.ista (optional, use ista instead of fista, false by default).
%            param.subgrad (optional, if not param.ista, use subradient descent instead of fista, false by de
%            param.a, param.b (optional, if param.subgrad, the gradient step is a/(t+b)
%            also similar options as mexProximalTree
%
%            the function also implements the ADMM algorithm via an option param.admm=true. It is not documen
%            and you need to look at the source code to use it.
%
%
% Output:  W:  double dense p x n matrix or p x Nn matrix (for multi-logistic loss)
%
% Author: Julien Mairal, 2010
```

6 Miscellaneous Functions

6.1 Function mexConjGrad

Implementation of a conjugate gradient for solving a linear system $\mathbf{Ax} = \mathbf{b}$ when $\mathbf{A}$ is positive definite. In some cases, it is faster than the Matlab function `pcg`, especially when the library uses the Intel Math Kernel Library.

```
%
% Usage:   x =mexConjGrad(A,b,x0,tol,itermax)
%
% Name: mexConjGrad
%
% Description: Conjugate gradient algorithm, sometimes faster than the
%    equivalent Matlab function pcg. In order to solve Ax=b;
%
% Inputs: A:  double square n x n matrix. HAS TO BE POSITIVE DEFINITE
%         b:  double vector of length n.
%         x0: double vector of length n. (optional) initial guess.
%         tol: (optional) tolerance.
%         itermax: (optional) maximum number of iterations.
%
% Output: x: double vector of length n.
%
% Author: Julien Mairal, 2009
```

6.2 Function mexBayer

Apply a Bayer pattern to an input image

```
%
% Usage:   Y =mexBayer(X,offset);
%
% Description: mexBayer applies a Bayer pattern to an image X.
%       There are four possible offsets.
%
% Inputs: X:  double m x n matrix
%         offset: scalar, 0,1,2 or 3
%
% Output: Y: double m x m matrix
%
% Author: Julien Mairal, 2009
```

6.3 Function mexCalcAAt

For an input sparse matrix $\mathbf{A}$, this function returns $\mathbf{A}\mathbf{A}^T$. For some reasons, when $\mathbf{A}$ has a lot more columns than rows, this function can be much faster than the equivalent matlab command `X*A'`.

```
%
% Usage:   AAt =mexCalcAAt(A);
%
% Name: mexCalcAAt
%
% Description: Compute efficiently AAt = A*A', when A is sparse
%   and has a lot more columns than rows. In some cases, it is
%   up to 20 times faster than the equivalent Matlab expression
%   AAt=A*A';
%
% Inputs: A:  double sparse m x n matrix
%
% Output: AAt: double m x m matrix
%
% Author: Julien Mairal, 2009
```

6.4 Function mexCalcXAt

For an input sparse matrix $\mathbf{A}$ and a matrix $\mathbf{X}$, this function returns $\mathbf{X}\mathbf{A}^T$. For some reasons, when $\mathbf{A}$ has a lot more columns than rows, this function can be much faster than the equivalent matlab command `X*A'`.

```
%
% Usage:   XAt =mexCalcXAt(X,A);
%
% Name: mexCalcXAt
%
% Description: Compute efficiently XAt = X*A', when A is sparse and has a
%   lot more columns than rows. In some cases, it is up to 20 times
%   faster than the equivalent Matlab expression XAt=X*A';
%
% Inputs: X:  double m x n matrix
%         A:  double sparse p x n matrix
%
% Output: XAt: double m x p matrix
%
% Author: Julien Mairal, 2009
```

6.5 Function mexCalcXY

For two input matrices $\mathbf{X}$ and $\mathbf{Y}$, this function returns $\mathbf{X}\mathbf{Y}$.

```
%
% Usage:   Z =mexCalcXY(X,Y);
%
% Name: mexCalcXY
%
```

```
% Description: Compute Z=XY using the BLAS library used by SPAMS.
%
% Inputs: X:  double m x n matrix
%         Y:  double n x p matrix
%
% Output: Z: double m x p matrix
%
% Author: Julien Mairal, 2009
```

6.6 Function mexCalcXYt

For two input matrices $\mathbf{X}$ and $\mathbf{Y}$, this function returns $\mathbf{X}\mathbf{Y}^T$.

```
%
% Usage:   Z =mexCalcXYt(X,Y);
%
% Name: mexCalcXYt
%
% Description: Compute Z=XY' using the BLAS library used by SPAMS.
%
% Inputs: X:  double m x n matrix
%         Y:  double p x n matrix
%
% Output: Z: double m x p matrix
%
% Author: Julien Mairal, 2009
```

6.7 Function mexCalcXtY

For two input matrices $\mathbf{X}$ and $\mathbf{Y}$, this function returns $\mathbf{X}^T\mathbf{Y}$.

```
%
% Usage:   Z =mexCalcXtY(X,Y);
%
% Name: mexCalcXtY
%
% Description: Compute Z=X'Y using the BLAS library used by SPAMS.
%
% Inputs: X:  double n x m matrix
%         Y:  double n x p matrix
%
% Output: Z: double m x p matrix
%
% Author: Julien Mairal, 2009
```

6.8 Function mexInvSym

For an input symmetric matrices $\mathbf{A}$ in $\mathbb{R}^{n\times n}$, this function returns $\mathbf{A}^{-1}$.

```
%
% Usage:   B =mexInvSym(A);
%
% Name: mexInvSym
%
```

```
% Description: returns the inverse of a symmetric matrix A
%
% Inputs: A:  double n x n matrix
%
% Output: B: double n x n matrix
%
% Author: Julien Mairal, 2009
```

6.9 Function mexNormalize

```
%
% Usage:   Y =mexNormalize(X);
%
% Name: mexNormalize
%
% Description: rescale the columns of X so that they have
%        unit l2-norm.
%
% Inputs: X:  double m x n matrix
%
% Output: Y: double m x n matrix
%
% Author: Julien Mairal, 2010
```

6.10 Function mexSort

```
%
% Usage:   Y =mexSort(X);
%
% Name: mexSort
%
% Description: sort the elements of X using quicksort
%
% Inputs: X:  double vector of size n
%
% Output: Y: double  vector of size n
%
% Author: Julien Mairal, 2010
```

6.11 Function mexDisplayPatches

Print to the screen a matrix containing as columns image patches.

A Duality Gaps with Fenchel Duality

This section is taken from the appendix D of Julien Mairal's PhD thesis [17]. We are going to use intensively Fenchel Duality (see [2]). Let us consider the problem

$$\min_{\mathbf{w}\in\mathbb{R}^p} [g(\mathbf{w}) \stackrel{\triangle}{=} f(\mathbf{w}) + \lambda\psi(\mathbf{w})], \tag{45}$$

We first notice that for all the formulations we have been interested in, $g(\mathbf{w})$ can be rewritten

$$g(\mathbf{w}) = \tilde{f}(\mathbf{X}^\top \mathbf{w}) + \lambda\psi(\mathbf{w}), \tag{46}$$

where $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n]$ are training vectors, and $\tilde{f}$ is an appropriated smooth real-valued function of $\mathbb{R}^n$, and ψ one of the regularization functions we have introduced.

Given a primal variable $\mathbf{w}$ in $\mathbb{R}^p$ and a dual variable $\boldsymbol{\kappa}$ in $\mathbb{R}^n$, we obtain using classical Fenchel duality rules [2], that the following quantity is a duality gap for problem (45):

$$\delta(\mathbf{w}, \boldsymbol{\kappa}) \triangleq g(\mathbf{w}) + \tilde{f}^*(\boldsymbol{\kappa}) + \lambda\psi^*(-\mathbf{X}\boldsymbol{\kappa}/\lambda),$$

where $\tilde{f}^*$ and ψ^* are respectively the Fenchel conjugates of $\tilde{f}$ and ψ. Denoting by $\mathbf{w}^\star$ the solution of Eq. (45), the duality gap is interesting in the sense that it upperbounds the difference with the optimal value of the function:

$$\delta(\mathbf{w}, \boldsymbol{\kappa}) \geq g(\mathbf{w}) - g(\mathbf{w}^\star) \geq 0.$$

Similarly, we will consider pairs of primal-dual variables $(\mathbf{W}, \mathbf{K})$ when dealing with matrices.

During the optimization, sequences of primal variables $\mathbf{w}$ are available, and one wishes to exploit duality gaps for estimating the difference $g(\mathbf{w}) - g(\mathbf{w}^\star)$. This requires the following components:

- being able to efficiently compute $\tilde{f}^*$ and ψ^*.
- being able to obtain a "good" dual variable $\boldsymbol{\kappa}$ given a primal variable $\mathbf{w}$, such that $\delta(\mathbf{w}, \boldsymbol{\kappa})$ is close to $g(\mathbf{w}) - g(\mathbf{w}^\star)$.

We suppose that the first point is satisfied (we will detail these computations for every loss and regularization functions in the sequel), and explain how to choose $\boldsymbol{\kappa}$ in general (details will also be given in the sequel).

Let us first consider the choice that associates with a primal variable $\mathbf{w}$, the dual variable

$$\boldsymbol{\kappa}(\mathbf{w}) \triangleq \nabla\tilde{f}(\mathbf{X}^\top \mathbf{w}), \tag{47}$$

and let us compute $\delta(\mathbf{w}, \boldsymbol{\kappa}(\mathbf{w}))$. First, easy computations show that for all vectors $\mathbf{z}$ in $\mathbb{R}^n$, $\tilde{f}^*(\nabla\tilde{f}(\mathbf{z})) = \mathbf{z}^\top\nabla\tilde{f}(\mathbf{z}) - \tilde{f}(\mathbf{z})$, which gives

$$\begin{aligned}
\delta(\mathbf{w}, \boldsymbol{\kappa}(\mathbf{w})) &= \tilde{f}(\mathbf{X}^\top \mathbf{w}) + \lambda\psi(\mathbf{w}) + \tilde{f}^*(\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w})) + \lambda\psi^*(-\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w})/\lambda), & (48)\\
&= \lambda\psi(\mathbf{w}) + \mathbf{w}^\top\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w}) + \lambda\psi^*(-\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w})/\lambda). & (49)
\end{aligned}$$

We now use the classical Fenchel-Young inequality (see, Proposition 3.3.4 of [2]) on the function ψ, which gives

$$\delta(\mathbf{w}, \boldsymbol{\kappa}(\mathbf{w})) \geq \mathbf{w}^\top\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w}) - \mathbf{w}^\top\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w}) = 0,$$

with equality if and only if $-\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w})$ belongs to $\partial\psi(\mathbf{w})$. Interestingly, we now that first-order optimality conditions for Eq. (46) gives that $-\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top \mathbf{w}^\star) \in \partial\psi(\mathbf{w}^\star)$. We have now in hand a non-negative function $\mathbf{w} \mapsto \delta(\mathbf{w}, \boldsymbol{\kappa}(\mathbf{w}))$ of $\mathbf{w}$, that upperbounds $g(\mathbf{w}) - g(\mathbf{w}^\star)$ and satisfying $\delta(\mathbf{w}^\star, \boldsymbol{\kappa}(\mathbf{w}^\star)) = 0$.

This is however not a sufficient property to make it a good measure of the quality of the optimization, and further work is required, that will be dependent on $\tilde{f}$ and ψ. We have indeed proven that $\delta(\mathbf{w}^\star, \boldsymbol{\kappa}(\mathbf{w}^\star))$ is always 0. However, for $\mathbf{w}$ different than $\mathbf{w}^\star$, $\delta(\mathbf{w}^\star, \boldsymbol{\kappa}(\mathbf{w}^\star))$ can be infinite, making it a non-informative duality-gap. The reasons for this can be one of the following:

- The term $\psi^*(-\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top\mathbf{w})/\lambda)$ might have an infinite value.
- Intercepts make the problem more complicated. One can write the formulation with an intercept by adding a row to $\mathbf{X}$ filled with the value 1, add one dimension to the vector $\mathbf{w}$, and consider a regularization function ψ that does regularize the last entry of $\mathbf{w}$. This further complexifies the computation of ψ^* and its definition, as shown in the next section.

Let us now detail how we proceed to solve these problems, but first without considering the intercept. The analysis is similar when working with matrices $\mathbf{W}$ instead of vectors $\mathbf{w}$.

A.0.1 Duality Gaps without Intercepts

Let us show how to compute the Fenchel conjugate of the functions we have introduced. We now present the Fenchel conjugate of the loss functions $\tilde{f}$.

- **The square loss**
$$\tilde{f}(\mathbf{z}) = \tfrac{1}{2n}\|\mathbf{y}-\mathbf{z}\|_2^2,$$
$$\tilde{f}^*(\boldsymbol{\kappa}) = \tfrac{n}{2}\|\boldsymbol{\kappa}\|_2^2 + \boldsymbol{\kappa}^\top\mathbf{y}.$$
- **The logistic loss**
$$\tilde{f}(\mathbf{z}) = \tfrac{1}{n}\sum_{i=1}^n \log(1+e^{-y_i\mathbf{z}_i})$$
$$\tilde{f}^*(\boldsymbol{\kappa}) = \begin{cases} +\infty \text{ if } \exists\ i \in [1;n] \ \text{ s.t. } \ y_i\boldsymbol{\kappa}_i \notin [-1,0], \\ \sum_{i=1}^n (1+y_i\boldsymbol{\kappa}_i)\log(1+y_i\boldsymbol{\kappa}_i) - y_i\boldsymbol{\kappa}_i\log(-y_i\boldsymbol{\kappa}_i) \text{ otherwise.} \end{cases}$$
- **The multiclass logistic loss (or softmax)**. The primal variable is now a matrix $\mathbf{Z}$, in $\mathbb{R}^{n\times r}$, which represents the product $\mathbf{X}^\top\mathbf{W}$. We denote by $\mathbf{K}$ the dual variable in $\mathbb{R}^{n\times r}$.
$$\tilde{f}(\mathbf{Z}) = \tfrac{1}{n}\sum_{i=1}^n \log\Big(\sum_{j=1}^r e^{\mathbf{Z}_{ij}-\mathbf{Z}_{i\mathbf{y}_i}}\Big)$$
$$\tilde{f}^*(\mathbf{K}) = \begin{cases} +\infty \text{ if } \exists i \in [1;n] \ \text{ s.t. } \ \{\mathbf{K}_{ij} < 0 \text{ and } j \neq \mathbf{y}_i\} \text{ or } \mathbf{K}_{i\mathbf{y}_i} < -1, \\ \sum_{i=1}^n \Big[\sum_{j\neq\mathbf{y}_i} \mathbf{K}_{ij}\log(\mathbf{K}_{ij}) + (1+\mathbf{K}_{i\mathbf{y}_i})\log(1+\mathbf{K}_{i\mathbf{y}_i})\Big]. \end{cases}$$

Our first remark is that the choice Eq. (47) ensures that $\tilde{f}(\boldsymbol{\kappa})$ is not infinite.

As for the regularization function, except for the Tikhonov regularization which is self-conjugate (it is equal to its Fenchel conjugate), we have considered functions that are norms. There exists therefore a norm $\|.\|$ such that $\psi(\mathbf{w}) = \|\mathbf{w}\|$, and we denote by $\|.\|_*$ its dual-norm. In such a case, the Fenchel conjugate of ψ for a vector $\boldsymbol{\gamma}$ in $\mathbb{R}^p$ takes the form

$$\psi^*(\boldsymbol{\gamma}) = \begin{cases} 0 & \text{if } \|\boldsymbol{\gamma}\|_* \leq 1, \\ +\infty & \text{otherwise.} \end{cases}$$

It turns out that for almost all the norms we have presented, there exists (i) either a closed form for the dual-norm or (ii) there exists an efficient algorithm evaluating it. The only one which does not conform to this statement is the tree-structured sum of ℓ_2-norms, for which we do not know how to evaluate it efficiently.

One can now slightly modify the definition of $\boldsymbol{\kappa}$ to ensure that $\psi^*(-\mathbf{X}\boldsymbol{\kappa}/\lambda) \neq +\infty$. A natural choice is

$$\boldsymbol{\kappa}(\mathbf{w}) \triangleq \min\Big(1, \frac{\lambda}{\|\mathbf{X}\nabla\tilde{f}(\mathbf{X}^\top\mathbf{w})\|_*}\Big)\nabla\tilde{f}(\mathbf{X}^\top\mathbf{w}),$$

which is the one we have implemented. With this new choice, it is easy to see that for all vectors $\mathbf{w}$ in $\mathbb{R}^p$, we still have $\tilde{f}^*(\boldsymbol{\kappa}) \neq +\infty$, and finally, we also have $\delta(\mathbf{w}, \boldsymbol{\kappa}(\mathbf{w})) < +\infty$ and $\delta(\mathbf{w}^\star, \boldsymbol{\kappa}(\mathbf{w}^\star)) = 0$, making it potentially a good duality gap.

A.0.2 Duality Gaps with Intercepts

Even though adding an intercept does seem a simple modification to the original problem, it induces difficulties for finding good dual variables.

We recall that having an intercept is equivalent to having a problem of the type (46), by adding a row to $\mathbf{X}$ filled with the value 1, adding one dimension to the vector $\mathbf{w}$ (or one row for matrices $\mathbf{W}$), and by using a regularization function that does not depend on the last entry of $\mathbf{w}$ (or the last row of $\mathbf{W}$).

Suppose that we are considering a problem of type (46) of dimension $p+1$, but we are using a regularization function $\tilde{\psi} : \mathbb{R}^{p+1} \to \mathbb{R}$, such that for a vector $\mathbf{w}$ in $\mathbb{R}^{p+1}$, $\tilde{\psi}(\mathbf{w}) \triangleq \psi(\mathbf{w}_{[1;p]})$, where $\psi : \mathbb{R}^p \to \mathbb{R}$ is one of the regularization function we have introduced. Then, considering a primal variable $\mathbf{w}$, a dual variable $\boldsymbol{\kappa}$, and writing $\boldsymbol{\gamma} \triangleq -\mathbf{X}\boldsymbol{\kappa}/\lambda$, we are interested in computing

$$\tilde{\psi}^*(\boldsymbol{\gamma}) = \begin{cases} +\infty \text{ if } \boldsymbol{\gamma}_{p+1} \neq 0 \\ \psi^*(\boldsymbol{\gamma}_{[1;p]}) \text{ otherwise,} \end{cases}$$

which means that in order the duality gap not to be infinite, one needs in addition to ensure that $\boldsymbol{\gamma}_{p+1}$ be zero. Since the last row of $\mathbf{X}$ is filled with ones, this writes down $\sum_{i=1}^{p+1} \boldsymbol{\kappa}_i = 0$. For the formulation with matrices $\mathbf{W}$ and $\mathbf{K}$, the constraint is similar but for every column of $\mathbf{K}$.

Let us now detail how we proceed for every loss function to find a "good" dual variable $\boldsymbol{\kappa}$ satisfying this additional constraint, given a primal variable $\mathbf{w}$ in $\mathbb{R}^{p+1}$, we first define the auxiliary function

$$\boldsymbol{\kappa}'(\mathbf{w}) \triangleq \nabla \tilde{f}(\mathbf{X}^\top \mathbf{w}),$$

(which becomes $\mathbf{K}'(\mathbf{W}) \triangleq \nabla \tilde{f}(\mathbf{X}^\top \mathbf{W})$ for matrices), and then define another auxiliary function $\boldsymbol{\kappa}''(\mathbf{w})$ as follows, to take into account the additional constraint $\sum_{i=1}^{p+1} \boldsymbol{\kappa}_i = 0$.

- **For the square loss**, we define another auxiliary function:

  $$\boldsymbol{\kappa}''(\mathbf{w}) \triangleq \boldsymbol{\kappa}'(\mathbf{w}) - \frac{1}{n} \mathbf{1}_{p+1}^\top \boldsymbol{\kappa}'(\mathbf{w}) \mathbf{1}_{p+1}$$

  where $\mathbf{1}_{p+1}$ is a vector of size $p+1$ filled with ones. This step, ensures that $\sum_{i=1}^{p+1} \boldsymbol{\kappa}''(\mathbf{w})_i = 0$.

- **For the logistic loss**, the situation is slightly more complicated since additional constraints are involved in the definition of $\tilde{f}^*$.

  $$\boldsymbol{\kappa}''(\mathbf{w}) \triangleq \operatorname*{arg\,min}_{\boldsymbol{\kappa} \in \mathbb{R}^n} \|\boldsymbol{\kappa} - \boldsymbol{\kappa}'(\mathbf{w})\|_2^2 \text{ s.t. } \sum_{i=1}^{n} \boldsymbol{\kappa}_i = 0 \text{ and } \forall i \in [1;n],\ \boldsymbol{\kappa}_i \in [-1, 0].$$

  This problem can be solved in linear-time [3, **?**] using a similar algorithm as for the projection onto the ℓ_1-ball, since it is an instance of a *quadratic knapsack problem*.

- **For the multi-class logistic loss**, we proceed in a similar way, for every column $\mathbf{K}^j$ of $\mathbf{K}$, $j \in [1;r]$:

  $$\mathbf{K}''^j(\mathbf{w}) \triangleq \operatorname*{arg\,min}_{\boldsymbol{\kappa} \in \mathbb{R}^n} \|\mathbf{K}'^j - \boldsymbol{\kappa}'(\mathbf{w})\|_2^2 \text{ s.t. } \sum_{i=1}^{n} \boldsymbol{\kappa}_i = 0 \text{ and } \forall i \in [1;n],\ \{\boldsymbol{\kappa}_i \geq 0 \text{ if } j \neq \mathbf{y}_i\} \text{ and } \{\boldsymbol{\kappa}_i \geq -1 \text{ if } \mathbf{y}_i = j\}.$$

When the function ψ is the Tykhonov regularization function, we end the process by setting $\boldsymbol{\kappa}(\mathbf{w}) = \boldsymbol{\kappa}''(\mathbf{w})$. When it is a norm, we choose, as before for taking into account the constraint $\|\mathbf{X}\boldsymbol{\kappa}\|_* \leq \lambda$,

$$\boldsymbol{\kappa}(\mathbf{w}) \stackrel{\triangle}{=} \min\Big(1, \frac{\lambda}{\|\mathbf{X}\boldsymbol{\kappa}''(\mathbf{w})\|_*}\Big)\boldsymbol{\kappa}''(\mathbf{w}),$$

with a similar formulation for matrices $\mathbf{W}$ and $\mathbf{K}$.

Even though finding dual variables while taking into account the intercept requires quite a lot of engineering, notably implementing a quadratic knapsack solver, it can be done efficiently.

References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[2] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: Theory and examples*. Springer, 2006.

[3] P. Brucker. An O(n) algorithm for quadratic knapsack problems. 3:163–166, 1984.

[4] E. J. Candès, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *Journal of Fourier Analysis and Applications*, 14:877–905, 2008.

[5] B. V. Cherkassky and A. V. Goldberg. On implementing the push-relabel method for the maximum flow problem. *Algorithmica*, 19(4):390–410, 1997.

[6] S. F. Cotter, J. Adler, B. Rao, and K. Kreutz-Delgado. Forward sequential algorithms for best basis selection. In *IEEE Proceedings of Vision Image and Signal Processing*, pages 235–244, 1999.

[7] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2):407–499, 2004.

[9] J. Friedman, T. Hastie, H. Hölfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of statistics*, 1(2):302–332, 2007.

[10] W. J. Fu. Penalized regressions: The bridge versus the Lasso. *Journal of computational and graphical statistics*, 7:397–416, 1998.

[11] A. V. Goldberg and R. E. Tarjan. A new approach to the maximum flow problem. In *Proc. of ACM Symposium on Theory of Computing*, pages 136–146, 1986.

[12] P. O. Hoyer. Non-negative sparse coding. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, 2002.

[13] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[14] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

[15] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, 2001.

[16] N. Maculan and J. R. G. Galdino de Paula. A linear-time median-finding algorithm for projecting a vector on the simplex of Rn. *Operations research letters*, 8(4):219–222, 1989.

[17] J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, Ecole Normale Supérieure, Cachan, 2010.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[19] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[20] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010.

[21] S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[22] N. Meinshausen and P. Buehlmann. Stability selection. Technical report. ArXiv:0809.2932.

[23] G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22.

[24] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar. Collaborative hierarchical sparse modeling. Technical report, 2010. Preprint arXiv:1003.0400v1.

[25] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, 67(1):91–108, 2005.

[26] J. A. Tropp. Algorithms for simultaneous sparse approximation. part ii: Convex relaxation. *Signal Processing, special issue "Sparse approximations in signal and image processing"*, 86:589–602, April 2006.

[27] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. *Signal Processing, special issue "sparse approximations in signal and image processing"*, 86:572–588, April 2006.

[28] S. Weisberg. *Applied Linear Regression*. Wiley, New York, 1980.

[29] T. T. Wu and K. Lange. Coordinate descent algorithms for Lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.

[30] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 2006.

[31] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, 2005.