Adaptation of Reference Library and Structured Sparse Representations for Spectroscopic Imaging

Jack Xin

Department of Mathematics
University of California, Irvine

Collaborators and Acknowledgements

- Ernie Esser, Mathematics, UC Irvine
- Yifei Lou, Mathematics, UC Irvine
- Penghang Yin, Mathematics, UC Irvine
- Partially supported by NSF
- Thanks to B. Finlayson-Pitts and L. Wingen (Atmospheric Chemistry, UC Irvine)
- Thanks to J. Stutz (Atmospheric Chemistry, UCLA).

Outline

- Demixing problems in spectroscopic imaging when uncertainties exist in reference spectra.
- Sparse representation with group structured (adaptive) basis, limitations of l_1.
- Group sparsity, l_1/l_2 regularization, geometric and analytical properties.
- Optimization algorithms based on l_1/l_2, $l_1 - l_2$ penalties.
- Applications to differential optical absorption spectroscopy (DOAS) and hyperspectral imaging (HSI).

Introduction

- Demixing data b with given reference spectra (columns of full rank matrix A):

$$x = \mathrm{argmin}_{x \geq 0} \|Ax - b\|^2$$

  a nonnegative least squares problem. Uniqueness under sparsity for the under-determined case if columns of A are incoherent enough (Candes, Tao et al 2005; Bruckstein et al 2008). Greedy and L_1 methods (Tibshinani 96, Donohu et al. 98, Tropp 04, Osher et al 08, among others).

- Due to measurement error (e.g. wave length misalignment in DOAS), columns of A contain uncertainty. Multiple reference spectra correspond to the same material, measured under different conditions. No clear which one is optimal to use for mixture data b.

Introduction

- Putting all reference spectra to form A leads to coherence of columns.
- Model uncertainty by enlarging dictionary: including translations and scaled versions of each standard dictionary element (Lou, Bertozzi, Soatto 2011).
- Goal: Select one vector from each group (1 intra-sparsity), minimal number of groups (inter-sparsity).
- Optimization with sparsity promoting penalties (l_1/l_2, $l_1 - l_2$), comparing with l_1 and greedy l_0 method. The ratio norm l_1/l_2 has been used in nonnegative matrix factorization (Hoyer 2002), blind deconvolution and deblurring (Fergus et al; J. Cai, Z. Shen et al; 2011-2012).

Example of Coherent Dictionary and Sparsity

- Let $p \in (0, 1]$ and two distinct dense b^1, $b^2 \in R^n$ $(n \geq 2)$ so that $b = b^1 + b^2$ is also dense; $a = \|(b^1, b^2)\|_p$, $A = [b^1, b^2, a\, I_n, a\, I_n]$, $I_n = n \times n$ identity matrix. Consider $A\,x = b$, $x \in R^{2+2n}$, sparse solutions and their p-norms are:

$$x_s = [1, 1, 0, \cdots, 0]', \;\; sp = 2, \quad \|x_s\|_p = 2^{1/p},$$

$$x'_* = [0, 1, b^1/a, 0]', \quad x''_* = [1, 0, 0, b^2/a]', \; sp \geq 3,$$

$$x_* = [0, 0, b^1/a, b^2/a]', \; sp \geq 4.$$

$$\|x'_*\|_p = (1 + \|b^1\|_p^p / a^p)^{1/p} \in (1, 2^{1/p}),$$

$$\|x''_*\|_p = (1 + \|b^2\|_p^p / a^p)^{1/p} \in (1, 2^{1/p}),$$

$$\|x_*\|_p = \|(b^1, b^2)\|_p / a = 1.$$

Example of Coherent Dictionary and Sparsity

- x_s cannot be recovered by minimizing l_p norm st. $Ax = b$. At least three solutions exist with less sparsity and smaller l_p norm than $\|x_s\|_p$.

- The l_1/l_2 ratio norm for nonnegative x:

$$\|x\|_1/\|x\|_2 = \mathbf{1} \cdot x/\|x\|_2 = \|\mathbf{1}\|_2 \cos \angle(\mathbf{1}, x).$$

  Minimization moves x towards coordinate planes, helping sparsity.

- However, minimizing l_1/l_2 does not give the sparsest solution in general.

Example of Coherent Dictionary and Sparsity

Ratio norms are:

$$\|x_s\|_1/\|x_s\|_2 = \sqrt{2},$$

$$\|x'_*\|_1/\|x'_*\|_2 = \|(a, b^1)\|_1/\|(a, b^1)\|_2,$$

$$\|x''_*\|_1/\|x''_*\|_2 = \|(a, b^2)\|_1/\|(a, b^2)\|_2,$$

$$\|x_*\|_1/\|x_*\|_2 = \|(b^1, b^2)\|_1/\|(b^1, b^2)\|_2.$$

We want $\|x'_*\|_1/\|x'_*\|_2 > \sqrt{2}$ or:

$$(2\|b^1\|_1 + \|b^2\|_1)/((\|b^1\|_1 + \|b^2\|_1)^2 + \|b^1\|_2^2)^{1/2} > \sqrt{2},$$

$$(2\|b^1\|_1 + \|b^2\|_1)^2 > 2(\|b^1\|_1^2 + \|b^2\|_1^2 + 2\|b^1\|_1\|b^2\|_1 + \|b^1\|_2^2),$$

$$2\|b^1\|_1^2 > \|b^2\|_1^2 + 2\|b^1\|_2^2.$$

- Likewise $\|x''_*\|_1/\|x''_*\|_2 > \sqrt{2}$ requires:

$$2\|b^2\|_1^2 > \|b^1\|_1^2 + 2\|b^2\|_2^2.$$

The above inequalities reduce to: $\|b^i\|_1 > \sqrt{2}\|b^i\|_2, \;\; i = 1, 2$, if the first two columns of A satisfy $\|b^1\|_1 = \|b^2\|_1$, $b^1 \neq b^2$.

- Kashin-Garnaev-Gluskin inequality: there exist a set S of $[n/2]$-dimensional subspaces of R^n with probability at least $1 - \exp\{-c_0 n\}$, such that for any $b^i \in S$ $(i = 1, 2)$, $b^i \neq 0$:

$$\|b^i\|_1/\|b^i\|_2 \geq c_1\sqrt{n}/\sqrt{1 + \log 2},$$

where c_0 and c_1 are positive constants independent of n. If $n >$ an absolute constant, x_s has the smallest ratio norm among the 4 solutions, ruling out the counter example to l_1 minimization.

- Minimizing the ratio norm does not always give the sparsest solution. First, for any $y \in \mathbb{R}^n$

$$\mathbf{Ker}(A) = \mathbf{span}\{[1, 0, -b^1/a, 0]', [0, 1, -b^2/a, 0]', [0, 0, -y, y]'\}.$$

Let
$x^* = x_s + [1, 0, -b^1/a, 0]' - [0, 1, -b^2/a, 0]' = [2, 0, (b^2 - b^1)/a, 0]'$.

$$\|x^*\|_1/\|x^*\|_2 \leq \frac{2 + \|b^2 - b^1\|_1/a}{2} < \|x_s\|_1/\|x_s\|_2 = \sqrt{2},$$

if

$$\|b^2 - b^1\|_1/a < 2\sqrt{2} - 2 \approx 0.828.$$

- Not stringent, as $\|b^2 - b^1\|_1/a \leq (\|b^2\|_1 + \|b^1\|_1)/a = 1$.

Example of Coherent Dictionary and Sparsity

- In summary, $x^* = [2, 0, (b^2 - b^1)/a, 0]'$ is a less sparse solution than $x_s = [1, 1, 0, \cdots, 0]'$ with smaller ratio of l_1/l_2 norm (if $b^1 - b^2$ is small enough). Minimization of l_1/l_2 does not yield x_s.
- On the other hand, x^* contains a large peak (height 2), and many smaller peaks ($(b^1 - b^2)/a$) if $b^1 \approx b^2$, resembling a perturbation of 1-sparse solution $[2, 0, \cdots, 0]'$ when $b^1 = b^2$.
- (Continuity) The minimizer of l_1/l_2 goes from exact 1-sparse structure when $b^1 = b^2$ to an approximate 1-sparse structure when $b^1 \approx b^2$.
- (Discreteness) the l_0 minimizer x_s experiences a jump from $[2, 0, 0, 0]'$ to $[1, 1, 0, 0]'$.

Example of Coherent Dictionary and Sparsity

- Discrete character of l_0 makes it subtle to recover the least l_0 solution by minimizing l_1/l_2.
- If we view b^1 and b^2 as dictionary members in a group, minimizing l_1/l_2 selects only one of them (intra sparsity).
- Similarly, if we view corresponding columns (1st and $(n+1)$-th, 2nd and $(n+2)$-th, etc) of $[\alpha I_n \ \ \alpha I_n]$ as vectors in a group (of 2 elements), then x^* selects one member out of each group.
- *Minimizing l_1/l_2 has the tendency of removing redundencies or preferring intra-sparsity in a coherent and over-determined dictionary.* L1 minimization does not do as well in terms of intra-sparsity, using all group elements except for knocking out the b^1, b^2 group.

Exact Recovery of l_1/l_2

- For $x \geq 0 \in \mathbb{R}^n$, let $S = \{i : x_i > 0\}, Z = \{i : x_i = 0\}$, sparsity of x is $|S| = \|x\|_0 = k > 0$.
  Define **uniformity** of x:

$$U(x) = \frac{\min_{i \in S} x_i}{\max_{i \in S} x_i} \leq 1$$

  $U(x) = 1$, if $k = 1$.

- Consider the following two problems:

$$P_0 : \quad \min \|x\|_0, \; s.t. \; Ax = Ax_0$$

$$P_1 : \quad \min \|x\|_1 / \|x\|_2, \; s.t. \; Ax = Ax_0$$

Exact Recovery of l_1/l_2

Theorem

Let $x_0 \geq 0 \in \mathbb{R}^n$, $\|x_0\|_0 = k$, *the unique solution to* P_0. *If*

$$U(x) > \max\{\frac{\sqrt{\|x\|_0} - \sqrt{\|x\|_0 - k}}{\sqrt{\|x\|_0} + \sqrt{\|x\|_0 - k}}, 1/2\}$$

for all $x \neq x_0$ *satisfying* $Ax = Ax_0$, *then* x_0 *uniquely solves* P_1.
In particular, if any feasible solution x *is a binary vector with entries 0 or 1, then the above inequality holds b.c.* $U(x) = 1$. *Clearly,* P_0 *and* P_1 *are equivalent if all* x *are binary, since* $\|x\|_1/\|x\|_2 = \sqrt{\|x\|_0}$.

Exact Recovery of l_1/l_2

- If $U = U(x) \geq 1/2$,

$$\frac{2\sqrt{U}}{1+U}\sqrt{\|x\|_0} \leq \frac{\|x\|_1}{\|x\|_2}$$

- The lower bound condition on U without $1/2$ gives:

$$\sqrt{k} < \frac{2\sqrt{U}}{1+U}\sqrt{\|x\|_0}$$

- Combining:

$$\frac{\|x_0\|_1}{\|x_0\|_2} \leq \sqrt{k} < \frac{2\sqrt{U}}{1+U}\sqrt{\|x\|_0} \leq \frac{\|x\|_1}{\|x\|_2}$$

Variational Models for Group Sparsity

Let the dictionary A have l_2 normalized columns, consist of M groups, each with m_j elements. Write $A = [A_1 \cdots A_M]$ and $x = [x_1 \cdots x_M]^T$, each $x_j \in \mathbb{R}^{m_j}$, $N = \sum_{j=1}^{M} m_j$. The general non-negative least squares problem with sparsity constraints

$$\min_{x \geq 0} F(x) := \frac{1}{2}\|Ax - b\|^2 + R(x), \tag{1}$$

where

$$R(x) = \sum_{j=1}^{M} \gamma_j R_j(x_j) + \gamma_0 R_0(x). \tag{2}$$

Functions R_j represent intra group sparsity penalties applied to each group of coefficients x_j, $j = 1, ..., M$, and R_0 is the inter group sparsity penalty applied to x.

Variational Models for Group Sparsity

- We choose

$$R_j(x_j) = \gamma_j \frac{\|x_j\|_1}{\|x_j\|_2}, \qquad R_0(x) = \gamma_0 \frac{\|x\|_1}{\|x\|_2}$$

or

$$S_j(x_j) = \gamma_j(\|x_j\|_1 - \|x_j\|_2), \qquad S_0(x) = \gamma_0(\|x\|_1 - \|x\|_2)$$

- For analysis of convergence, $\|x_j\|_2$ and $\|x\|_2$ are smoothed near origin, e.g. replacing $\|x\|_2$ by $\phi(x,\epsilon) + \frac{\epsilon}{2}$,

$$\phi(x,\epsilon) = \inf_y \|y\|_2 + \frac{1}{2\epsilon}\|y - x\|^2 = \begin{cases} \frac{\|x\|_2^2}{2\epsilon} & \text{if } \|x\|_2 \leq \epsilon \\ \|x\|_2 - \frac{\epsilon}{2} & \text{otherwise} \end{cases} \quad (3)$$

so called Huber function. Alternatively, the approach of adding dummy variables.

Abstract Model, Convex-Concave Splitting

- Our model is in the form:

$$\min_{x \in X} F(x) := \frac{1}{2}\|Ax - b\|^2 + R(x)$$

where:
(1) X is a convex set,
(2) $R(x) \in C^2(X, R)$ and the eigenvalues of $\nabla^2 R(x)$ are bounded on X.
(3) F is coercive on X: for any $x^0 \in X$, $\{x \in X : F(x) \leq F(x^0)\}$ is a bounded set.

- Convex-concave splitting, $F = F^C + F^E$,

$$F^C(x) = \frac{1}{2}\|Ax - b\|^2 + \|x\|_C^2, \quad F^E(x) = R(x) - \|x\|_C^2$$

for an appropriately chosen positive definite matrix C.

Minimize Upper Bound of Difference

- Quadratic Upper Bound:
  Let λ_r and λ_R be lower and upper bounds resp. of eigenvalues of $\nabla^2 R(x)$ for $x \in X$.

Theorem

Let C be symmetric positive definite and let λ_c denote the smallest eigenvalue of C. If $\lambda_c \geq \lambda_R - \frac{1}{2}\lambda_r$, then for $x, y \in X$,

$$F(y) - F(x) \leq (y - x)^T (\frac{1}{2} A^T A + C)(y - x) + (y - x)^T \nabla F(x).$$

- Iterate

$$x^{n+1} = \arg\min_{x \in X} (x - x^n)^T (\frac{1}{2} A^T A + C_n)(x - x^n) + (x - x^n)^T \nabla F(x^n)$$

for C_n chosen to guarantee a sufficient decrease in F.

Convergence

- $F(x^n)$ is non-increasing $\Longrightarrow$ x^n is bounded.
- $\|x^{n+1} - x^n\| \longrightarrow 0$.
- Any limit point x^* of the sequence $\{x^n\}$ satisfies $(y - x^*)^T \nabla F(x^*) \geq 0$ for all $y \in X$, implying x^* is a stationary point of F.
- Quadratic programming by Alternative Direction Method of Multipliers (ADMM).

DOAS Data

- DOAS is an imaging technique for studying air pollution. It estimates the concentrations of gases in an air mixture by measuring (over a range of wavelengths) the reduction in the intensity of light shined through it.
- Based on Beer's law, given the mixture absorption data $J(\lambda)$ and reference spectra $\{y_j(\lambda)\}$, estimate fitting coefficients $\{a_j\}$ and the deformations $\{v_j(\lambda)\}$ from the model,

$$J(\lambda) = \sum_{j=1}^{M} a_j y_j\big(\lambda + v_j(\lambda)\big) + \eta(\lambda)\ , \tag{4}$$

where $M =$ total number of gases, η Gaussian noise.

DOAS Data

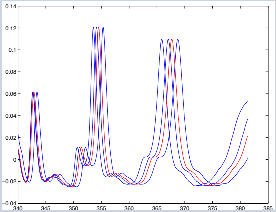
- Construct a dictionary by deforming each y_j with a set of possible deformations.
- Approximate deformation by linear functions $v_j(\lambda) = p_j\lambda + q_j$, enumerate all possible deformations by choosing p_j, q_j from two pre-determined sets $\{P_1, \cdots, P_K\}$, $\{Q_1, \cdots, Q_L\}$.
- Let A_j be a matrix whose columns are deformations of the jth reference $y_j(\lambda + P_k\lambda + Q_l)$ for $k = 1, \cdots, K$ and $l = 1, \cdots, L$. Rewrite the model as:

$$J = [A_1, \cdots, A_M] \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix} + \eta \, , \tag{5}$$

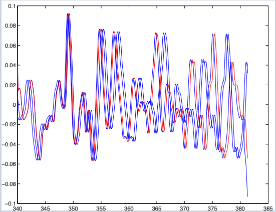
where $x_j \in \mathbb{R}^{KL}$ and $J \in \mathbb{R}^W$.

Dictionary Elements on HONO, NO2, O3

HONO



NO2



O3

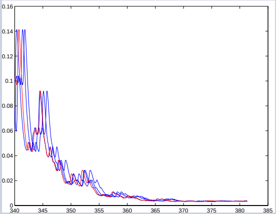


Figure: reference spectrum in red, three deformed spectra are in blue.

Test and Comparison

- Total 441 linearly deformed references in each of the three groups.
- Randomly select one element for each group with random magnitude plus additive zero mean Gaussian noise to synthesize the input data $J(\lambda) \in \mathbb{R}^W$ for $W = 1024$.
- random mixing magnitudes chosen at different orders with mean values of 1, 0.1, 1.5 for HONO, NO2 and O3 respectively.
- std of noise $\eta = 0.05$, $\epsilon_j = 0.05$ for all three groups, $\gamma_j = 0.1$.

Comparison of Sparse Selection

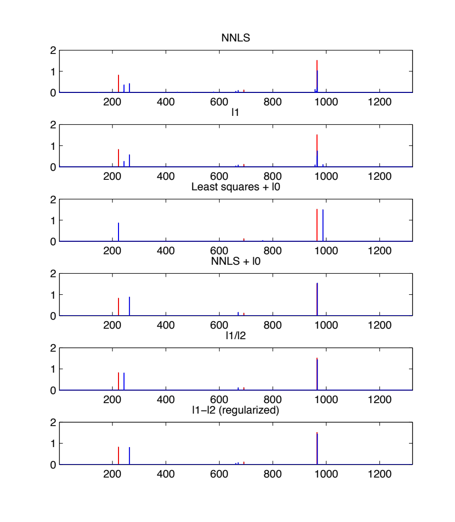
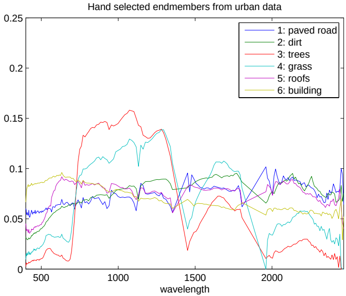


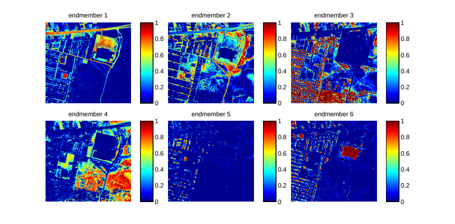
Figure: computed (blue) on top of ground truth (red).

Urban hyperspectral image and dictionary elements

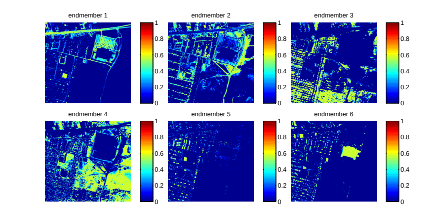


Images of Rows of Abundance Matrix —Fraction Planes

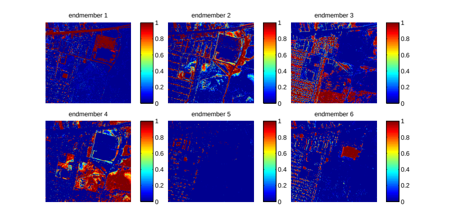
NNLS



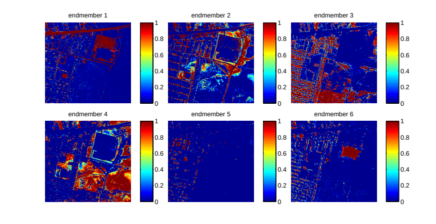
l_1



l_1/l_2



l_1 - l_2



Comparison of Concentration (Abundance)

- l_1 penalty promotes sparse solutions by trying to move coefficient vectors (concentration or abundance values) perpendicular to the positive face of the l_1 ball, shrinking the magnitudes of all elements.
- l_1/l_2 penalty, to some extent l_1 - l_2, promote sparsity by trying to move tangent to the l_2 ball. They are better at preserving the magnitudes of abundances while enforcing a similarly sparse solution. This is reflected in their lower sum of squares errors.
- Fraction nonzero (NNLS, L1, L1/L2, L1-L2) = (0.4752, 0.2683, 0.2645, 0.2677).
- Sum of Sq Error = (1111.2, 19107, 1395.3, 1335.6).

Conclusion and Future Work

- Studied variational method for linear demixing problems where the dictionary contains multiple references for each material and we want to collaboratively choose the best one for each material present.
- Analyzed and used l_1/l_2 and $l_1 - l_2$ penalties to obtain structured sparse solutions to non-negative least squares problems, reformulated as constrained minimization problems with differentiable but non-convex objectives.
- Exact recovery of l_1/l_2 minimization and convergence properties of algorithms.
- Study how to include relative likelihood of candidate references if certain prioir information is known, explore alternative sparsity penalties that can be adapted to the data set.