Supplementary Material for *High-resolution Hyperspectral Imaging via Matrix Factorization*

November 18, 2010

In this supplementary note, we describe in more detail our approach to constructing a factorization $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} \in \mathbb{R}^{m\times p}$ of the given observation into a sparsifying basis $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ of sparse coefficients. We formulate this problem as a constrained optimization which seeks a basis $\boldsymbol{A}$ which best sparsifies the observations, subject to the data constraint $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X}$. We also impose the additional constraint that each of the columns of $\boldsymbol{A}$ have unit ℓ^2-norm, to eliminate scale ambiguities between the two unknowns $\boldsymbol{A}$, $\boldsymbol{X}$. This leaves us with a manifold of possible solution pairs

$$\mathcal{M} = \{(\boldsymbol{A}, \boldsymbol{X}) \mid \boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X},\ \|\boldsymbol{A}\boldsymbol{e}_i\|_2 = 1\ \forall\ i\} \subset \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p}. \tag{0.1}$$

Motivated by the success of ℓ^1-minimization in recovering signals which are sparse in a known basis [BDE08], as well as emerging theoretical results on the good properties of ℓ^1-minimization for sparse matrix factorization [GS10], relax the search for the best sparsifying basis $\boldsymbol{A}$ to the following optimization problem:

$$\text{minimize } \|\boldsymbol{X}\|_1 \quad \text{subject to} \quad (\boldsymbol{A}, \boldsymbol{X}) \in \mathcal{M}. \tag{0.2}$$

This problem asks that we minimize a convex function, subject to a nonlinear (nonconvex) constraint. In the worst case, this is a very difficult problem. However, surprisingly encouraging empirical results suggest that there are large classes of problems that actually can be solved globally via this approach. These problems correspond to instances where the input $\boldsymbol{Y}$ is indeed generated by a sufficiently sparse $\boldsymbol{X}$, and the dimension is sufficiently large (say, $m \geq 20$ for a 3-sparse $\boldsymbol{X}$). Figure 1 (right) shows one example of this behavior: for well-structured simulated problems, locally minimizing the ℓ^1-norm correctly recovers the globally optimal solution $(\boldsymbol{A}, \boldsymbol{X})$ with high probability. We will report on this phenomenon more extensively in a separate, forthcoming work.

However, this still leaves us with a very challenging optimization problem: the number of variables is very large – potentially in the hundreds of thousands or even millions – and the objective function is nonsmooth. Our approach to this problem is a very simple and natural one, which has been studied in the optimization literature at least since the 1970's, as a nonsmooth Gauss-Newton method [Cro78, JO80]. This approach repeatedly linearizes the constraint $\boldsymbol{x} \doteq (\boldsymbol{A}, \boldsymbol{X}) \in \mathcal{M}$, replacing $\mathcal{M}$ with its tangent space at the current iterate. This yields a simple iteration:[1]

$$\boldsymbol{\delta}_k = \arg\min \left\{ \|\boldsymbol{X} + \boldsymbol{\Delta}_X\|_1 \mid (\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X) \in T_{\boldsymbol{x}_k}\mathcal{M},\ \|\boldsymbol{\Delta}_A\|_F^2 + \|\boldsymbol{\Delta}_X\|_F^2 \leq \eta_k \right\} \tag{0.3}$$

$$\boldsymbol{x}_{k+1} = \mathcal{P}_{\mathcal{M}}[\boldsymbol{x}_k + \boldsymbol{\delta}_k], \tag{0.4}$$

[1] We have identified $T_{\boldsymbol{x}_k}\mathcal{M}$ with a subspace of $\mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p}$ in the natural way.

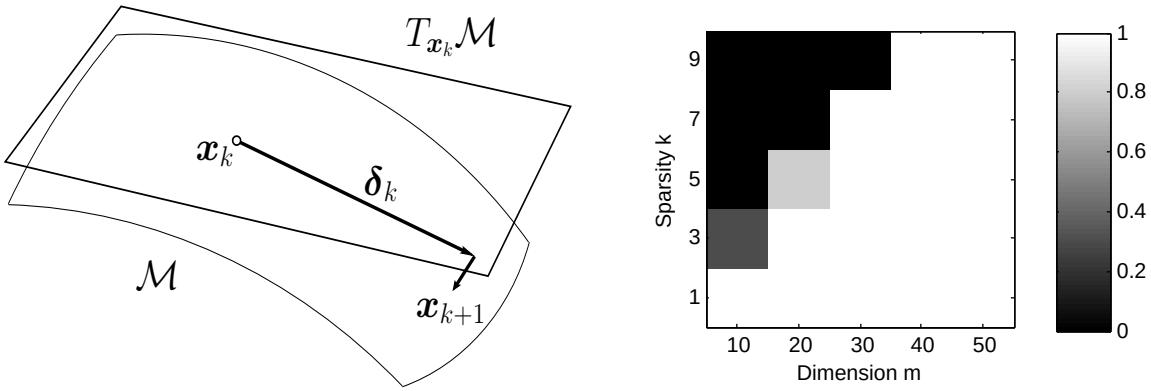

Figure 1: **Left: Illustration of our method.** At each iteration k, our algorithm linearizes the constraint manifold $\mathcal{M}$, and finds the perturbations $\boldsymbol{\delta}_k = (\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X)$ within this linearized space that make the objective $\|\boldsymbol{X} + \boldsymbol{\Delta}_X\|_1$ as small as possible. This minimizing point, $\boldsymbol{x}_k + \boldsymbol{\delta}_k$, is then projected back onto $\mathcal{M}$ to yield the next iterate, $\boldsymbol{x}_{k+1}$. **Right: Illustration of correct recovery.** We generate synthetic examples with randomly chosen $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, for varying $m = 10, 20, \ldots, 50$, and $\boldsymbol{X} \in \mathbb{R}^{m\times p}$ of varying sparsity $k = 1, 3, \ldots, 9$. Recovery is considered correct if the relative errors in $\boldsymbol{A}$ and $\boldsymbol{X}$ are both smaller than 10^{-6}. The figure plots the fraction of correct recoveries over 10 independent trials - white corresponds to perfect recovery in all trials. We observe that when the dimension is large enough, the algorithm indeed correctly recovers $(\boldsymbol{A}, \boldsymbol{X})$ with high probability.

where $\mathcal{P}_{\mathcal{M}} : \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p}$ is a projection operator onto $\mathcal{M}$. The basic idea of this approach is illustrated in Figure 1. This algorithm is known to converge quadratically in the neighborhood of any strict local minimum [JO80]. We describe the two key steps of the algorithm in more details below.

Linearized Subproblem: Solving the linearized subproblem (0.3) amounts to solving a large, equality-constrained ℓ^1-norm minimization problem. It is not difficult to show that the tangent space $T_{\boldsymbol{x}}\mathcal{M}$ at a given point $\boldsymbol{x} = (\boldsymbol{A}, \boldsymbol{X})$ can be expressed as the set of pairs $(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X)$ satisfying

$$\boldsymbol{A}\boldsymbol{\Delta}_X + \boldsymbol{\Delta}_A\boldsymbol{X} = 0, \quad \langle \boldsymbol{A}\boldsymbol{e}_i\boldsymbol{\Delta}_A, \boldsymbol{e}_i\rangle = 0, \ \forall\, i. \tag{0.5}$$

Hence, in more concrete form, the subproblem that we need to solve at each step is

$$\begin{aligned} \text{minimize } \ \|\boldsymbol{X} + \boldsymbol{\Delta}_X\|_1 \quad \text{subject to} \qquad & \boldsymbol{A}\boldsymbol{\Delta}_X + \boldsymbol{\Delta}_A\boldsymbol{X} = 0, \\ & \langle \boldsymbol{A}\boldsymbol{e}_i\boldsymbol{\Delta}_A, \boldsymbol{e}_i\rangle = 0 \ \forall\ i, \\ & \|\boldsymbol{\Delta}_A\|_F^2 + \|\boldsymbol{\Delta}_X\|_F^2 \le \eta^2. \end{aligned} \tag{0.6}$$

For this purpose, we employ an Augmented Lagrange Multiplier (ALM) algorithm [Ber82], which was introduced into the literature on ℓ^1-minimization by Yin, Osher and collaborators [YOGD08], and studied by those authors as a Bregman iterative algorithm. To describe this algorithm more fully, it is helpful to introduce a linear operator $\Psi : \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p} \to \mathbb{R}^{m\times p} \times \mathbb{R}^n$ corresponding to the above tangent space constraint:

$$\Psi(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X) = (\boldsymbol{A}\boldsymbol{\Delta}_X + \boldsymbol{\Delta}_A\boldsymbol{X}, \ \mathrm{diag}[\boldsymbol{A}^*\boldsymbol{\Delta}_A]). \tag{0.7}$$

Introduce the convex function $J : \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times p}$ via

$$J(\boldsymbol{\Delta}_A, \boldsymbol{\Delta}_X) \doteq \begin{cases} \|\boldsymbol{X} + \boldsymbol{\Delta}_X\|_1, & \|\boldsymbol{\Delta}_X\|_F^2 + \|\boldsymbol{\Delta}_A\|_F^2 \leq \eta^2 \\ \infty, & \text{else.} \end{cases} \tag{0.8}$$

The Bregman iterative approach solves a sequence of convex programs in unknowns $\boldsymbol{h} = (\boldsymbol{H}_A, \boldsymbol{H}_X)$, with varying inputs $\boldsymbol{r}_j = (\boldsymbol{R}_A, \boldsymbol{R}_X)$, starting from $\boldsymbol{r}_0 = (\boldsymbol{0}, \boldsymbol{0})$:

$$\boldsymbol{h}_{j+1} \quad = \quad \arg\min_{\boldsymbol{h}} \; J(\boldsymbol{h}) \; + \; \lambda\|\Psi(\boldsymbol{h}) - \boldsymbol{r}_j\|, \tag{0.9}$$

and updating the residual via $\boldsymbol{r}_{j+1} = \boldsymbol{r}_j + \Psi(\boldsymbol{h}_{j+1}) - \Psi(\boldsymbol{h}_j)$. This sequence of iterates converges to the solution to (0.6). Each of these problems (0.9) is solved via a simple soft-thresholding algorithm, which repeatedly sets

$$\boldsymbol{z}_{i+1} \; = \; \arg\min_{\boldsymbol{z}} \; J(\boldsymbol{z}) + \frac{\lambda}{2\beta}\|\boldsymbol{z} - \beta\Psi(\boldsymbol{z}_i - \boldsymbol{r}_j)\|. \tag{0.10}$$

This problem is a variant of shrinkage or soft-thresholding that arises in ℓ^1-minimization. It can be solved via an efficient projection algorithm similar to the one given in [MBPS10].

Projection: We project a given pair $\boldsymbol{A}, \boldsymbol{X}$ onto $\mathcal{M}$ as follows. We first scale the columns of $\boldsymbol{A}$ to obtain a matrix $\boldsymbol{A}'$ whose columns have unit ℓ^2 norm. We then select $\boldsymbol{X}'$ to be the matrix $\tilde{\boldsymbol{X}}$ satisfying $\boldsymbol{Y} = \boldsymbol{A}'\tilde{\boldsymbol{X}}$ which closest to $\boldsymbol{X}$ in Frobenius norm. In notation:

$$\boldsymbol{X}' = \boldsymbol{X} + (\boldsymbol{A}')^\dagger(\boldsymbol{Y} - \boldsymbol{A}'\boldsymbol{X}). \tag{0.11}$$

We then set $\mathcal{P}(\boldsymbol{A}, \boldsymbol{X}) = (\boldsymbol{A}', \boldsymbol{X}')$.

It should be noted that this operator does not find a pair $(\boldsymbol{A}', \boldsymbol{X}')$ that is closest to $(\boldsymbol{A}, \boldsymbol{X})$ in terms of the ℓ^2 norm $\|(\boldsymbol{A}, \boldsymbol{X})\| = (\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{X}\|_F^2)^{1/2}$. Finding a closest point on $\mathcal{M}$ in terms of this (rather natural) norm would require solving a separate nonconvex optimization problem. Rather, our operator $\mathcal{P}(\boldsymbol{A}, \boldsymbol{X})$ can be viewed as the limit (as $t \to \infty$) of projection operators with respect to weighted ℓ^2 norms $\|(\boldsymbol{A}, \boldsymbol{X})\| = (t\|\boldsymbol{A}\|_F^2 + \|\boldsymbol{X}\|_F^2)^{1/2}$. This substitution leaves us with a simple, tractable operator $\mathcal{P}$, and does not affect the qualitative behavior of the algorithm.

Initialization: We find it surprisingly effective to simply start the algorithm from a random initialization. We choose $\boldsymbol{A}_0$ to be a matrix whose columns are independent random samples from the uniform distribution on the sphere $\mathbb{S}^{m-1}$. We then simply set $\boldsymbol{X}_0 = \boldsymbol{A}_0^\dagger\boldsymbol{Y}$.

References

[BDE08] A.M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *To appear in SIAM Review*, 2008.

[Ber82] D. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982.

[Cro78] L. Cromme. Strong uniqueness: A far-reaching criterion for the convergence analysis of iterative procedures. *Numerishe Mathematik*, 29:179–193, 1978.

[GS10] R. Gribonval and K. Schnass. Dictionary identification - sparse matrix factorization via ℓ_1-minimization. *http://arxiv.org/abs/0904.4774*, 2010.

[JO80] K. Jittorntrum and M. Osborne. Strong uniqueness and second order convergence in nonlinear discrete approximation. *Numerische Mathematik*, 34:439–455, 1980.

[MBPS10] 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.

[YOGD08] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ^1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences (SJIS)*, 2008.