Nonnegative matrix factorization with regularization and sparsity-enforcing terms

Henri Lantéri
Laboratoire H.Fizeau - UMR 6525
Universite de Nice-Sophia Antipolis
06108 Nice Cedex 2 - France
Email: Henri.Lanteri@unice.fr

Céline Theys
Laboratoire H.Fizeau - UMR 6525
Universite de Nice-Sophia Antipolis
06108 Nice Cedex 2 - France
Email: Celine.Theys@unice.fr

Cédric Richard
Laboratoire H.Fizeau - UMR 6525
Universite de Nice-Sophia Antipolis
06108 Nice Cedex 2 - France
Email: Cedric.Richard@unice.fr

***Abstract*—The aim of this paper is to present several multiplicative algorithms for nonnegative matrix factorization. We show how to obtain such algorithms in the case where non-negativity and flux conservation constraints are imposed, and how to regularize such problems by introducing smoothness or sparsity properties. Application to hyperspectral imagery is finally considered.**

I. Introduction of the physical context

Hyperspectral imaging has received considerable attention in the last few years. See for instance [1], [2] and references therein. It consists of data acquisition with high sensitivity and resolution in hundreds contiguous spectral bands, geo-referenced within the same coordinate system. With its ability to provide extremely detailed data regarding the spatial and spectral characteristics of a scene, this technology offers immense new possibilities in collecting and managing information for civilian and military application areas.

Each vector pixel of an hyperspectral image characterizes a local spectral signature. Usually, one consider that each vector pixel can be modeled accurately as a linear mixture of different pure spectral components, called endmembers. Referring to our notations in Equation (1), each column of $\mathbf{V}$ can thus be interpreted as a spectral signature obtained by linear mixing of the spectra of endmembers, i.e., the columns of $\mathbf{W}$. Nonnegative matrix factorization (NMF) consists of estimating the endmember spectra $\mathbf{W}$ and the abundance coefficients $\mathbf{H}$ from the spectral signatures $\mathbf{V}$, subject to non-negative constraints on the entries of $\mathbf{W}$ and $\mathbf{H}$, and sum-to-one constraints on the columns of $\mathbf{W}$.

In this paper, we propose multiplicative interior-point algorithms that can be used within this context. We also show how to regularize the problem by introducing smoothness or sparsity constraints on the columns of $\mathbf{W}$ and $\mathbf{H}$ respectively. The paper is organized as follows. In Section 2, we describe the problem at hand and notations for non-negative matrix factorization. In Section 3, we recall the main steps of the Split Gradient Method (SGM). In Sections 4 and 5, we show how to handle non-negativity and sum-to-one constraints, and how to incorporate regularization. Finally, application to hyperspectral imagery is briefly considered in section 6.

II. Nonnegative matrix factorization

Here we consider the problem of nonnegative matrix factorization, which is now a popular dimension reduction technique, employed for non-subtractive, part-based representation of non-negative data. Given a data matrix $\mathbf{V}$ of dimension $F \times N$ with nonnegative entries, the NMF problem consists of seeking a factorization of the form

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

where $\mathbf{W}$ and $\mathbf{H}$ are two matrices with non negative entries, of dimensions $F \times K$ and $K \times N$, respectively. Dimension K is usually chosen such that $FK + KN \ll FN$, hence reducing the data dimensionality.

Factorization (1) is usually sought through the minimization problem

$$\min_{\mathbf{W},\mathbf{H}} \mathcal{D}(\mathbf{V}, \mathbf{WH}) \quad \text{s.t.} \quad [\mathbf{W}]_{ij} \geq 0, [\mathbf{H}]_{ij} \geq 0 \tag{2}$$

where $\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \sum_{i,j} \delta([\mathbf{V}]_{ij}|[\mathbf{WH}]_{ij})$ is a separable measure of fit. We shall consider that $\delta(x|y)$ is a convex positive function of $y \in \mathbb{R}_+$ given $x \in \mathbb{R}_+$. In this paper, to restrict the solutions space, we shall show how this framework can be extended to deal with flux constraints of the form:

$$\sum_i [\mathbf{W}]_{ij} = 1 \qquad \sum_i [\mathbf{H}]_{ij} = \sum_i [\mathbf{V}]_{ij} \tag{3}$$

The algorithmic approach described hereafter is alternately applied on $\mathbf{W}$ and $\mathbf{H}$ in order to solve (2), (3).

III. Introduction to the Split Gradient Method

In [3], we proposed the Split Gradient Method (SGM) to derive multiplicative algorithms dedicated to NMF for any convex cost function $\mathcal{D}$. As the SGM only takes into account non-negativity constraints, we also presented the flux-constrained SGM in order to satisfy constraints (3).

We shall now briefly recall the main steps of the SGM within the NMF context. See [3], [4], [5] for more details. The SGM is based on the Karush-Kuhn-Tucker conditions at the optimum $\mathbf{W}^*$ and $\mathbf{H}^*$. Restricting our attention

to minimization w.r.t $\mathbf{H}$ (minimization w.r.t $\mathbf{W}$ follows exactly the same scheme), we have

$$[\mathbf{H}^*]_{ij}[\nabla_{\mathbf{H}}\mathcal{D}(\mathbf{V},\mathbf{W}\mathbf{H}^*)]_{ij}=0 \\ \Leftrightarrow [\mathbf{H}^*]_{ij}[-\nabla_{\mathbf{H}}\mathcal{D}(\mathbf{V},\mathbf{W}\mathbf{H}^*)]_{ij}=0 \quad (4)$$

The second expression in Equation (4), while trivial, makes apparent the descent direction with the minus sign. In order to solve this equation iteratively, three points have to be noticed. The first one is that $\mathbf{M}\cdot(-\nabla_{\mathbf{H}}\mathcal{D})$ is a gradient-related descent direction of criterion $\mathcal{D}$ if $\mathbf{M}$ is a matrix with positive entries, where $\cdot$ denotes the Hadamard product. The second one is that $[-\nabla_{\mathbf{H}}\mathcal{D}]_{ij}$ can always be decomposed as

$$[-\nabla_{\mathbf{H}}\mathcal{D}]_{ij}=[\mathbf{R}]_{ij}-[\mathbf{S}]_{ij}, \quad (5)$$

where $[\mathbf{R}]_{ij}$ and $[\mathbf{S}]_{ij}$ are positive entries. Last but not least, the third one is that equations of the form $\varphi(\mathbf{H})=0$ can be solved with a fixed-point algorithm, under some conditions on φ, by considering $\mathbf{H}=\mathbf{H}+\varphi(\mathbf{H})$. Implementing this fixed-point strategy with (4) and using

$$[\mathbf{M}]_{ij}=\frac{1}{[\mathbf{S}]_{ij}} \quad (6)$$

we obtain the gradient-related descent algorithm

$$[\mathbf{H}^{k+1}]_{ij}=[\mathbf{H}^k]_{ij}+\alpha^k\frac{[\mathbf{H}^k]_{ij}}{[\mathbf{S}^k]_{ij}}\left([\mathbf{R}^k]_{ij}-[\mathbf{S}^k]_{ij}\right) \quad (7)$$

with α^k a positive step size that allows to control convergence of the algorithm. This step size can be determined by simplified search methods, [6], in the range $[0,\alpha^k_{\max}]$ where $\alpha^k_{\max}$ is explicitly computed from (7) to ensure the non-negativity of the components of $\mathbf{H}^{k+1}$. It can be easily shown that $\alpha^k_{\max}$ is always greater than one. For $\alpha^k=1$, we obtain the simple and well-known multiplicative form

$$[\mathbf{H}^{k+1}]_{ij}=[\mathbf{H}^k]_{ij}\frac{[\mathbf{R}^k]_{ij}}{[\mathbf{S}^k]_{ij}}. \quad (8)$$

The positiveness of $[\mathbf{H}^{k+1}]_{ij}$ is thus ensured if $[\mathbf{H}^0]_{ij}>0$. Obviously, this setup does not guarantee the convergence of the algorithm.

IV. Flux conservation constraints

We propose to extend the SGM algorithm in order to deal with flux conservation constraints (3). This can be achieved by using the following normalized variables

$$[\mathbf{W}]_{ij}=\frac{[\mathbf{Z}]_{ij}}{\sum_m[\mathbf{Z}]_{mj}} \quad (9)$$

$$[\mathbf{H}]_{ij}=[\mathbf{T}]_{ij}\times\frac{\sum_m[\mathbf{V}]_{mj}}{\sum_m[\mathbf{T}]_{mj}}. \quad (10)$$

The optimization problem becomes unconstrained with respect to the flux conservation constraints, and we can proceed as with the SGM algorithm. Note that the measure of fit $\mathcal{D}$ is still convex in the normalized variables. The gradient versus these new variables expresses as:

$$-\frac{\partial\mathcal{D}}{\partial[\mathbf{Z}]_{\ell j}}=\frac{1}{\sum_m[\mathbf{Z}]_{mj}} \\ \left(\left(-\frac{\partial\mathcal{D}}{\partial[\mathbf{W}]_{\ell j}}\right)-\sum_i[\mathbf{W}]_{ij}\left(-\frac{\partial\mathcal{D}}{\partial[\mathbf{W}]_{ij}}\right)\right) \quad (11)$$

and

$$-\frac{\partial\mathcal{D}}{\partial[\mathbf{T}]_{\ell j}}=\frac{\sum_m[\mathbf{V}]_{mj}}{\sum_m[\mathbf{T}]_{mj}} \\ \left(\left(-\frac{\partial\mathcal{D}}{\partial[\mathbf{H}]_{\ell j}}\right)-\frac{\sum_i[\mathbf{T}]_{ij}}{\sum_m[\mathbf{T}]_{mj}}\left(-\frac{\partial\mathcal{D}}{\partial[\mathbf{H}]_{ij}}\right)\right) \quad (12)$$

Let us note that any shift, for all (i,j), of the form

$$(-\partial\mathcal{D}/\partial[\mathbf{W}]_{ij})_{\text{shift}}= \\ (-\partial\mathcal{D}/\partial[\mathbf{W}]_{ij})-\min_{ij}(-\partial\mathcal{D}/\partial[\mathbf{W}]_{ij})+\epsilon \quad (13)$$

and

$$(-\partial\mathcal{D}/\partial[\mathbf{H}]_{ij})_{\text{shift}}= \\ (-\partial\mathcal{D}/\partial[\mathbf{H}]_{ij})-\min_{ij}(-\partial\mathcal{D}/\partial[\mathbf{H}]_{ij})+\epsilon \quad (14)$$

where ϵ is a (small) positive constant, leaves (11) and (12) unchanged. This shift, however, ensures the positiveness of $(-\partial\mathcal{D}/\partial[\mathbf{W}]_{ij})_{\text{shift}}$ and $(-\partial\mathcal{D}/\partial[\mathbf{H}]_{ij})_{\text{shift}}$. Considering again the SGM algorithm, and restricting our attention to minimization w.r.t $\mathbf{H}$, decomposition (5) can be obtained as

$$[\mathbf{R}]_{ij}=\frac{\sum_m[\mathbf{V}]_{mj}}{\sum_m[\mathbf{T}]_{mj}}\left(-\frac{\partial\mathcal{D}}{\partial[\mathbf{H}]_{ij}}\right)_{\text{shift}} \quad (15)$$

$$[\mathbf{S}]_{ij}=\frac{\sum_m[\mathbf{V}]_{mj}}{\sum_m[\mathbf{T}]_{mj}}\frac{\sum_i[\mathbf{T}]_{ij}}{\sum_m[\mathbf{T}]_{mj}}\left(-\frac{\partial\mathcal{D}}{\partial[\mathbf{H}]_{ij}}\right)_{\text{shift}} \quad (16)$$

Let us note that $[\mathbf{S}]_{ij}$ does not dependent on i. The optimization algorithm on $[\mathbf{T}]_{ij}$ deduced from (7) writes

$$[\mathbf{T}^{k+1}]_{\ell j}=[\mathbf{T}^k]_{\ell j}+\alpha^k[\mathbf{T}^k]_{\ell j} \\ \times\left(\frac{(-\partial\mathcal{D}/\partial[\mathbf{H}^k]_{\ell j})_{\text{shift}}}{\frac{\sum_i[\mathbf{T}^k]_{ij}}{\sum_m[\mathbf{T}^k]_{mj}}(-\partial\mathcal{D}/\partial[\mathbf{H}^k]_{ij})_{\text{shift}}}-1\right) \quad (17)$$

It can be checked that

$$\sum_\ell[\mathbf{T}^{k+1}]_{\ell j}=\sum_\ell[\mathbf{T}^k]_{\ell j}. \quad (18)$$

for all α^k. Transforming back to the initial variables, we finally have

$$[\mathbf{H}^{k+1}]_{\ell j}=[\mathbf{H}^k]_{\ell j}+\alpha^k[\mathbf{H}^k]_{\ell j} \\ \times\left(\sum_m[\mathbf{V}]_{mj}\frac{(-\partial\mathcal{D}/\partial[\mathbf{H}^k]_{\ell j})_{\text{shift}}}{\sum_i[\mathbf{H}^k]_{ij}(-\partial\mathcal{D}/\partial[\mathbf{H}^k]_{ij})_{\text{shift}}}-1\right) \quad (19)$$

We observe that

$$\sum_i [\mathbf{H}^0]_{ij} = \sum_i [\mathbf{V}]_{ij} \Longrightarrow \sum_i [\mathbf{H}^k]_{ij} = \sum_i [\mathbf{V}]_{ij} \quad \forall k \quad (20)$$

At each iteration, the entries of each column of $\mathbf{H}^k$ are positive and their sum constant. The ℓ_1-norm of the columns of $\mathbf{H}^k$ is thus fixed, which encourages sparsity. Analogous derivations can be done for $\mathbf{W}$, with or without flux conservation constraints. We shall now show how to incorporate constraints on the columns of matrices $\mathbf{H}$ and $\mathbf{W}$ via Tikhonov regularization and sparsity-enforcing terms. Other penalty functions are, however, possible.

V. Regularization

The regularization penalty terms are incorporated separately on the columns of $\mathbf{W}$ and $\mathbf{H}$, and are added to the data consistency term $\mathcal{D}(\mathbf{V}, \mathbf{WH})$. The penalized objective function expresses as

$$\mathcal{D}_{\text{reg}}(\mathbf{V}, \mathbf{WH}) = \mathcal{D}(\mathbf{V}, \mathbf{WH}) + \gamma_1 \mathcal{F}_1(\mathbf{W}) + \gamma_2 \mathcal{F}_2(\mathbf{H}) \quad (21)$$

where $\mathcal{F}_1(\mathbf{W})$ and $\mathcal{F}_2(\mathbf{H})$ are penalty functions, and γ_1 and γ_2 the regularization factors.

A. Tikhonov smoothness regularization

Such a regularization, which expresses some smoothness property of the solution, mainly applies to the endmember spectra, that is, on the columns of $\mathbf{W}$. We can consider

$$\mathcal{F}_1(\mathbf{W}) = \frac{1}{2} \sum_{ij} ([\mathbf{W}]_{ij} - c)^2 \quad (22)$$

possibly with $c = 0$, or

$$\mathcal{F}_1(\mathbf{W}) = \frac{1}{2} \sum_{ij} [\partial_{1,2} \mathbf{W}]_{ij}^2 \quad (23)$$

where $\partial_{1,2}$ is the first or second-order derivative operator. For simplicity, we approximate $\partial_{1,2}\mathbf{W}$ in closed numerical form as

$$[\partial_{1,2}\mathbf{W}]_{ij} = [\mathbf{W}]_{ij} - [\mathbf{AW}]_{ij} \quad (24)$$

where $\mathbf{AW}$ stands for the convolution of each column of matrix $\mathbf{W}$ by a mask, e.g., [1 0 0] and $[\frac{1}{2}\ 0\ \frac{1}{2}]$ for the first and second-order derivative operators, respectively. The opposite of the gradient can be expressed in matrix form as follows

$$-[\nabla_{\mathbf{W}} \mathcal{F}_1]_{ij} = [(\mathbf{A} + \mathbf{A}^\top)\mathbf{W}]_{ij} - [(\mathbf{A}^\top \mathbf{A} + \mathbf{I})\mathbf{W}]_{ij}. \quad (25)$$

This expression makes the decomposition (5) required by the SGM algorithm explicit since it consists of a difference between two positive terms.

Note that Tikhonov regularization with the basic SGM algorithm was initially associated to the basic SGM algorithm in [7], i.e., without flux constraint. The interested reader is invited to consult this reference for an overview of the results that have been obtained.

B. Sparsity-enforcing function

Such a penalty, which expresses that most of information may be concentrated in a few coefficients, mainly applies to the abundance coefficients, that is, to the columns of $\mathbf{H}$. Keeping in mind that the algorithm satisfies flux conservation constraint, see (20), we are ready to consider the following sparsity measure σ introduced by Hoyer [8]

$$\sigma = \frac{\sqrt{K} - \frac{\|[\mathbf{H}]_{\bullet j}\|_1}{\|[\mathbf{H}]_{\bullet j}\|_2}}{\sqrt{K} - 1}, \qquad 0 \leq \sigma \leq 1 \quad (26)$$

with K the number of rows of $\mathbf{H}$, and $[\mathbf{H}]_{\bullet j}$ its j-th row. This clearly defines a relation between the ℓ_2-norm and the ℓ_1-norm of $[\mathbf{H}]_{\bullet j}$, which remains constant along iterations.

$$\|[\mathbf{H}]_{\bullet j}\|_2^2 = \alpha^2 \|[\mathbf{H}]_{\bullet j}\|_1^2 \quad (27)$$

with

$$\alpha = \frac{1}{\sqrt{K} - \sigma(\sqrt{K} - 1)}, \qquad \frac{1}{\sqrt{K}} \leq \alpha \leq 1. \quad (28)$$

Note that only two values for σ lead to unambiguous situations; If $\alpha = 1$, only one entry of $[\mathbf{H}]_{\bullet j}$ is nonzero; If $\alpha = 1/\sqrt{K}$, all the entries of $[\mathbf{H}]_{\bullet j}$ are equal. Any other value for α can correspond to different sets of entries. As a consequence, we suggest to consider the following penalty function[1]

$$\mathcal{F}_2(\mathbf{H}) = \frac{1}{2} \sum_j \left(\|[\mathbf{H}]_{\bullet j}\|_2^2 - \alpha^2 \|[\mathbf{H}]_{\bullet j}\|_1^2 \right)^2 \quad (29)$$

with $\alpha = 1$, and use the regularization factor γ_2 in (21) to push $[\mathbf{H}]_{\bullet j}$ toward a sparse solution. For convenience, let us provide the opposite of the gradient of $\mathcal{F}_2(\mathbf{H})$

$$\begin{aligned} -[\nabla_{\mathbf{H}} \mathcal{F}_2]_{ij} = (\alpha^2 \|[\mathbf{H}]_{\bullet j}\|_1^2 - \|[\mathbf{H}]_{\bullet j}\|_2^2) \\ ([\mathbf{H}]_{ij} - \alpha^2 \|[\mathbf{H}]_{\bullet j}\|_1) \end{aligned} \quad (30)$$

to be used in (19) with (21). In the next section, we shall test this algorithm for hyperspectral data unmixing.

VI. Application to hyperspectral data unmixing

The experiments presented in this paper were performed with 20 linear mixtures of 6 endmembers, the length of each spectrum being 826. The six endmembers used in this example were extracted from the ENVI library [9], and correspond to the construction concrete, green grass, micaceous loam, olive green paint, bare red brick and galvanized steel metal.

In order to characterize the performance of our approach, and show that it tends to provide sparse solutions, we considered a matrix $\mathbf{H}$ with only one nonzero entry per column. This entry was selected randomly and set to one. See Figure 1. Each observed spectrum was corrupted by an additive white Gaussian noise at a signal-to-noise ratio equal to 20 dB. The Frobenius norm

$$\mathcal{D}(\mathbf{V}, \mathbf{WH}) = \frac{1}{2} \sum_i \sum_j ([\mathbf{V}]_{ij} - [\mathbf{WH}]_{ij})^2 \quad (31)$$

[1] Using (19), note that $\|[\mathbf{H}]_{\bullet j}\|_1^2$ remains constant along iterations.

was used as criterion for the quality of the factorization. The following ratio was considered to stop the algorithm

$$\frac{\mathcal{D}(\mathbf{V},\mathbf{W}^k\mathbf{H}^k)-\mathcal{D}(\mathbf{V},\mathbf{W}^{k-1}\mathbf{H}^{k-1})}{\mathcal{D}(\mathbf{V},\mathbf{W}^{k-1}\mathbf{H}^{k-1})}\leq 10^{-10}. \qquad (32)$$

Regularization factor γ_1 was set to 0. See [7] for simulations dedicated to Tikhonov regularization. The matrices $\mathbf{H}$ obtained for $\gamma_2 = 0$ and $\gamma_2 = 10^{-3}$, respectively, are presented in Figures 2 and 3.

We clearly observe that the sparsity-enforcing function allowed us to recover, in most cases, the endmember involved in each observed spectrum. On the contrary, when no sparsity penalty term was used, all the entries of the estimated matrix $\mathbf{H}$ were nonzero. Finally, we checked that normalization of the columns of the matrix $\mathbf{W}$, as well as the flux conservation between $\mathbf{V}$ and $\mathbf{H}$, were satisfied at each iteration in both cases.

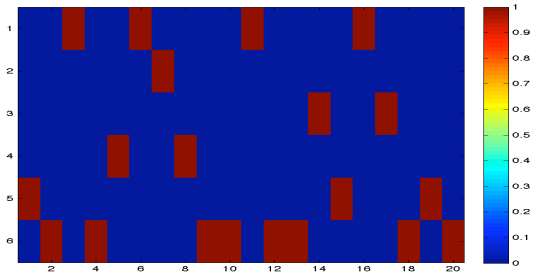

Fig. 1. True $\mathbf{H}$

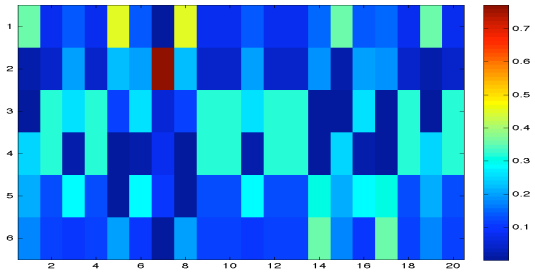

Fig. 2. Estimated $\mathbf{H}$ without sparsity constraint, $\gamma_2 = 0$

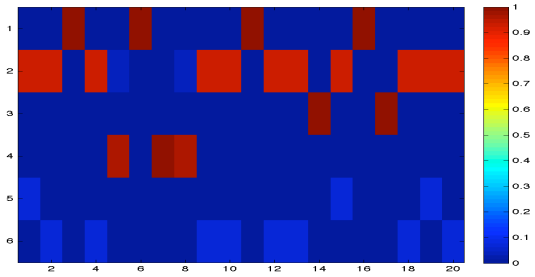

Fig. 3. Estimated $\mathbf{H}$ with sparsity constraint, $\gamma_2 = 10^{-3}$

VII. Conclusion

In this paper, we proposed a (split) gradient-descent method to solve the nonnegative matrix factorization problem subject to flux conservation constraints on each column of the estimated matrices. Tikhonov smoothness and sparsity-enforcing regularization terms were also considered. We illustrated our approach with an application within the context of hyperspectral data unmixing.

References

[1] C. I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York: Plenum Publishing Co., 2003.

[2] D. A. Landgrebe, *Signal Theory Methods in Multispectral Remote Sensing*. New York: Wiley, 2003.

[3] H. Lantéri, C. Theys, C. Richard, and C. Févotte, "Split gradient method for nonnegative matrix factorization," in *EUSIPCO, Aalborg*, 2010.

[4] H. Lantéri, M. Roche, and C. Aime, "Penalized maximum likelihood image restoration with positivity constraints- multiplicative algorithms," *Inverse problems*, vol. 18, pp. 1397–1419, 2002.

[5] H. Lantéri, M. Roche, O. Cuevas, and C. Aime, "A general method to devise maximum likelihood signal restoration multiplicative algorithms with non-negativity constraints," *Signal Processing*, vol. 54, no. 81, pp. 945–974, 2001.

[6] L. Armijo, "Minimization of functions having continuous derivatives," *Pacific Journal of Mathematics*, no. 16, pp. 1–3, 1966.

[7] H. Lantéri, C. Theys, and C. Richard, "Regularized split gradient method for non negative matrix factorization," in *ICASSP, Prague*, 2011.

[8] P. O. Hoyer, "Nonnegative matrix factorization with sparseness constraint," *Journal Machine Learning*, vol. 5, pp. 1457–1469, 2004.

[9] RSI (Research Systems Inc.), *ENVI User's guide Version 4.0*, Boulder, CO 80301 USA, Sep. 2003.