

SPECTRAL UNMIXING VIA MINIMUM VOLUME SIMPLICES: APPLICATION TO NEAR INFRARED SPECTRA OF COUNTERFEIT TABLETS

Marta B. Lopes^{†,‡}, José M. Bioucas-Dias[‡], Mário A. T. Figueiredo[‡]

Jean-Claude Wolff

[†]Institute for Biotechnology and Bioengineering

[‡]Instituto de Telecomunicações

Instituto Superior Técnico, Lisboa, Portugal

GlaxoSmithKline

Medicines Research Centre

Stevenage, UK

ABSTRACT

Counterfeit pharmaceutical products pose a serious public health problem. It is thus important not only to detect them, but also to identify their composition and assess the risk for the patient. Identifying the spectral signatures of the pure compounds present in a (maybe counterfeit) tablet of unknown origin is clearly a hyperspectral unmixing problem. In fact, under a linear mixing model, the hyperspectral vectors belong to a simplex whose vertices are the spectra of the pure compounds in the mixture. *Minimum volume simplex analysis* (MVSA) and *minimum-volume enclosing simplex* (MVES) are recently proposed algorithms, exploiting the idea of finding a simplex of minimum volume fitting the observed data. This work gives evidence of the usefulness of MVES and MVSA for unmixing near infrared (NIR) hyperspectral data of tablets of unknown composition. Experiments reported in this paper show that MVES and MVSA strongly outperform the state-of-the-art method in analytical chemistry for spectral unmixing: *multivariate curve resolution - alternating least squares* (MCR-ALS). These experiments are based on synthetic data (studying the effect of noise and of the presence/absence of pure pixels) and on a real dataset composed of NIR hyperspectral images of counterfeit tablets.

Index Terms— Hyperspectral unmixing, source separation, minimum volume simplex, alternating least squares, counterfeit tablets, near infrared imaging.

1. INTRODUCTION

The pharmaceutical industry has been devoting considerable attention to the problem of detecting counterfeit drugs and determining their composition, given the recent spread of counterfeiting operations. The absence of the active ingredient and/or the introduction of other, non-genuine, materials may reveal very harmful to the patients. Thus, the determination and quantification of the compounds present in these drugs is of the utmost importance for determining their negative impact. Moreover, the identification of compounds with a known narrow geographical distribution may help authorities find and shut down counterfeiting operations. Developing fast, non destructive, and effective methods for the identification of the composition of counterfeit drugs constitutes thus an active research front.

Near infrared (NIR) hyperspectral imaging has been recently used for the chemical study of counterfeit drugs [1, 2, 3]. NIR reflectance hyperspectral images of genuine and counterfeit tablets re-

veal clear spatial and spectral differences. However, when it comes to identifying the composition, with no prior knowledge about the samples, difficulties arise and spectral unmixing holds the promise of playing an important role. This promise is justified by the fact that a spectral unmixing method allows estimating the number of reference materials present, their spectral signatures, and the corresponding abundance fractions, as formally stated in the next paragraph.

Let $M = [m_1, \dots, m_p]$ be an $l \times p$ matrix, where each column $m_i \in \mathbb{R}_+^l$, for $i = 1, \dots, p$, represents the spectral signature (with l bands) of each of the p pure materials (the so-called endmembers). Under a noiseless linear mixing hypothesis, each observed spectrum $y \in \mathbb{R}_+^l$ is modeled as $y = M\alpha$, where $\alpha = [\alpha_1, \dots, \alpha_p]^T$ is the vector of abundance fractions of each material, which necessarily belongs to the canonical simplex $\Delta_p = \{x \in \mathbb{R}_+^p : 1_p^T x = 1\}$ (where $1_p = [1, 1, \dots, 1]^T$ denotes a vector of p ones, thus $1_p^T x = \sum_{i=1}^p x_i$). Given a set of n observed spectra (hyperspectral image pixels), collected in an $l \times n$ matrix Y , the noiseless linear mixing model for this data is thus

$$Y = MA, \quad (1)$$

where A is a $p \times n$ matrix containing the N abundance fraction vectors, thus satisfying

$$A \geq 0, \quad \text{and} \quad A^T 1_p = 1_n; \quad (2)$$

in (2), $A \geq 0$ means that all elements of A are non-negative [4]. Notice that, under this noiseless linear model, each column of Y (*i.e.*, the spectrum in each pixel) belongs to the simplex $S_M = \{y \in \mathbb{R}_+^l : y = M\alpha, \alpha \in \Delta_p\}$, the vertices of which correspond to the pure reference materials in the mixture. Spectral unmixing aims at estimating the number of pure materials p , their spectra M , and the corresponding abundance fractions A , from the observed data Y .

Multivariate curve resolution - alternating least squares (MCR-ALS) has been widely used for spectral unmixing in chemical applications [5]. The method works by alternatingly minimizing a least squares error criterion with respect to M and A , under the above explained positivity additivity constraints. The main disadvantage of MCR-ALS is the so-called rotational ambiguity problem, *i.e.*, a set of simplices with different orientations, all comprising the data points, are minimizers of the least squares criterion.

When there are pure pixels present in the data, *i.e.*, pixels containing only one of the pure materials in the mixture, the data contains the vertices of the simplex S_M . Many algorithms have been designed to exploit this scenario, aiming at finding the spectra of the vertices of the simplex: the *pixel purity index* [6], N-FINDR [7], *vertex component analysis* [4], *simplex growing algorithm* [8] and *automated morphological endmember extraction* [9].

M. B. Lopes gratefully acknowledges financial support from the Portuguese Foundation for Science and Technology (research grant no. SFRH/BD/31084/2006).

For datasets without pure pixels, recent algorithms have been proposed which follow the approach proposed in [10]. The unmixing problem is formulated as finding the vertices of the simplex of minimum volume containing the data. In order to yield good results, these algorithms require, however, that at least $p - 1$ spectral vectors stay in each of the facets of the simplex. The *minimum volume simplex analysis* (MVSA) method [11] works by fitting a minimum volume simplex to the data, while constraining the abundance fractions to be non-negative and sum to one. The resulting optimization problem is solved as a sequence of quadratically constrained sub-problems. In a final step, the hard constraint on the abundance fractions is replaced with a hinge type loss function to account for outliers and/or noise. The *minimum-volume enclosing simplex* (MVES) method [12] adopts a similar optimization problem, and addresses it by solving a sequence of linear programs.

In this work, we bring these recent hyperspectral unmixing approaches based on minimizing the simplex volume into chemical and pharmaceutical problems. In particular, we consider two simulated datasets and one real dataset composed of NIR hyperspectral images of five counterfeit tablets found in a market survey. The performance of the MVSA and MVES techniques is assessed in comparison with the MCR-ALS method (considered the state-of-the-art in analytical chemistry for source separation), showing that, for these datasets, the minimum-simplex-based methods clearly outperform MCR-ALS.

2. METHODS

2.1. Dimensionality Reduction

A standard first step used in hyperspectral unmixing consists in identifying the p -dimensional subspace spanned by the columns of matrices Y and M and project the data onto this subspace. Notice that if Y is indeed generated according to (1), the rank of Y is at most p . Assuming that this preprocessing step is always done, one may assume without loss of generality that that $l = p$, thus Y is now an $p \times n$ matrix and M is a square $p \times p$ matrix. Of course, in many practical problems, p is unknown and subspace identification methods have to be used [13]; in this paper, we assume that p is known.

2.2. MVSA

The implementation of the minimum volume concept is based on the fact that the volume of the simplex S_M (defined in Section 1) is a proportional to $|\det M|$. Assuming that M does have rank p (the pure spectral signatures are linearly independent), let $Q = M^{-1}$, and recall that $\det Q = 1/\det M$. Using Q , the constraints in (2) are written as $QY \geq 0$ and $Y^T Q^T 1_p = 1_n$. A minimum simplex volume criterion can then be written as the following constrained optimization problem:

$$\max_Q |\det Q|, \quad \text{subject to: } QY \geq 0 \quad \text{and} \quad Y^T Q^T 1_p = 1_n, \quad (3)$$

where, as above, $QY \geq 0$ denotes an element-wise inequality. It was shown in [11] that the constraint $Y^T Q^T 1_p = 1_n$ can be simplified to $Q^T 1_p = q$, where q is a suitably defined vector which can be computed from the data Y . Finally, the strict constraint $QY \geq 0$ is relaxed into a soft constraint, to accommodate for noise and/or outliers, leading to

$$\max_Q \log |\det Q| - \lambda 1_p^T \text{hinge}(QY) 1_n, \quad \text{subject to: } Q^T 1_p = q, \quad (4)$$

where λ is a regularization parameter and $\text{hinge}(\cdot)$ denotes the element-wise application of $\text{hinge}(x) = \max\{-x, 0\}$. The optimization problem (4) is solved via a sequential quadratic programming method, after being initialized by the VCA algorithm [4]; for more details, see [11].

2.3. MVES

The MVES approach is also based on the criterion (3), but does not consider the simplification of the constraint $Y^T Q^T 1_p = 1_n$ into $Q^T 1_p = q$, and uses a different optimization strategy. Notice that the constraint simplification is important, since $Y^T Q^T 1_p = 1_n$ is in fact a set of n scalar constraints, while $Q^T 1_p = q$ is a set of p scalar constraints, and usually $n \gg p$ (e.g., in the experiments reported below, $p = 4$ and $n = 10^4$). Finally, the optimization problem is addressed using a cyclic method based on the cofactor expansion of the determinant, with each intermediate optimization problem being equivalent to a pair of linear programs; for more details, see [12].

2.4. MCR-ALS

MCR-ALS formulates the problem of estimating M and A as that of finding a bilinear factorization of Y (as expressed by (1)) in a mean squared error sense, under the constraints in (2) and also positivity constraints on the elements of M . As for MVSA and MVES, the preprocessing step explained in Subsection 2.1 is applied to the data.

The method is initialized with either an estimate of M or an estimate of A and then cyclically updates these estimates using a minimum mean squared error criterion under the constraints explained in the previous paragraph, until a convergence criterion is satisfied. Specifically, this convergence criterion stops the algorithm when the change in the relative root mean squared error falls below some threshold, i.e., when

$$\frac{1}{\|Y\|_F} \left| \|Y - \widehat{M}^{(t)} \widehat{A}^{(t)}\|_F - \|Y - \widehat{M}^{(t-1)} \widehat{A}^{(t-1)}\|_F \right| < \delta,$$

where $\|B\|_F = (\sum_{i,j} B_{ij}^2)^{1/2}$ denotes the Frobenius matrix norm, and $\widehat{M}^{(t)}$ and $\widehat{A}^{(t)}$ are the estimates of M and A at iteration t of the algorithm. In the experiments reported below, we use the MCR-ALS implementation available at <http://www.ub.edu/mcr/>.

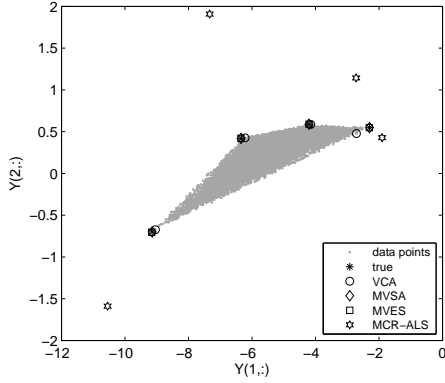
3. DATASETS

3.1. Simulated Dataset I

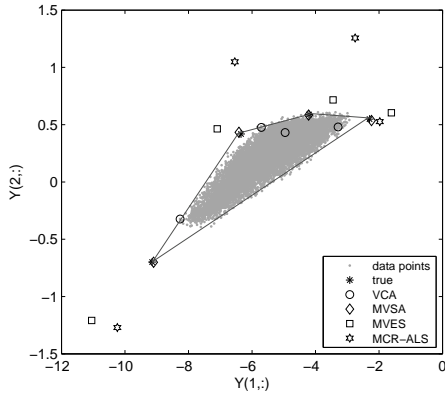
The first simulated dataset has size $n = 10000$ spectra/pixels and $p = 4$ pure spectra. The data is generated according to (1). The abundance fractions are Dirichlet distributed with unit parameter (i.e., uniformly distributed over the canonical simplex). The spectral signatures (with $l = 126$ spectral bands) of the endmembers are NIR spectra of four pure pharmaceutical substances: lamivudine, microcrystalline cellulose (MCC), rice starch, and talc; for more details about acquisition of the spectral signatures see [1, 14].

3.2. Simulated Dataset II

This second dataset aims at assessing the performance of the methods with noisy data and the absence of pure pixels. This dataset was generated from Dataset I as follows. First, all pixels with any abundance fractions above 0.7 were discarded; let Y denote the resulting data matrix. Second, a zero-mean Gaussian noise was added to the spectral vectors, such that $\text{SNR} \equiv 20 \log \|Y\|_F / \|N\|_F = 20$ dB (where N denotes the noise matrix).



(a)



(b)

Fig. 1: Estimated endmembers of simulated datasets (a) Simulated Dataset I (noise free, pure pixels present); (b) Simulated Dataset II (noisy and with pixels close to pure excluded).

3.3. Real Dataset

This dataset contains hyperspectral NIR images of five counterfeit Heptodin™ tablets. The data matrix Y contains $n = 5 \times 10^5$ pixels/spectra, with $l = 126$ bands. Given previous knowledge about composition of these tablets [1, 14], the above set of pure tablets was used to assess the accuracy of the estimates of the spectral signature.

4. RESULTS AND DISCUSSION

The three datasets were preprocessed as described in Subsection 2.1, taking $p = 4$. The regularization parameter λ was set to $100/n$. The endmembers estimated by MVSA, MVES and MCR-ALS from the Simulated Dataset I are represented Fig. 1(a) (the 2D representation uses the first two components of the 4D space obtained by the preprocessing step). Notice the very accurate MVSA and MVES estimates, closely followed by VCA; given the presence of pure pixels/spectra in the data, VCA performs well. For the Simulated Dataset II (which is noisy and has no pure spectra), MVSA still yields very accurate estimates, as shown in Fig. 1(b); in this case, MVES shows its higher sensitivity to noise, although it still

Table 1: Quantitative results for the three methods with the three datasets, as described in the text.

Method	$\ \widehat{M} - M\ _F$		
	Simulated data I	Simulated data II	Real data
MVSA	0.00	0.15	3.96
MVES	0.02	2.41	4.10
MCR-ALS	5.22	5.24	5.74

leads to reasonable estimates, while MCR-ALS provided again the worst estimates.

After an estimate \widehat{M} of M is obtained, if we are in possession of the true M and wish to compute some error measure (e.g., $\|\widehat{M} - M\|_F$), it is necessary to establish the best correspondence between the columns of M and those of \widehat{M} . This is done simply by trying all $p!$ permutations of the columns of \widehat{M} and choosing the one yielding the lowest values of the adopted error measure. Table 1 reports quantitative results, in terms of $\|\widehat{M} - M\|_F$. These results show a dramatic superiority of MVSA and MVES over MCR-ALS, for the noiseless data, and a clear superiority of MVSA over both MVES and MCR-ALS, in the synthetic noisy data case. For the real dataset, again MVSA and MVES clearly outperform MCR-ALS.

Fig. 2 depicts the estimates of the four endmembers of the real dataset produced by the three methods, alongside with the true pure spectra (for the best correspondence, as explained in the previous paragraph). The closeness between MVSA and MVES spectra is due to the very low noise in the real dataset. The MCR-ALS estimates are clearly worse, particularly for the pure spectra of lamivudine, MCC and starch.

Finally, regarding computational time, while MVSA and MVES run in just a few seconds, MCR-ALS took several minutes until convergence was achieved.

5. CONCLUSIONS

We have shown the advantages of using minimum-volume-based unmixing techniques (MVSA and MVES), over the traditional alternating least squares (MCR-ALS) approach, for the analysis of hyperspectral NIR data of counterfeit tablets. Both MVSA and MVES produced very good pure spectra estimates, from both simulated and real datasets, with MVSA proving to be the best method for noisy data with no pure pixels. Moreover, MVSA and MVES are computationally much faster, when compared to MCR-ALS approach.

Finding new, faster, and effective methodologies for spectral unmixing is crucial, particularly in applications with a high impact in public health. Combined with NIR imaging, a fast and non-destructive technique, spectral unmixing allows, in a few seconds, to obtain estimates of the compounds spectral signatures. Whenever a counterfeit drug appears, it is urgent not only to detect it and remove it from the market, but also to identify its composition, thus assessing the level of risk associated to it. The methods described in this paper may be contributing to the computational arsenal in the war against counterfeit drugs.

ACKNOWLEDGEMENT

We thank Tsung-Han Chan, Chong-Yung Chi, Yu-Min Huang, and Wing-Kin Ma, for providing us the MVES matlab code.

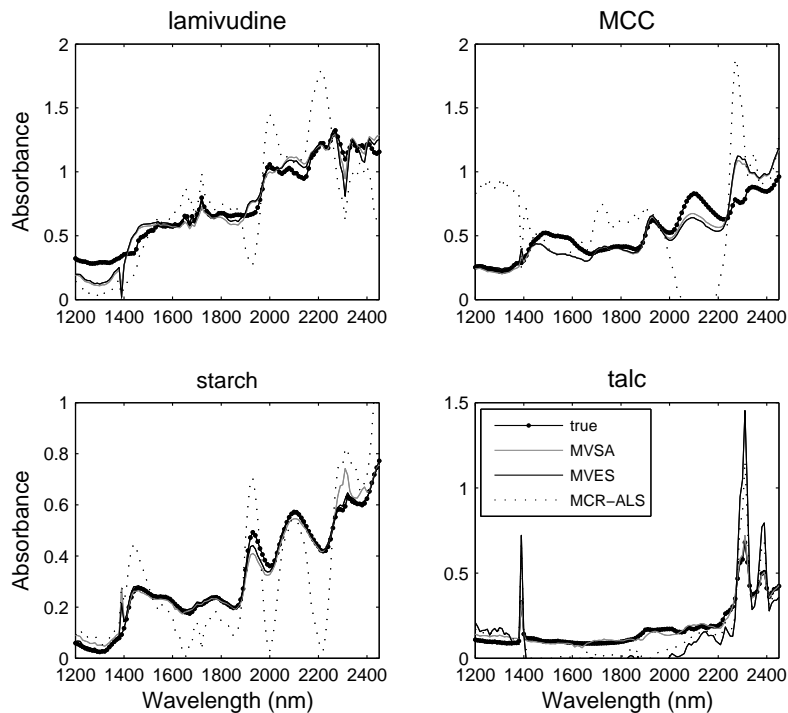


Fig. 2: Estimated signatures of real dataset (counterfeit tablets) by MVSA, MVES and MCR-ALS.

6. REFERENCES

- [1] M. B. Lopes and J.-C. Wolff, "Investigation into classification/sourcing of suspect counterfeit Heptodin™ tablets by near infrared chemical imaging," *Analytica Chimica Acta*, vol. 633, pp. 149–155, 2009.
- [2] J. Dubois, J.-C. Wolff, J. K. Warrack, J. Schoppelrei, and E. N. Lewis, "NIR chemical imaging for counterfeit pharmaceutical products analysis," *Spectroscopy*, vol. 22, pp. 36–41, 2007.
- [3] J.-C. Wolff, J. K. Warrack, L. Lidder, and E. N. Lewis, "NIR-based chemical imaging as an anticounterfeiting tool," *Pharmaceutical Manufacturing*, vol. 7, 2008.
- [4] J. M. P. Nascimento and J. M. Bioucas-Dias, "Vertex component analysis: a fast algorithm to unmix hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 898–910, 2005.
- [5] R. Tauler, A. Smilde, and B. Kowalski, "Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution," *Journal of Chemometrics*, vol. 32, pp. 31–58, 1995.
- [6] J. Boardman, "Automating spectral unmixing of aviris data using convex geometry concepts," In *4th Annual JPL Airborne Geoscience Workshop*, vol. 1, JPL Pub. 93–26, pp. 11–14, 1993.
- [7] M. E. Winter, "N-FINDR: an algorithm for fast autonomous spectral endmember determination in hyperspectral data," In *SPIE Conference Imaging Spectrometry V*, vol. 3753, pp. 266–275, 1999.
- [8] C.-I. Chang, C.-C. Wu, W. Liu, and Y.-C. Ouyang, "A new growing method for simplex-based endmember extraction algorithm," In *IEEE International Geoscience and Remote Sensing Symposium*, vol. 44, pp. 2804–2819, 2006.
- [9] A. Plaza, P. Martinez, and R. Perez, "Spatial/spectral endmember extraction by multidimensional morphological operations," In *IEEE International Geoscience and Remote Sensing Symposium*, vol. 40, pp. 2025–2041, 2002.
- [10] M. D. Craig, "Minimum-volume transforms for remotely sensed data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, pp. 542–552, 1994.
- [11] J. Li and J. M. Bioucas-Dias, "Minimum volume simplex analysis: a fast algorithm to unmix hyperspectral data," In *IEEE International Geoscience and Remote Sensing Symposium*, Boston, EUA, 2008.
- [12] T.-H. Chan, C.-Y. Chi, Y.-M. Huang, and W.-K. Ma, "A convex analysis based minimum-volume enclosing simplex algorithm for hyperspectral unmixing," In *IEEE International Conference in Acoustics, Speech and Signal Processing*, Taiwan, 2009.
- [13] J. M. Bioucas-Dias and J. M. P. Nascimento, "Hyperspectral subspace identification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 2435–2445, 2008.
- [14] M. B. Lopes, J.-C. Wolff, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "Determination of the composition of counterfeit Heptodin™ tablets by near infrared chemical imaging and classical least squares estimation," *Analytica Chimica Acta*, vol. 641, pp. 46–51, 2009.