Algorithm validation using multicolor phantoms

Daniel V. Samarov[1,*], **Matthew L. Clarke**[2,3], **Ji Youn Lee**[2], **David W. Allen**[2], **Maritoni Litorja**[2], **and Jeeseong Hwang**[2]

[1] *National Institute of Standards and Technology, Information Technology Laboratory, Statistical Engineering Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA*
[2] *National Institute of Standards and Technology, Radiation and Molecular Physics Division, 100 Bureau Drive, Gaithersburg, MD 20899, USA*
[3] *Currently with the National Gallery of Art, Washington, DC, 20565, USA*

[*] *daniel.samarov@nist.gov*

Abstract: We present a framework for hyperspectral image (HSI) analysis validation, specifically abundance fraction estimation based on HSI measurements of water soluble dye mixtures printed on microarray chips. In our work we focus on the performance of two algorithms, the Least Absolute Shrinkage and Selection Operator (LASSO) and the Spatial LASSO (SPLASSO). The LASSO is a well known statistical method for simultaneously performing model estimation and variable selection. In the context of estimating abundance fractions in a HSI scene, the "sparse" representations provided by the LASSO are appropriate as not every pixel will be expected to contain every endmember. The SPLASSO is a novel approach we introduce here for HSI analysis which takes the framework of the LASSO algorithm a step further and incorporates the rich spatial information which is available in HSI to further improve the estimates of abundance. In our work here we introduce the dye mixture platform as a new benchmark data set for hyperspectral biomedical image processing and show our algorithm's improvement over the standard LASSO.

OCIS codes: Probability theory, stochastic processes, and statistics (000.5490); Multispectral and hyperspectral imaging (110.4234); Instrumentation, measurement, and metrology (120.0120); Medical optics and biotechnology (170.0170); Medical and biological imaging (170.3880); Microscopy (180.0180); Optical standards and testing (350.4800).

References and links

1. B. Sorg, B. Moeller, O. Donovan, Y. Cao, and M. Dewhirst, "Hyperspectral imaging of hemoglobin saturation in tumor microvasculature and tumor hypoxia development," J. Biomed. Opt. **10**, 044004 (2005).
2. M. Martin, M. Wabuyele, P. Chen, M. Panjehpour, M. Phan, B. Overholt, G. Cunningham, D. Wilson, R. DeNovo, and T. Vo-Dinh, "Development of an advanced hyperspectral imaging (hsi) system with applications for cancer detection," Ann. Biomed. Eng. **34**, 1061–1068 (2006).
3. K. Zuzak, R. Francis, E. Wehner, M. Litorja, J. Cadeddu, and E. Livingston, "Active dlp hyperspectral illumination: a noninvasive, in vivo, system characterization visualizing tissue oxygenation at near video rates," Anal. Chem. **83**, 7424–7430 (2011).
4. B. Pogue and M. Patterson, "Review of tissue simulating phantoms for optical spectroscopy, imaging and dosimetry," J. Biomed. Opt. **16**, 16272–16283 (2006).
5. M. Clarke, D. Allen, D. Samarov, and J. Hwang, "Characterization of hyperspectral imaging and analysis via microarray printing of dyes," Proc. SPIE. **7891**, 78910W (2011).
6. M. Clarke, J. Lee, D. Samarov, D. Allen, M. Litorja, and J. Hwang, "Designing microarray phantoms for hyperspectral imaging validation," Biomed. Opt. Express (to be published).

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

7. L. Nieman, M. Sinclair, J. Timlin, H. Jones, and D. Haaland, “Hyperspectral imaging system for quantitative identification and discrimination of fluorescent labels in the presence of autofluorescence,” in *3rd IEEE International Symposium on Biomedical Imaging: Nano to Macro, 2006* (2006), pp. 1288–1291.
8. R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. R. Stat. Soc. B **58**, 267–288 (1996).
9. D. Samarov, J. Hwang, J. Lee, and M. Clarke, “The spatial lasso with applications to unmixing hyperspectral images,” Tech. Rep., National Institute of Standards and Technology (2012).
10. F. Green, *The Sigma-Aldrich Handbook of Stains, Dyes and Indicators* (Aldrich Chem Co Library, 1990).
11. D. Samarov, M. Clarke, J. Lee, D. Allen, M. Litorja, and J. Hwang, “Validating the lasso algorithm by unmixing spectral signatures in multicolor phantoms,” Proc. SPIE **8229**, 82290Z (2012).
12. C.-I. Chang and Q. Du, “Estimation of the number of spectrally distinct signal sources in hyperspectral imagery,” IEEE Trans. Geosci. Remote Sens. **42**, 608–619 (2004).
13. J. Bioucas-Dias and J. Nascimento, “Hyperspectral subspace identification,” IEEE Trans. Geosci. Remote Sens. **46**, 2435–2445 (2008).
14. J. Nascimento and J. Dias, “Vertex component analysis: A fast algorithm to unmix hyperspectral data,” IEEE Trans. Geosci. Remote Sens. **43**, 898–910 (2005).
15. J. Bioucas-Dias, “A variable splitting augmented lagrangian approach to linear spectral unmixing,” in *First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS ’09* (2009), pp. 1–4.
16. L. Breiman, “Better subset rergression using the nonnegative garotte,” em Technometrics **37**, 373–384 (1995).
17. J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” J. Am. Stat. Assoc. **96**, 1348–1360 (2001).
18. M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, “Sparse unmixing of hyperspectral data,” IEEE Trans. Geosci. Remote Sens. **49**, 2014–2039 (2011).
19. J. Bioucas-Dias and A. Plaza, “Hyperspectral unmixing: geometrical, statistical, and sparse regression approaches,” Proc. SPIE **7830** 78300A (2010).
20. B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, “Least angle regression,” Ann. Stat. **32**, 407–499 (2004).
21. J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani, “Pathwise coordinate optimization,” Ann. Appl. Stat. **1**, 302–332 (2007).
22. A. Zymnis, S.-J. Kim, J. Skaf, M. Parente, and S. Boyd, “Hyperspectral image unmixing via alternating projected subgradients,” in *Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers, 2007. ACSSC 2007* (2007), pp. 1164–1168.
23. A. Zare, “Spatial-spectral unmixing using fuzzy local information,” in *2011 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* (IEEE, 2011), pp. 1139–1142.

1. Introduction

Hyperspectral imaging (HSI) is a technology commonly used in remote sensing but has recently found increased use in biomedicine, from investigations at the cellular [1] to the tissue level[2, 3] and often captured in reflectance mode. Measurement accuracy depends on radiometric proficiency and the robustness of the algorithm used to extract the desired quantity. Radiometric challenges are both instrument-related (e.g. drift, noise, sensor inhomogeneities) and scene-related (e.g. glare, topography, light field non-uniformity). These can be minimized by performing radiometric calibration using artifacts such as wavelength, intensity and reflectance standards and applying the appropriate corrections. The algorithm on the other hand, needs to accurately extract the analyte quantity in the presence of signal contaminants such as scattering and signal interferers arising from undesired absorption and spatial and spectral correlations. It requires validation against known preparations, closely resembling in properties to the real samples. In biomedicine, tissue phantoms are prepared for this purpose[4]. These ground-truth artifacts tend to have short shelf life and are often prepared with other functional considerations besides optical property. They may also be difficult to prepare reproducibly.

Recently a novel, custom-tailored microarray printing platform was developed[5, 6] to create a testbed for characterizing hyperspectral imaging systems and validating algorithm performance. This microarray printing system allows for precise sample composition through variation of dye concentrations and mixture proportions, and high spatial control. The printed sample spots are imaged using a hyperspectral imager. Details of the preparation of the microarray prints and image acquisition have been described previously[5]. Briefly, a spotting robot (Spot-

#164628 - $15.00 USD   Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

Bot2, ArrayIt, Sunnyvale, CA) is programmed to print prescribed mixtures of dyes dissolved in acqueous polyethylene glycol (PEG) on a microscope slide. The spots ($\sim 100\ \mu$m in diameter) are spaced 250 μm apart to ensure no signal crosstalk between spots.

The printed dye microarray is a semi-stable, scalable template of mixtures of known composition for testing algorithm accuracy and to some extent, instrumental limitations. Components and concentrations can be tailored to approximate optical properties of the specific species in the actual biomedical measurement. Spatial control allows us to determine the degree of spatial correlations given a specific optical detection geometry. The dye microarray printing platform is a method of preparing surrogate validation samples especially for systems which are not easy to prepare or handle. This method has been shown to be useful in preparing validation artifacts for fluorescence[7] and absorbance[5] microscopy of biological samples.

In the work presented here we focus on endmember (a signature belonging to a chemical species, or simply a radiometric artifact such as glare) *abundance fraction* (i.e. dye concentration) estimation. This benchmarking sample allows us to see how well an algorithm is capable of performing with various issues present. It also provides insight into some of the instrument and acquisition limitations such as the imager's limit of detection for various endmember mixture proportions and combinations. Similar to complex tissue phantoms, this sample also exhibits effects of scattering and edge diffraction despite the intent of creating a simple system of variable absorbance spectra.

In our work here we look to validate the well known LASSO (Least Absolute Shrinkage and Selection Operator [8]) and a novel method we call SPLASSO (Spatial LASSO [9]) against the microarray multicolor phantom. The LASSO and SPLASSO are both "sparse" regression methods (often referred to as basis pursuit in other fields), i.e. they have the property that some of the regression coefficients (i.e. abundances) are set exactly equal to 0. This property will be discussed in more detail in Section 2. Methods like these are particularly well suited for abundance estimation in hyperspectral imaging as they exploit the fact that most pixels in an image are only composed of a subset of the total number of endmembers present (i.e. some will have 0 abundance).

1.1. Sample Description

Fig. 1 shows the layout of one of these microarrays consisting of two different dyes. The two dyes are acid red 1 (AR) and new coccine (NC)[10]. The image on the left is what the array looks like and on the right is the layout of the locations, relative proportions and concentrations of each of the dyes. Dye samples are initially prepared in water. Dyes were dissolved in water to create stock solutions that will provide absorption spectra of similar magnitude (w/w: 4.9% AR; 7.0% NC). These stock solutions are then further diluted and mixed with 75% poly(ethylene glycol, MW = 600 D) (PEG) for a final PEG concentration of 50% (v/v). For ease of comparison, the highest concentration of each dye sample present in images is designated as 100% relative concentration. Here, the two right-most columns with "AR" and "NC" written above, correspond to individual dyes at 100% down to 1% relative concentration. The columns to the left of this show mixtures of these dyes at varying proportions and concentrations. All together, three replicate phantoms were generated using these two dyes and the layout described. We focus on the second and third replicates due to issues with acquisition of the first.

The two dyes we selected for this study were chosen because of their spectral distributions, shown in Fig. 2. Here relative absorbance measurements (y-axis) were taken at 61 wavelengths between 400 nm and 700 nm (x-axis) with 5 nm spacing. While the dyes are somewhat spectrally distinct there is considerable overlap between them. This overlap presents a considerable challenge when trying to determine the concentration and proportion of a particular dye at a given spatial location where two dyes have been mixed. While not presented here, multicolor

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

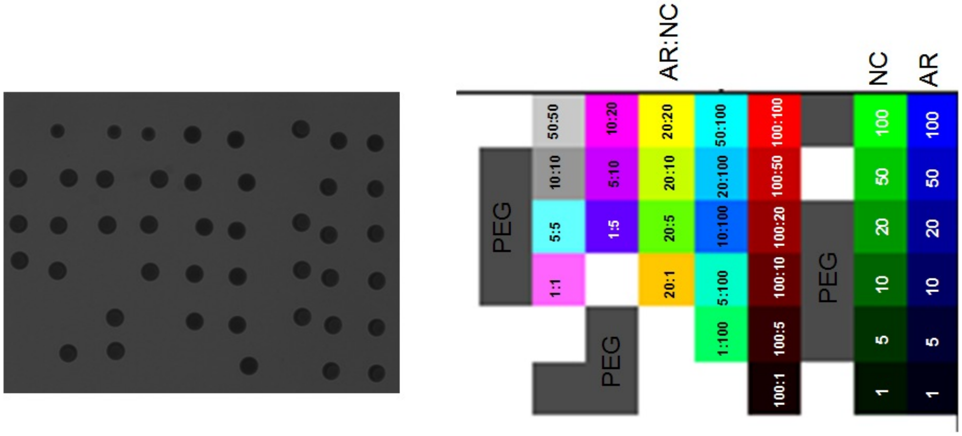

Fig. 1. The design of the microarray printing platform for two dyes. The image on the left depicts the actual array and the image on the right shows the location, concentrations and proportion of each of the dyes.

phantoms have been developed where the spectral signatures are more distinct, making estimation of the abundance fraction (somewhat) easier[11, 5, 6]. The two remaining signatures correspond to the PEG and background spectra.

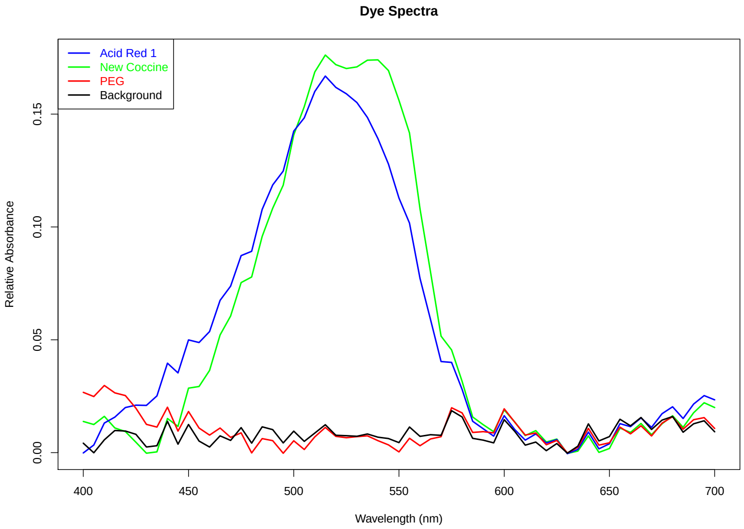

Fig. 2. The spectral signatures of the two dyes used in this experiment, as well as the PEG and background signatures.

The spectral signatures presented in Fig. 2 were estimated from the HSI measurements by taking the average of a 7×7 pixel region within the 100% dye spot locations for AR and NC. Similar averages were taken for the PEG spots and the background. Note, prior to obtaining these spectra, the data cubes were preprocessed using a baseline correction algorithm. These

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

four spectra AR, NC, PEG and background are used as the endmembers in our models.

It is important to note that in most cases simply taking the average of a region believed to contain an endmember, as we do here (where we happen to have an understanding of what the "ground truth" is) will not yield very reliable results. A widely studied problem in the HSI literature is estimating how many and which endmembers are present in an image. Some well known algorithms for determining the number of endmembers include virtual dimensionality (VD[12]) and hyperspectral signal identification by minimum error (HYSIME[13]). For estimating the endmembers themselves, two commonly used approaches are vertex component analysis (VCA[14]) and the simplex identification via split augmented Lagrangian (SISAL[15]) algorithms. While both these problems are extremely important, and the data presented here can certainly act as test bed for them. However, since our focus is on abundance estimation we take what we know to be "ground truth" for the endmembers and leave the exploration of these other problems to future work.

1.2. Results and discussion

Algorithm performance is measured by looking at how well the estimated dye concentrations (whose calculation we describe next) and corresponding ratios match up with the design in Fig. 1. Fig. 3 shows a heatmap of the estimated abundances using the LASSO and SPLASSO for each of the two dyes (we have zoomed in on the dye spots and removed most of the background regions to ease visualization of the results) for the second of the three replicates (results are similar for the third). The vertical white lines delineate the different dye mixtures shown at the bottom and the color bar to the right of each figure shows the range of the estimated abundance fractions. Note that in each case the fractions are less than 1. There are a number of factors at play here which could cause this to happen. In particular the complexity of the background and other features all have an effect on the measured absorbance and resulting estimate of abundance fractions.

From a qualitative standpoint the SPLASSO algorithm produces a more visually appealing result as compared to the standard LASSO (less salt-and-pepper artifacts). From a practical standpoint, this reduces the number of false-positive readings, i.e. saying a dye is present when it is not and reduces the overall variability in the measurements (as we will see in the following).

In order to estimate the concentration at each dye location we begin by estimating the 100% concentration values. This is done by calculating the average estimated abundance fractions from a 7×7 pixel region centered at the 100% concentration locations for AR and NC (see Fig. 1 and 3); we call these values AR_{100}, and NC_{100}. Similar 7×7 regions are then selected and averages calculated at each spot; we call these values AR_i, and NC_i, where $i \in \{1, \ldots, 45\}$ denotes the spot location shown in Fig. 4. The final estimated calibrated abundance fraction for each dye at each location are then calculated as

$$\frac{\mathrm{AR}_i}{\mathrm{AR}_{100}} \text{ and } \frac{\mathrm{NC}_i}{\mathrm{NC}_{100}}.$$

The results for the first replicate is shown in Fig. 5 (again, results for the second are similar). Each of the barcharts shows the absolute error of our estimate (which we refer to as calibrated abundance fraction or CAF) from ground truth along the x-axis. The locations of the barcharts throughout the figure correspond to the dye design shown in Fig. 1. The blue bars display the results of the SPLASSO and the green bars the LASSO with the bars on the top half of each plot showing the results for NC and the lower bars AR. The segments at the top of each bar represent the standard errors of the CAF in the 7×7 region used to find the estimates. The true concentration and mixing proportions can be found just above each of the plots. As an example,

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

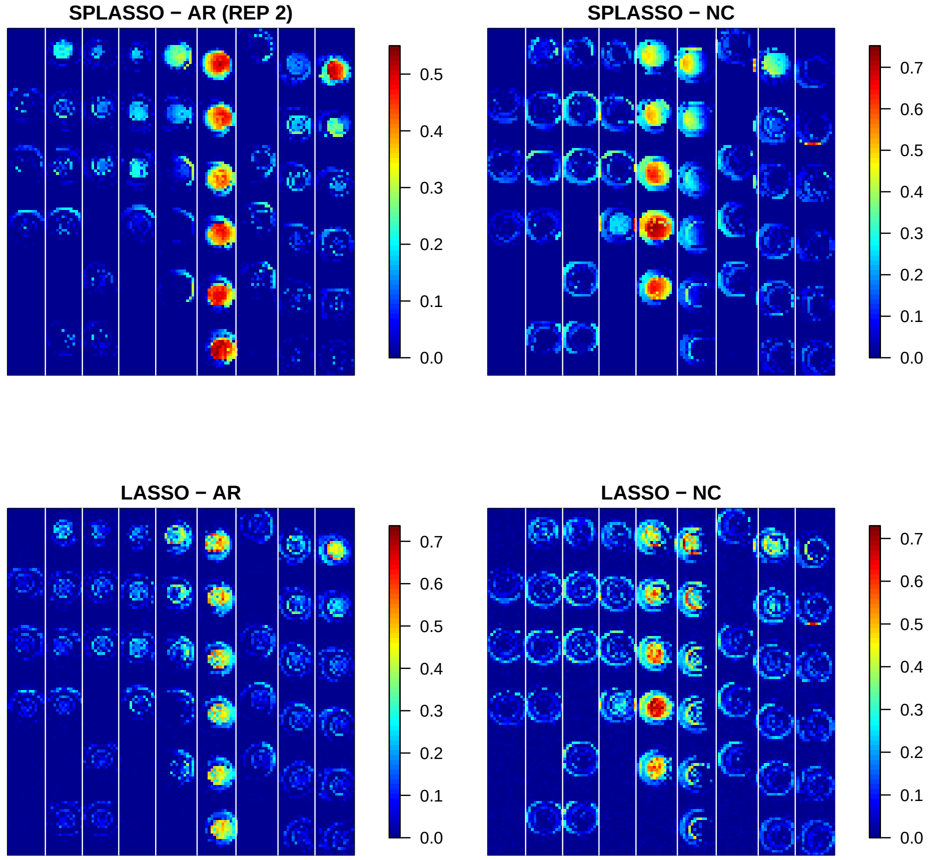

Fig. 3. The abundance estimates of AR and NC for the second replicate data set. The color scale on the right indicates the estimated abundance fractions.

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

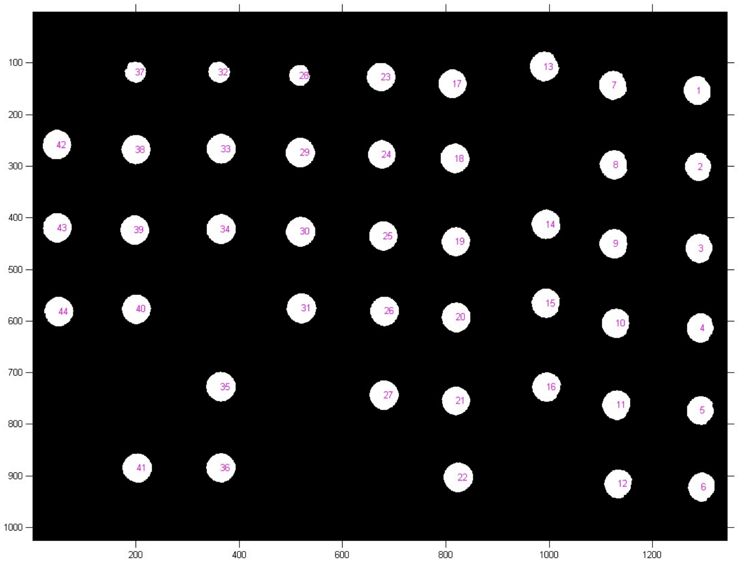

Fig. 4. Indices of the pure and mixed dye locations.

if we were to look at the top right barchart, which corresponds to the 100% AR and 0% NC dye spot location, we see that the absolute error for LASSO and SPLASSO for AR the CAF is 0% (a result of normalizing against the abundance estimate at this location) and is approximately 10% for the SPLASSO and 40% for the LASSO for the NC CAF. Similarly, the barchart just below that displays the absolute errors for the 50% AR and 0% NC spot and so on.

From these results we can see that both the LASSO and SPLASSO are able to generally capture the dilution curves of the pure AR (rightmost column) and NC (second from the right) dyes. In the locations where the two dyes have been mixed, the overall accuracy of the measurements decrease, however the relationships are still generally captured. On the whole the absolute error of the SPLASSO estimates as compared to ground truth are smaller and less variable than the corresponding LASSO estimates (with the exception of a few locations). Additionally the SPLASSO shows a higher overall sensitivity for estimation of the CAF in lower concentration mixtures. This improvement in performance is further reflected by the smaller sum of absolute errors across dye spot locations for the SPLASSO, 8.6 as compared to the the LASSO, 16. Of interest is that in the regions where one or both dyes are *not* present, the SPLASSO tends to produce estimates closer to 0% CAF. In biomedical imaging applications this is particularly important as avoiding false positives can be as important as identifying true positives.

Given the challenging nature of the two dye mixture data these results are quite promising and show that both the LASSO and SPLASSO are strong candidates for abundance fraction estimation in HSI. There are several possible reasons why the senstitivity of these algorithms was a bit low in certain parts of the image. To start, part of the issue may involve a need for improvement in the image acquisition process itself. However, further development of pre-processing steps (e.g. background correction) would almost certainly help improve results. Experimental errors may also play a role. For example, as the relative ratio of one dye increases, dye segregation due to different surface wetting may introduce local heterogeneity within a single microarray

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

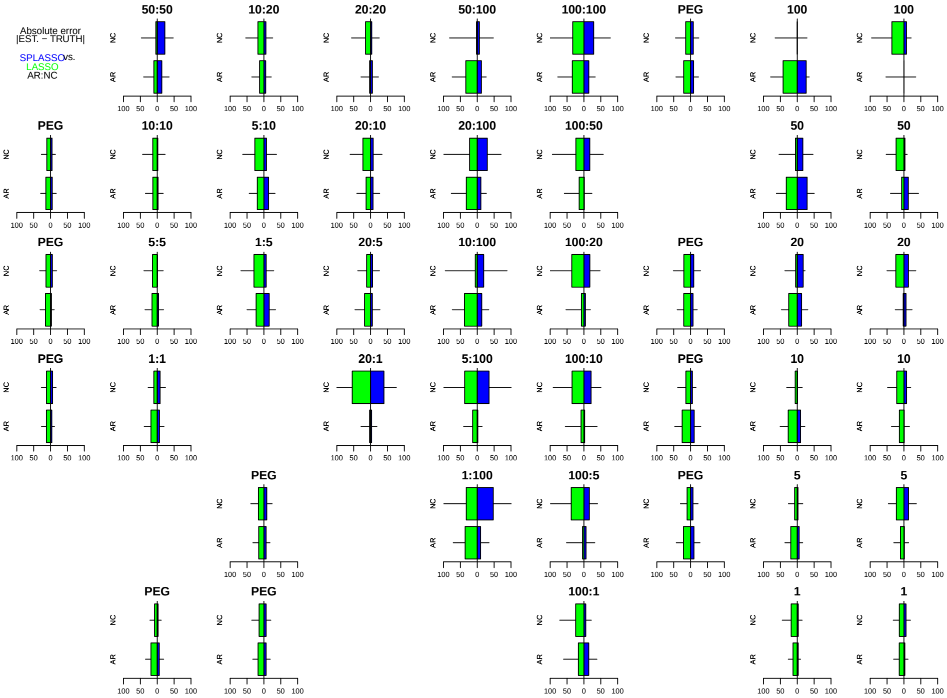

Fig. 5. Results from CAF estimation using the SPLASSO and LASSO at each dye location. In every subplot the corresponding barplots show the absolute errors (x-axis) of the CAF estimate from ground truth and their corresponding standard errors (vertical line) for the AR and NC dyes.

spot resulting in differential reflective scattering properties. It is important to emphasize that these issues will also be present in biomedical samples; having the ability to generally assess how an algorithm might perform under these circumstances is critical.

For this particular benchmark data set, we conclude that the SPLASSO algorithm is capable of extracting endmembers of two different chemical substances within the ratio range of 100:50 to 100:100. This also reflects the HSI camera's general level of sensitivity for measuring mixture proportions. For the sake of space we have not included results for the remaining replicate measurements; however similar results hold.

2. Methods

While not always the case in practice, it is commonly assumed (as it is here) that at each pixel the spectral signature is a linear mixture of each of the endmembers present in the scene. Before providing a more formal description of linear mixing we begin by introducing some notation: define $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})^T$, $i = 1, \ldots, n$ to be the set of spectral response vectors, n corresonding to the total number of pixels in the image. Let $\mathbf{x}_j = (x_{1j}, \ldots, x_{pj})^T$, $j = 1, \ldots, m$ be the set endmembers (where each of the p entries maps to a specific wavelength), which are collected in the matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$. Finally let $\beta_i = (\beta_{i1}, \ldots, \beta_{im})$, $i = 1, \ldots, n$ be the set of abundance vectors whose entries tell us the proportion and concentration of an endmember at a pixel. In order to ensure that these abundances have a physical meaning it is typically required that each element of β_i be nonnegative and that the sum of the elements of β_i are less than or equal to one. More generally

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

$$\mathbf{y}_i = \mathbf{X}\beta_i, \text{ subject to } \beta_{il} \geq 0 \text{ and } \sum_{l=1}^{m} \beta_{il} \leq 1. \tag{1}$$

Note, it is more commonly assumed in the HSI literature that $\sum_i \beta_{il} = 1$. This constraint reflects the assumption that we have captured all or most of the relevant endmembers; however in the majority of cases this will not be true. One of the major issues with forcing the β_{il}'s to sum to one is that it has the potential to introduce noise artifacts into the estimates; for example, if an endmember actually present in the image has not been included, a sum-to-one constraint may artificially inflate the estimate abundance of another endmember in order to compensate. Allowing for an inequality as in (1) will help avoid such situations and will generally be more robust to noise.

As mentioned in the introduction "sparsity" in the abundance vectors β_i (i.e. possibly many β_{ij}'s being equal to 0) arises naturally in hyperspectral imaging as most pixels are typically composed of only a subset of the m endmembers. For example, in the the dye mixture data we know that many of the spots are made up of one or a mixture of two dyes. In some applications large dictionaries of endmembers specific to the types of objects being analyzed are available, with only a subset of the endmembers in the dictionary being present in the image at all. By explicitly taking into account the sparse nature of the endmember abundance vectors we are able to reduce the number of false positives (saying an endmember is present in a pixel when it is not) and therefore the accuracy of the estimation.

In Sections 2.1 and 2.2 we outline the LASSO and SPLASSO models respectively. Both of these approaches produce sparse estimates of the abundances and as illustrated in Section 1 produce very good results.

2.1. LASSO

Standard approaches to model building, such as ordinary least squares (OLS) do not produce sparse results and variable selection procedures traditionally used in conjunction with OLS, such as best subset selection, encounter difficulties when there are more than a few variables (as the number of possible combinations to consider quickly becomes intractable). Other shortcomings of subset selection methods are related to the discrete nature in which variables are added or removed from the model [16], [17].

In order to effectively deal with these challenges regularization techniques which incorporate an l_1 penalty on the coefficient (abundance) vector, such as the LASSO [8] were developed. Through the introduction of the penalty term these methods are able to simultaneously perform prediction and variable selection. Sparse regression methods have been shown to be effective in practice across a wide range of applications and the LASSO and related methods have seen extensive use in the HSI literature[18], [19]. The form of the LASSO is quite similar to the linear mixing model described in (1) with the additional constraint that $|\beta|_1 = \sum_{j=1}^{m} \beta_j \leq c$, for some constant c (where $|\cdot|_1$ is the l_1 norm). The loss function can then be expressed as

$$\hat{\beta}_i(\text{LASSO}) = \arg\min_{\beta_i} \left\| \mathbf{y}_i - \sum_{j=1}^{m} \mathbf{x}_j \beta_{ij} \right\|^2 + \lambda |\beta_i|_1, \tag{2}$$

where λ is a nonnegative regularization parameter. The l_1 penalty term has the effect of continuously shrinking the coefficients toward 0 as λ increases and, for λ sufficiently large it can be shown that some coefficients are set exactly to 0. Extending this to the abundance estimation problem requires that the above estimation procedure be repeated for each i, $i = 1, \ldots, n$.

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

To gain some insight into how the LASSO is able to obtain estimates which are exactly 0, we consider the following special case: suppose that the matrix of endmembers, $\mathbf{X}$ is orthonormal, i.e. $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ and $\mathbf{I}$ is the identity matrix. Then it can be shown that the solution of the LASSO problem in (2) has the closed form solution

$$\hat{\beta}_{il}(\text{LASSO}) = \text{sgn}(\hat{\beta}_{il}(\text{OLS}))(|\hat{\beta}_{il}(\text{OLS})| - \lambda/2)_+, \; l = 1, \ldots, m \qquad (3)$$

where $\hat{\beta}_{il}(\text{OLS}) = \mathbf{x}_l^T\mathbf{y}_i$ is the OLS estimate, and $(u)_+ = \max(0,u)$. Thus for $\lambda/2 \geq |\hat{\beta}_{il}(\text{OLS})|$, $\hat{\beta}_{il}(\text{LASSO}) = 0$.

For the more general case where we do not have orthonormality two fast and efficient algorithms have been proposed in the literature to solve (2): one based on least angle regression (LARS) [20] and a more recent adaptation based on the coordinate descent algorithm [21]. Both these methods provide a significant improvement in computational speed over standard linearly constrained, quadratic programming approaches; in particular coordinate descent[21] has been shown to be very efficient for working with larger data sets. Straightforward modifications of either approach allow us to incorporate the positivity and sum to less than or equal to one constraints of linear mixing.

2.2. *SPLASSO*

One of the drawbacks of existing LASSO methods in the context of hyperspectral imaging is that they ignore the smoothly varying, spatial relationships between pixels and abundances. While numerous methods exist which capture spatial information in the unmixing and abundance estimation process[22], [18], [23], there are none which combines both the spatial and sparse behavior found in HSI. In order to effectively leverage this we introduce a spatial penalty term of the form $\sum_{j \in N(\mathbf{y}_i)} ||\beta_i - \beta_j||^2 w_{ij}$ into the LASSO objective (2) giving us the SPLASSO loss function

$$\hat{\beta}_j(\text{SPLASSO}) = \arg\min_{\beta_j} \sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{X}\beta_i||^2 + \lambda_1|\beta_i|_1 + \lambda_2 \sum_{j \in N(\mathbf{y}_i)} ||\beta_i - \beta_j||^2 w_{ij}. \qquad (4)$$

Here λ_1 and λ_2 are nonnegative regularization parameters, $N(\mathbf{y}_i)$ is the set of neighboring pixels about $\mathbf{y}_i$ and $w_{ij} \in [0,1]$ is a spatial weight function capturing the similarity between observation i and its neighbors $j \in N(\mathbf{y}_i)$. The neighborhood defined by $N(\cdot)$ can take on a number of different forms; for our purposes we take $N = N_k$, the symmetric k-neighborhood on a regular 2D grid. To illustrate the form of N_k, suppose we are at grid point g_{rs} in a $M_1 \times M_2$ image, $1 \leq r \leq M_1$, $1 \leq s \leq M_2$. For $k = 1$ our neighborhood would be defined as the set of points $N_1 = \{g_{r-1,s}, g_{r+1,s}, g_{r,s-1}, g_{r,s+1}, g_{r-1,s+1}, g_{r+1,s+1}, g_{r-1,s-1}, g_{r+1,s-1}\}$.

The introduction of the penalty term $\sum_{j \in N_k(\mathbf{y}_i)} ||\beta_i - \beta_j||^2 w_{ij}$ in (4) has the effect of "encouraging" the β_i's to be similar to their k-neighbors, introducing a smoothness to the coefficient vectors. In hyperspectral unmixing this has several appealing aspects: in particular it allows our estimates to be more robust to instrument and sample variability. Intuitively this makes sense as the variability introduced from these different sources will tend to be smoothed out. Of course, as in any smoothing method, care needs to be taken to avoid removing actual features by oversmoothing.

For this reason appropriate selection of the weights w_{ij} and regularization parameters λ_1 and λ_2 are extremely important. In the application of the SPLASSO to hyperspectral imaging it is desirable to have a weight function which uses both spatial and spectral information. Let us suppose that the spectral signature, $\mathbf{y}_i$ whose abundances we are estimating corresponds to

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012

the rs^{th} pixel in the image (for illustrative purpose we refer to this point as $\mathbf{y}_{rs}$). The spatial component of the weight function can then be captured by

$$b_{rs}(lm) = \begin{cases} \frac{1}{(r-l)^2+(s-m)^2}, l \in [r-k, r+k], m \in [s-k, s+k] & \text{if } (l,m) \notin (r,s), \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Our decision to use (5) is because it provides a decrease in the effect a neighboring pixel has the further we move out from the current observation being estimated. However, the decrease is not so rapid as to make the contribution of the surrounding observations negligible. Next, to leverage spectral information we use the weights

$$c_{rs}(lm) = \frac{\mathbf{y}_{rs}^T \mathbf{y}_{lm}}{||\mathbf{y}_{rs}||||\mathbf{y}_{lm}||}, l \in [r-k, r+k], m \in [s-k, s+k], \tag{6}$$

which is the cosine of the angle between the spectra. This is a similarity measure commonly used in hyperspectral image analysis applications. For our purposes it is appealing because it allows our spatial weight function to be adaptive to local features in the image, e.g. if we are at the edge of an object. We have also found it useful in practice to include a threshold on the angle between spectra, so that if $\text{acos}(c_{rs}(lm)) > t$, $t \in [0, \pi]$ then $c_{rs}(lm) = 0$. Putting the spatial (5) and spectral (6) weights together the weight function is defined as

$$w_{rs}(lm) = b_{rs}(lm)c_{rs}(lm).$$

To gain insight into the properties of (4) and the role of the regularization parameter λ_2 we once again considering the case where $\mathbf{X}$ is taken to be orthonormal. Let $\gamma = 1/(1+\lambda_2)$, $\sum_{j \in N_k(\mathbf{y}_i)} w_{ij} = 1$, (note, the latter does not need to hold in general, we do so here for illustrative purposes), $\alpha_{i,l} = \sum_{j \in N_k(\mathbf{y}_i)} \beta_{j,l} w_{ij}$ and

$$\hat{b}_{i,l} = \gamma \hat{\beta}_{i,l}(\text{OLS}) + (1-\gamma)\alpha_{i,l}$$

then it can be shown that

$$\hat{\beta}_{i,l}(\text{SPLASSO}) = \text{sgn}(\hat{b}_{i,l}) \left(|\hat{b}_{i,l}| - \frac{\lambda_1}{2}\gamma \right)_+. \tag{7}$$

Looking at (7) we can see that it is quite similar to (3) except that now it the parameter γ controls the tradeoff between the OLS estimate and a smoothly weighted average of its neighboring pixels.

Similar approaches to solving the LASSO can also be applied to solving the SPLASSO; for details see [9].

3. Conclusions

As HSI gains momentum in the biological and medical fields, and in particular as it begins to see a transition from benchtop to bedside, it is critical that the algorithms being used to analyze the resulting measurements be properly validated. To address these issues a collection of dye mixture phantoms have been developed which we have shown here can act as an example of a test bed for algorithm validation. In the context of hyperspectral medical imaging, as the resulting estimated abundances will be used to make important decisions, accurate quantitative identification of biochemical entities is critical. Here we have shown that the LASSO and in particular the SPLASSO, are two candidates well suited to such a task.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this manuscript are to foster understanding. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Acknowledgment

The authors would like to thank Drs. Antonio Possolo, Eric Shirley, and Kimberly Briggman for useful discussions and support. This work was supported by the National Institute of Standards and Technology Innovative Measurement Program on optical medical imaging.

#164628 - $15.00 USD Received 12 Mar 2012; revised 26 Apr 2012; accepted 26 Apr 2012; published 9 May 2012