PNNL-17810

U.S. DEPARTMENT OF
ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Large Spectral Library Problem

LK Chilton
SJ Walsh

September 2008

Pacific Northwest
NATIONAL LABORATORY

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-ACO5-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576 5728
email: reports@adonis.osti.gov

PNNL-17810

Large Spectral Library Problem

LK Chilton
SJ Walsh

September 2008

Prepared for
the U.S. Department of Energy
Under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Executive Summary

Hyperspectral imaging produces a spectrum or vector at each image pixel. These spectra can be used to identify materials present in the image. In some cases, spectral libraries representing atmospheric chemicals or ground materials are available. The challenge is to determine if any of the library chemicals or materials exist in the hyperspectral image. The number of spectra in these libraries can be very large, far exceeding the number of spectral channels collected in the field. Suppose an image pixel contains a mixture of p spectra from the library. Is it possible to uniquely identify these p spectra? We address this question in this paper and refer to it as the Large Spectral Library (LSL) problem. We show how to determine if unique identification is possible for any given library. We also show that if p is small compared to the number of spectral channels, it is very likely that unique identification is possible. We show that unique identification becomes less likely as p increases.

Contents

List of Figures

1 Introduction

Hyperspectral imaging produces a spectrum or vector at each image pixel. These spectra can be used to extract information from the image. In some cases, spectral libraries exist, and the challenge is to determine if any of the library spectra exist in the hyperspectral image. The number of spectra in these libraries can be very large, far exceeding the number of spectral channels collected in the field. A fundamental question is, "Given library spectra, each of length n, how many spectra can be uniquely identified in a hyperspectral image?" We address this question in this paper and refer to it as the Large Spectral Library (LSL) problem. The question is complicated by the presence of noise in spectral measurements as well as model uncertainty.

The "curse of dimensionality" refers to the difficulties associated with high-dimensional data. Yet, hyperspectral imaging is an extension from a low-dimensional view of a scene to a high-dimensional view. The expectation is that important features can only be perceived from the higher dimensional view. This paper explores how the dimension of the data impacts the LSL problem.

A first cut at the problem is to ignore two complicating factors that are the physical constraint of positive coefficients in spectral mixing and spectral noise. This allows us to address the problem with the tools of classical linear algebra, which is done in Section 2. In Section 3, we incorporate the constraint of positive coefficients. In Section 4, we explore how spectral noise affects spectral uniqueness. Section 5 is a summary of the impact of these findings on spectral identification in hyperspectral imaging analysis.

2 A Linear Algebra Perspective

In this section, we explore the LSL problem in the context of classical linear algebra. We ignore both the physical constraint of positive coefficients and spectral noise. These will be addressed in Sections 3 and 4, respectively.

Since individual spectra can be viewed as vectors, we consider the spectral library to be a collection of vectors. Two properties of the spectral library are of interest. The first is whether a pixel vector can be represented by a linear combination of spectra in the library.

The second property addresses the uniqueness of the representation. Of course, if the spectral library is large compared to the number of spectral channels and the library has a maximal linearly independent subset, then there will be infinitely many representations for any vector.

A more interesting question in the context of the LSL problem is the uniqueness of mixtures or linear combinations of length p, where p is small relative to the vector length. For example, let $p = 3$ and the vector length be $n = 100$, where n is the number of spectral channels. If a pixel is a linear combination of three spectra, are there any other linear combinations of three or fewer spectra that can produce the same pixel? We address this question in the following theorem. In the theorem, G represents the spectral library at the resolution of the fielded sensor, and w is an individual pixel vector.

Note: a $q-$tuple is a collection of q vectors.

Theorem 2.1. *Let $G = \{g_1, ..., g_N\}$, $g_i \in \mathbb{R}^n$ and $n \ll N$. Let $q \leq n$ and assume every q-tuple from G is linearly independent. Let $p \leq \frac{q}{2}$ and $v_1, v_2, \ldots, v_p \in G$. Finally, let $w = c_1v_1 + c_2v_2 + \cdots + c_pv_p$, $c_i \neq 0$ and $w \neq 0$. Then w has a unique representation of length p in G.*

Proof. Suppose there is another representation of w of length p or less. In other words, assume there exists $u_1, u_2, \ldots, u_k \in G$ such that

$$w = b_1u_1 + b_2u_2 + \cdots + b_ku_k,\ k \leq p\,.$$

We will show that this assumption leads to a contradiction. Let $\{v_1, v_2, \ldots, v_p\}$ and $\{u_1, u_2, \ldots, u_k\}$ have r elements in common where $0 \leq r < k$ and assume they are ordered such that the common elements are last. Then,

$$w = c_1v_1 + c_2v_2 + \cdots + c_pv_p = b_1u_1 + b_2u_2 + \cdots + b_ku_k$$

and

$$\begin{aligned} c_1v_1 + \cdots + c_{p-r}v_{p-r} + (c_{p-r+1} - b_{k-r+1})v_{p-r+1} + \cdots + \\ (c_p - b_k)v_p - (b_1u_1 + b_2u_2 + \cdots + b_{k-r}u_{k-r}) = 0 \end{aligned}$$

and at least $p + k - 2r$ coefficients in this equation are non-zero. Thus, by definition [Str88], this set of $p + k - 2r \leq q$ vectors is linearly dependent, which contradicts the hypothesis that every $q-$tuple in G is linearly independent. □

This theorem does not say that a longer representation does not exist. It does say there is only one representation of length p or less if the hypotheses of the theorem are satisfied.

It is important to notice in this theorem that for any set G, if every q-tuple is linearly independent, then every r-tuple is also linearly independent where $r \leq q$. In this sense, q is not unique.

For a given set G, one may want to find the largest q, but in the LSL problem, a more realistic concern would be to determine for a given p if all q-tuples are linearly independent where $q = 2p$. In theory, one could answer this question directly by computing the nullspace for every $q-$tuple in G. If the nullspace of every $q-$tuple is empty, then every $q-$tuple is linearly independent. Unfortunately, this approach is quickly overcome by combinatorics if N is large, but may be feasible for small p.

2.1 Interpretation in $\mathbb{R}^3$

Consider a set $G = \{g_1, \ldots, g_N\}$ with $g_i \in \mathbb{R}^3$ and let $p = 1$. Then the theorem hypothesis states that all 2-tuples (pairs) in G are linearly independent. Now, any linear combination of p vectors is the singleton $w = c_1 v_1$, and w points in the same direction as one of the original N vectors. Intuitively, the only way there can be more than one way to represent a single vector using a linear combination of size $p = 1$ is if two (or more) vectors in G point in the same direction. However, the hypothesis guarantees that every pair is linearly independent, which means that no two vectors in G can point in the same direction.

Now, consider $p = 2$. Then $q = 2p = 4 > 3$, and so the theorem does not address linear combinations of 2 or more vectors in $\mathbb{R}^3$.

In summary, in $\mathbb{R}^3$, the theorem only addresses the $p = 1$ case, which is single vectors. The only way single vectors are not unique is when two (or more) identical vectors occur in the original set G.

2.2 Interpretation in $\mathbb{R}^n, n \gg q = 2p$

The LSL problem is concerned with large n; typically $n > 100$ (recall that n is the number of spectral channels). Consider $p = 3$. Then all $q = 2p = 6$-tuples from G can be tested for linear independence. If all 6-tuples are linearly independent, then any vector w that is a linear combination of 3 vectors in G, i.e., $w = c_1v_1 + c_2v_2 + c_3v_3$ has no other representation of length 3 or less. There may be representations of length 4 or greater but none of length 3 or less.

We tested the Pacific Northwest National Laboratory (PNNL) Infrared Spectral Library (IRSL) in the long-wave infrared (LWIR) region to see if all q-tuples are linearly independent. We tested $q = 2, 4, 6$ on random subsets of the IRSL of size up to 200. We did not test the whole library (the IRSL has over 500 spectra) at the same time because of computational limitations. Every q-tuple we tested was linearly independent. For the IRSL this means that it is very likely that all linear combinations of 3 chemical spectra will have only one representation. Linear combinations larger than 3 may also be unique but we have not tested that case.

3 Positive Coefficients

In hyperspectral imaging, when pixel vectors are the result of linear spectral mixing, the mixing coefficients are physically constrained to be positive. So we consider linear combinations $w = c_1v_1 + c_2v_2 + \cdots + c_pv_p$ where every coefficient $c_i > 0$. As before, let $G = \{g_1, ..., g_N\}$ and let $G_{(i)}$ be every spectral vector in G except g_i. Now let $H_{i,2p-1}$ be the set of all $(2p-1)$-tuples from $G_{(i)}$ and let $A \in H_{i,2p-1}$ and $\beta \in \mathbb{R}^{2p-1}$. Claim: If $A\beta = g_i$ has no solution for all $A \in H_{i,2p-1}$, then every linear combination of size p or less from G is unique.

We show this by considering $w = c_1v_1 + c_2v_2 + \cdots + c_pv_p$ and making the assumption that there is another linear combination such that $w = b_1u_1 + b_2u_2 + \cdots + b_ku_k,\ b_i > 0,\ k \leq p$. We will show that this assumption leads to a contradiction. Since we have two representations of w, then

$$c_1v_1 + c_2v_2 + \cdots + c_pv_p = b_1u_1 + b_2u_2 + \cdots + b_ku_k$$

and solving for v_1 yields

$$v_1 = \frac{1}{c_1}[b_1u_1 + b_2u_2 + \cdots + b_ku_k - c_2v_2 - \cdots - c_pv_p] \ .$$

In other words, v_1 can be represented by a $p+k-1$ linear combination in G with coefficients that are both positive and negative. This is also true for each of $v_2, \ldots, v_p, u_1, u_2, \ldots, u_k$. In summary, we have shown that the assumption that there is a second representation of w leads to the conclusion that single vectors can be represented by linear combinations of size no greater than $2p-1$. Note that $2p-1 \geq p+k-1$.

We can test G, and if we can show that every $g_i \in G$ is not a linear combination of $2p-1$ or fewer elements in G, then the assumption that a second representation exists must be false.

We tested random subsets of the IRSL of size 100. We used $p = 3$ and found no cases where $g_i = A\beta$ for any $A \in H_{i,5}$. We did not check the whole library concurrently because of computational limitations. This guarantees that all positive coefficient linear combinations of size 3 or less from the subsets of the IRSL we tested have only one representation. It also provides evidence that this is also true for the whole library.

Another way to evaluate the positive coefficient case is related to Theorem 2.1. There we were interested in q-tuples where

$$c_1v_1 + \cdots + c_pv_p - b_1u_1 - \cdots - b_ku_k = 0 \tag{1}$$

and allowed the coefficients $c_1, \ldots, c_p, b_1, \ldots, b_k$ to be either positive or negative. In this section, we restrict our search to positive coefficients, guaranteeing that we are searching over a smaller set. Thus, if no q-tuples were found satisfying (1) when both positive and negative coefficients are allowed, then we are guaranteed to not find any when we only allow positive coefficients.

4 Spectral Noise

In this section, we explore how spectral noise affects the uniqueness of linear combinations of gases from the IRSL. As before, we are given a linear combination

$$w = c_1 v_1 + c_2 v_2 + \cdots + c_p v_p \,, \quad c_i > 0$$

from G (where G is the IRSL), and we want to know if there are other linear combinations $\tilde{w} = b_1 u_1 + \cdots + b_p u_p,\ b_i \geq 0$ from G that are close enough to w that they would produce essentially the same signal when measured by the same instrument. The signal measured due to w is $y = w + \epsilon$, and the signal measured due to $\tilde{w}$ is $\tilde{y} = \tilde{w} + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$ represents the measurement error. So if $\tilde{w}$ is close enough to w, $\tilde{y}$ will be indistinguishable from y if σ^2 is large enough. We will analyze the IRSL to indicate how likely nearly identical linear combinations are as a function of p, the size of the combination. We will use the minimum singular value (MSV) to measure how similar two sets of vectors are. How and why we use the MSV will be explained in what follows. We expect the likelihood of nearly identical cases to increase with p.

Let $G = \{g_1, ..., g_N\}$ represent the IRSL, and $w = c_1 v_1 + \cdots + c_p v_p,\ v_i \in G,\ c_i > 0$. Then $\tilde{w} = b_1 u_1 + \cdots + b_p u_p,\ u_i \in G,\ b_i \geq 0$ will be indistinguishable from w when $\|w - \tilde{w}\| \leq T$ or

$$\|c_1 v_1 + \cdots + c_p v_p - b_1 u_1 - \cdots - b_p u_p\| \leq T \tag{2}$$

where T is called a *threshold* and depends on environmental and instrument parameters and the identification algorithm. In this case $\| \circ \|$ denotes the Euclidean norm.

When $v_1, \ldots, v_p, u_1, \ldots, u_p$ are linearly dependent, there are linear combinations of them that equal 0, so the inequality in (2) could be satisfied for any threshold T. One way to test for linear independence of a set of vectors $v_1, \ldots, v_p, u_1, \ldots, u_p$ is to compute the rank of the $n \times 2p$ matrix $X = [v_1, \ldots, v_p, u_1, \ldots, u_p]$. Recall, we assume $n \gg 2p$. Now, if the rank equals the number of columns of X, then the columns are linearly independent. If the rank is less than the number of columns, then the columns are linearly dependent.

A very reliable method for computing the rank is singular value decomposition where the rank equals the number of non-zero singular values. However, if the MSV is non-zero but very close to 0, then even though the set of vectors is technically linearly independent, they

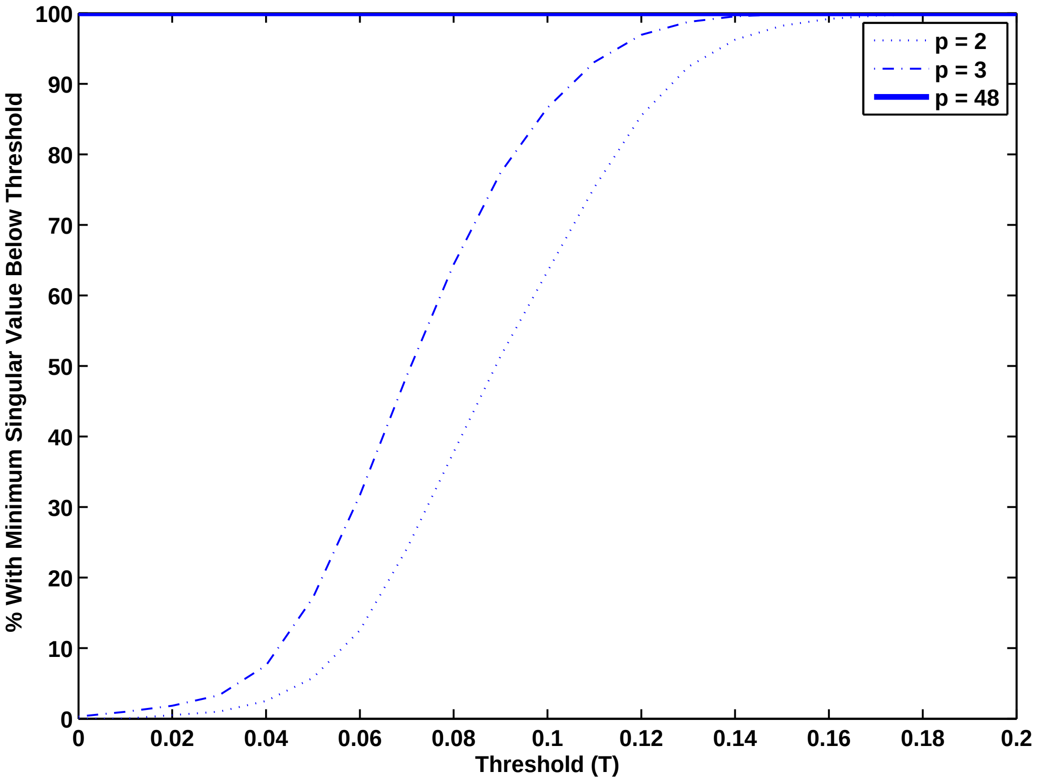

Figure 1: Percent of $2p$-tuples whose minimum singular value (MSV) is below the threshold (T) (indicating they are linearly dependent). For fixed T, the percent of $2p$-tuples with MSV below T increases as p increases. For each p, the percent of $2p$-tuples with MSV below T decreases as T decreases. The $p = 48$ case is hard to see because it is 1 for all values of T.

are effectively linearly dependent [Dem97]. This means that T in (2) is very close to zero. We use the MSV to indicate how close a $2p$-tuple is to being linearly dependent. If the MSV is not close to zero then the $2p$-tuple is linearly independent. If the MSV is zero or close to zero then the $2p$-tuple is linearly dependent.

Remark: Linear dependence of the set $v_1, \ldots, v_p, u_1, \ldots, u_p$ is a conservative test for existence of a $\tilde{w} = b_1u_1 + \cdots + b_pu_p$ that is close to $w = c_1v_1 + \cdots + c_pv_p$. Linear dependence says there is a linear combination of $v_1, \ldots, v_p, u_1, \ldots, u_p$ that equals zero. For a given $w = c_1v_1 + \cdots + c_pv_p$, it does not guarantee there is a $\tilde{w} = b_1u_1 + \cdots + b_pu_p$ such that $w - \tilde{w} = c_1v_1 + \cdots + c_pv_p - b_1u_1 - \cdots - b_pu_p = 0$.

Due to computational limitations, we analyze subsets of the IRSL of size 100. For each p, we compute the MSV of every $2p$-tuple from the subset. We count those $2p$-tuples whose MSV is below the threshold (T) (indicating the $2p$-tuple is nearly linearly dependent) for a set of threshold values from 0 to 0.2. The results are given in Figure 1. We see that for fixed T, the percent of $2p$-tuples with MSV below T increases as p increases. This represents the case for a given imaging scenario, where the instrument and environment are fixed for a specific image, indicating a specific (although unknown) value of T. This means that for a given T, as p increases the percent of p-tuples that have a unique representation decreases.

For each p, the percent of $2p$-tuples with MSV below T decreases as T decreases. The $p = 48$ case is hard to see because it is essentially 1 for all values of T, even when T is close to 0.

5 Summary and Conclusions

In this report, in Section 2 we have shown that for a set of spectral vectors G, if we ignore noise and if all $2p$-tuples in G are linearly independent, then all p-tuples have a unique representation.

When we consider the effect of noise, a good way to quantify how close two p-tuples are is to compute the MSV of the combined $2p$-tuple. If the MSV is very close to 0, then the two p-tuples will produce essentially the same response when measured by an instrument.

This is an important result for the LSL Problem. The MSV at which two p-tuples, w and $\tilde{w}$, become indistinguishable from each other depends on the instrument noise, environmental conditions, and the identification algorithm. For every combination of these three factors there is an effective threshold, T_e. If the MSV of a $2p$-tuple is below T_e, then the two p-tuples that were combined to form the $2p$-tuple will be indistinguishable.

For the PNNL IRSL, we see in Figure 1 that if T_e is about 0.2 then for any $p \geq 2$, all p-tuples can be duplicated by another p-tuple. On the other hand, if T_e is 0.02 then for $p = 2, 3$ almost all p-tuples are unique. For $p = 48$, no matter how small T_e is, every p-tuple will have another p-tuple that is nearly identical. These results emphasize the value of developing instruments and algorithms that reduce T_e.

References

[Dem97] James W. Demmel. *Applied Numerical Linear Algebra.* Society for Industrial and Applied Mathematics, Philadelphia, 1997.

[Str88] Gilbert Strang. *Linear Algebra and Its Applications.* Harcourt Brace Jovanovich, Orlando, FL, 3^{rd} edition, 1988.