This item is the archived peer-reviewed author-version of:

Estimation of the intrinsic dimensionality in hyperspectral imagery via the hubness phenomenon

Reference:

Heylen Rob, Parente Mario, Scheunders Paul.- Estimation of the intrinsic dimensionality in hyperspectral imagery via the hubness phenomenon

Proceedings of the 13th International Conference on Latent Variable Analysis and Signal Separation, LVA/ICA 2017, Grenoble, France, February 21-23, 2017 / Tichavský, Petr [edit.] - ISBN 978-3-319-53546-3 - Cham, Springer, 2017, p. 357-366

Full text (Publishers DOI): http://dx.doi.org/doi:10.1007/978-3-319-53547-0_34

Estimation of the intrinsic dimensionality in hyperspectral imagery via the hubness phenomenon

Rob Heylen¹, Mario Parente², and Paul Scheunders¹

¹ Visionlab, University of Antwerp, Belgium
{rob.heylen, paul.scheunders}@uantwerpen.be

² Electrical and computer engineering, University of Massachusetts, USA
mparente@ecs.umass.edu

Abstract. As hyperspectral images are high-dimensional data sets containing a lot of redundancy, a first important step in many applications such as spectral unmixing or dimensionality reduction is estimation of the intrinsic dimensionality of the data set. We present a new method for estimation of the intrinsic dimensionality in hyperspectral images based upon the hubness phenomenon, which is the observation that indegree distributions in a K-nearest neighbor graph will become skewed as the intrinsic dimensionality of the data set rises. The proposed technique is based upon comparing the indegree distributions of artificially generated data sets with the one from the target data set, and identifying the best match with some histogram metric. We show that this method obtains superior results compared to many alternatives, and does not suffer from the effects of interband and spectral correlations.

1 Introduction

A pixel in a hyperspectral image samples the full spectrum of the area in its field of view, and typically contains hundreds of spectral bands spanning the visible and infrared regions. Such images are typically obtained from a large distance (aerial or satellite), and the area captured in the field of view of each pixel becomes large, of the order of hundreds of meters squared. Therefore, the spectrum captured in one pixel contains information about all the objects in the field of view of that pixel, and one has to take this spectral mixing into account. A rich literature exists on "spectral unmixing" techniques, which attempt to reverse this spectral mixing operation. See [1, 2] for recent overviews.

While several exceptions exist, most unmixing techniques generally assume that a single spectrum, or "endmember", exists for every pure material in the scene, and each observed spectrum can then be considered to be composed of these endmember spectra. In the simplest model, one assumes that each spectrum can be described as a convex linear combination of endmember spectra, leading to the linear mixing model (LMM) [1]. Several more advanced models use higher-order polynomial representations to account for multiple reflections

between objects, or use physical modeling via radiative transfer theory or computer simulations to obtain even more accurate mixing models, but with the disadvantage of complex inversion strategies [2].

An important first step in any of these modeling and inversion attempts is to estimate the number of endmembers that one is going to employ. In most spectral mixing models, this number of endmembers (NOE) can be easily related to the intrinsic dimensionality (ID) of the data set. If one assumes that the data set lies on a smooth manifold in spectral space, the ID can be defined as the dimensionality of the Euclidean space to which this manifold is locally homeomorph. Several authors also describe the ID as the minimum number of degrees of freedom one requires to fully describe or model the data set. In the LMM, the NOE p is related to the ID d as $d = p - 1$, as the data points are constrained to a $(p-1)$ -dimensional simplex due to the convexity constraints, and in the absence of noise. In nonlinear models, the ID can be larger, as one might have additional hyperparameters. These results indicate that ID estimation can provide a good approximation for the NOE in many cases. Remark that proper estimation of the ID is also important in several other applications, such as dimensionality reduction, classification and clustering.

Many techniques exist in the hyperspectral unmixing literature for estimating the NOE, and these can be roughly divided into two classes: The first class attempts to estimate the dimensionality of the linear subspace that contains the data, while the second class employs manifold techniques. Examples of the former are the Harsanyi-Farrand-Chang virtual dimensionality (VD) technique [4] and hyperspectral signal identification by minimum error (HySime) [5]. An example of the latter is the Grassberger-Procaccia algorithm [6].

Any property that depends strongly on the ID of the data set and is invariant with respect to other parameters might be an appropriate estimator for the ID and the associated NOE in a hyperspectral data set. We present a new method for estimating the ID of a hyperspectral image, based on the hubness phenomenon [7]. The hubness phenomenon is an effect which occurs when the ID of a data set increases: If one creates a K -nearest neighbor (KNN) graph in a random data set, one observes that some of the data points will become highly connected to others when the ID grows, and start to act like hubs in the corresponding graph. The indegree distribution (IDD) of the graph, i.e. the distribution of the number of incoming connections from other data points, will become skewed to higher numbers. This hubness effect is present in numerous real-world high-dimensional data sets, and depends strongly on the ID of the data [7].

The proposed technique is based on generating IDD's for many simulated hyperspectral data sets with known parameters, and comparing the observed IDD with these artificially generated distributions. We show that this approach functions very well on real data sets, and is superior to many alternatives on artificially generated hyperspectral data. The technique possesses only one parameter K used for the KNN graph, but shows no strong dependence on this parameter. Furthermore, we also require a denoising procedure, as the ID depends strongly on the noise in the data set. Remark that techniques which exploit

the hubness phenomenon can be considered to belong to the class of manifold techniques, as they employ only a local property of the data, i.e. the KNN graph. This implies that the technique can also be employed for nonlinear mixing situations, where the data lies on a manifold instead of a linear subset. Furthermore, several alternative techniques show a severe underestimation of the NOE in real and simulated data sets due to the inherent correlations which are present between spectra on the one hand, and between adjacent spectral bands on the other hand. The proposed method does not suffer from these effects.

2 The hubness phenomenon

Let $X = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ be a data matrix containing N data vectors $\mathbf{x}_i \in \mathbb{R}^d$ columnwise, with d the embedding dimension of the data set. Let $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ be the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The first or nearest neighbor $\text{nn}_1(\mathbf{x}_i)$ of $\mathbf{x}_i$ is the point $\mathbf{x}_j$ of smallest Euclidean distance. This concept can be extended to the K nearest neighbor $\text{nn}_K(\mathbf{x}_i)$, given by the point with the K 'th smallest distance to $\mathbf{x}_i$. A KNN graph can be constructed by considering the set of directed edges $(\mathbf{x}_i, \text{nn}_j(\mathbf{x}_i))$, with $i = 1, \dots, N$ and $j = 1, \dots, K$, where the corresponding weights of each edge are given by its Euclidean length. The indegree $N_K(\mathbf{x})$ of a point $\mathbf{x}$ is defined as the number of times that $\mathbf{x}$ occurs as a neighbor of other points in the KNN graph of the data set X [7]. The distribution of N_K for all data points is the IDD of the data set.

Points with a high indegree relative to the average indegree act like hubs in the graph. The hubness phenomenon describes the observation that the relative number of hubs increases as the ID of the data set increases. In practice, this means that the IDD of the graph becomes skewed, and that more points with very high degrees show up as the ID increases. The cause for the occurrence of hubness is the concentration of measure phenomenon (see [7] for a detailed analysis): The average distance to the data centroid increases linearly with the intrinsic data dimensionality d , but its variance as $\sqrt{d}$. Therefore, the fraction of points which lie significantly closer to the centroid will decrease with increasing d . It is hypothesized that these points will become hubs, as they will lie closer than average to all other points, and hence will have a larger than average probability of being identified as a near neighbor.

As the intended application is estimation of the ID and the associated NOE in hyperspectral imagery, one can use simulated hyperspectral images to illustrate the hubness phenomenon. We employ the popular LMM to generate the artificial data sets, applied to spectra obtained in a laboratory setting. First, obtain p endmembers $E = (\mathbf{e}_1, \dots, \mathbf{e}_p)$ randomly from a spectral library, e.g. the USGS spectral library of minerals, and interpolate them at the same d wavelengths, leading to a data set with an ID equal to $\min(p-1, d)$. Next, create N abundance vectors $A = (\mathbf{a}_1, \dots, \mathbf{a}_N)$ randomly in a unit simplex by using the symmetric Dirichlet distribution (SDD):

$$P(\mathbf{a}|\alpha) = \begin{cases} \frac{\Gamma(\alpha p)}{\Gamma(\alpha)^p} \prod_{i=1}^p a_i^{\alpha-1} & \text{if } \mathbf{a} \in S_p \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

with S_p the unit simplex in p dimensions. The SDD will always return abundance vectors that obey convexity: They are non-negative and sum to one. Furthermore, it depends on a single hyperparameter α which can be used to skew the distribution: For $\alpha = 1$ one obtains a uniform distribution over the unit simplex. For $\alpha > 1$ the abundance vectors will be more probable close to the center of the simplex, while for $\alpha < 1$ the abundance vectors will become sparse, and will lie closer to the borders of the simplex. See Fig. 1 for an example.

Fig. 1: Abundance vectors generated by the SDD for different values of α

We use the LMM to generate the data set:

$$X = EA + \boldsymbol{\eta} \quad (2)$$

where $\boldsymbol{\eta}$ is a noise matrix of size (d, N) containing independent Gaussian noise with a given signal-to-noise ratio (SNR). The IDD is found by determining the KNN graph for a given value of K , counting how many times each point occurs as a neighbor of any other point, and aggregating these degrees in a normalized histogram. These IDDs are shown in Fig. 2 for a noiseless situation with uniform abundance values ($\alpha = 1$ in the SDD), $N = 10^4$, $d = 50$ and $K = 10$, and for several values of p . It can be observed that the shape of these IDDs will indeed shift to heavy-tailed distributions as p , and hence the ID, increases.

Fig. 2: The IDD for the artificial USGS data sets for $p \in [2, 50]$ (top to bottom), with $K = 10$ and $N = 10000$, each graph averaged over 100 random instances.

3 Properties of hubness in hyperspectral data sets

In order to better understand the properties of the hubness phenomenon in hyperspectral data sets, we present several results and observations acquired on artificial and simulated hyperspectral data sets. Due to lack of space, we cannot extensively discuss and illustrate these dependencies, but we refer to [8, 9] for more information on some of these items.

The hubness phenomenon shows strong dependencies on several aspects of the data set:

- The hubness phenomenon depends strongly on the **noise** in the data set, as noise is typically considered to be independent Gaussian noise, and hence acts as a full-dimensional additional component on top of the signal component of lower ID. This will obscure the actual ID, and stresses the need for a proper denoising procedure before estimation of ID or the NOE.
- Also the **coherence in the spectral library** which was employed to create the artificial data sets has a significant impact: If one uses a library where a high average pairwise correlation (APC) exists between spectra, then a subset of randomly chosen spectra has a large probability of pairs lying close together in spectral space. Such a pair might not introduce enough variability in the data set to be detected as an additional degree of freedom, and hence the hubness phenomenon will not be as explicit as the NOE rises when a highly coherent library is employed. An example of a highly coherent library is the USGS vegetation library, with an APC of 0.78, where the mineral and artificial materials libraries have an APC of 0.17 and 0.18 respectively.
- The value of K which was employed to generate the data set also impacts the hubness phenomenon. For low ID, the IDD will have a modus equal to K , and this modus will decrease for increasing ID. As long as the same value of K is used for all comparisons, this dependence is not an issue.

Several other aspects show only a very small or no dependence:

- The **size of the data set** has only a very small effect on the IDD.
- Also the **embedding dimension** does not influence the IDD when the noise is small or non-existent. This can be easily explained due to the independence of local interpoint distances on the embedding dimension when all points are restricted to a lower-dimensional manifold.
- Several data sets show an intrinsic **sparseness**, and also in hyperspectral data sets, it is often more realistic to assume that only a small subset of endmembers contributes to the spectrum observed in a given pixel. This leads to an ID which is locally lower than the ID of the full data set. This effect can be observed for instance by modifying the α parameter used to generate the abundance distributions. This effect stays very small.
- Also the **library size** has no impact on the IDD, as random subsets of different sizes yield identical IDD.

- Finally, a surprising result is that also the **metric** which was employed to create the KNN graphs has no influence at all on the IDD. We employed the L_1 , L_2 , L_∞ , correlation distance, spectral angle distance and spectral information divergence to create the KNN graphs in the same data sets, and obtained virtually identical IDDs in every case

These results indicate that the IDDs and the associated hubness phenomenon are indicative of the ID of the data set as long as the noise and library coherence are kept under control. Any other parameters used in the data creation will not affect the IDDs.

4 The algorithm

We want to exploit the strong dependence of the IDD on the ID of the data set to find an estimate of the NOE in a hyperspectral data set. The underlying idea is to generate a library of representative IDDs by artificially generating hyperspectral data sets with a known NOE, and then comparing the IDD of the target data set with the IDDs in this library by using some histogram metric. This way, the artificial data set which corresponds with the target data set can be identified, and the corresponding NOE can be derived.

The results in previous section 3 show that a denoising procedure has to be applied to the target data set in order to avoid overestimation of the NOE, as the presence of noise will typically increase the ID. Because no denoising technique can be expected to function perfectly, the obtained estimates must be considered as maximal values. This leads to the following algorithm:

1. Build an IDD library by generating a large number of random hyperspectral data sets with a known NOE, and determining the average IDD.
2. Denoise the target data set.
3. Determine the IDD of the target data set.
4. Compare this target IDD with those in the IDD library by using a histogram similarity metric to identify the data set with the largest correspondence.

To denoise the target data set, we employed the same method as in the HySime algorithm [5], where each spectral band is recreated from the others by linear regression. Remark that the results might depend significantly on the denoising method used, and that better denoising techniques might exist.

Many histogram or probability distribution metrics can be used to compare two histograms P and Q over the nonnegative integers. We chose to employ the total variation (TV) distance, as it is a simple and intuitive way to compare histograms.

5 Results

5.1 Artificial data sets

We have assessed the performance of the proposed algorithm on artificial data sets, and compared the obtained results with those obtained with several al-

alternatives from the literature. These are the HySime algorithm [5], the VD algorithm [4], and the denoised hyperspectral intrinsic dimensionality estimation with nearest-neighbor distance ratios (D-HIDENN) method, a technique we developed earlier based on the statistics of nearest-neighbor distance ratios [3]. Remark that the proposed method, HySime and the D-HIDENN method all use the same noise estimation technique. The HySime method has no parameters, while the VD algorithm depends on a false-alarm parameter (chosen to be 10^{-4}), and the D-HIDENN method has two integer parameters, chosen to be 1 and 3. We used $K = 10$ for the KNN graphs in the IDD method.

The data sets were generated using randomly selected spectra from the USGS database, which were mixed according to the LMM (2) with a uniformly and randomly chosen abundance matrix respecting the ASC and ANC (equation (1) with $\alpha = 1$). Remark that such data sets were also used for generating the artificial IDD, with the difference that now noise is present.

For every set of parameters, we averaged over 100 random data sets. The first experiment, displayed in Fig. 3 (a), assesses the performance as a function of the noise in the data set. These figures illustrate that for low SNR values, all methods show significant deviations from the true value, and also the variability in the results can be very high. For higher SNR however, the IDD method performs the best on average.

The dependence on the true value of p is illustrated in Fig. 3 (b), where the estimates are plotted as a function of p . Note that on average, the IDD method obtains the correct result for all values of p , although with a large variance which can be explained by the creation of these data sets: When a high number of endmembers are randomly drawn from a spectral library, certain subsets of these can be highly correlated. Depending on the number of correlated endmembers, different IDs can be expected in the resulting data sets. The alternative methods show the typical underestimating behavior for increasing p discussed in previous sections. These results indicate that the IDD method yields better results over the alternatives on these types of artificial data sets. Remark that the bad performance for the lowest values of p can be attributed to the denoising procedure which does not effectively remove all noise in this case, and that correct results are obtained here in a noiseless scenario. More examples of the p and SNR dependence can be found in [8, 9].

Furthermore, one can assess the performance of these NOE estimation techniques on nonlinearly mixed data as well. An easy way to create a nonlinear data set is to employ the PPNM model, a popular nonlinear mixing model which belongs to the family of bilinear mixing models. This model creates a new data set from an existing, linearly mixed, data set via a polynomial transformation. We employ a second-order transformation of the data set as $\mathbf{x} = \mathbf{y} + \mathbf{y}^2$, with $\mathbf{y}$ the result from the LMM. This yields a simple, nonlinearly mixed data set, but with the same ID as the original linearly mixed data set as one does not add degrees of freedom by this transformation. When the data sets employed in Figs. 3 are transformed with the PPNM and the experiments are run again, one obtains Fig. 4. This figure illustrates that the proposed technique obtains the correct

Fig. 3: The estimates for the NOE as (a) a function of the SNR and (b) as a function of p , for data sets mixed with the LMM with $N = 10^4$ pixels and (a) $p = 20$ and (b) SNR = 50.

answer on average also for nonlinearly mixed data sets as long as the noise is relatively small, while the alternative techniques still show large deviations.

Fig. 4: The estimates for the NOE as (a) a function of the SNR and (b) as a function of p , for data sets mixed with the PPNM model with $N = 10^4$ pixels and (a) $p = 20$ and (b) SNR = 50.

5.2 Real data sets

We have also executed the proposed IDD method on several well-known hyper-spectral data sets. The data sets under consideration are the Airborne Visible / Infrared Imaging Spectrometer (AVIRIS) Cuprite and Indian Pines data sets. These consist of 220-band images with an IFOV of $20 \times 20 m^2$, but we removed the water absorption bands ([1:4], [103:113], [148:166]) before use.

We have plotted the IDD along with the best match from the library in Fig. 5 (a) for the Cuprite data set, and in Fig. 5 (b) for the Indian Pines data set. These figures illustrate that the IDD from the data sets indeed closely resembles one of the artificially calculated IDD in the library. The TV distance is given in Fig. 5 (c) for both data sets. The minimal value of this distance identifies the NOE returned by the algorithm.

Fig. 5: The IDD and the best match for the (a) Cuprite and (b) Indian Pines data sets. (c) The TV distance as a function of p .

The estimates obtained by the different algorithms are displayed in table 1. The VD technique was used with several values for the false-alarm parameter. The HIDENN and D-HIDENN algorithms used the values $(1, 3)$ for their (k, k') parameter pair. The other algorithms do not depend on a parameter.

	Cuprite	Indian Pines
HFC VD 10^{-1}	41	56
HFC VD 10^{-2}	30	34
HFC VD 10^{-3}	22	26
HFC VD 10^{-4}	18	23
HySime	19	14
HIDENN	18	18
D-HIDENN	12	10
IDD method	25	18

Table 1: The estimates of the NOE obtained by the different techniques, on the AVIRIS Cuprite and Indian Pines data sets.

This table shows that a wide variety of estimates can be returned. While it is not known what the correct NOE would be in these data sets, classification results on these data sets give a hint to the NOE: 16 classes are employed in the Indian Pines data set, along with a background class. The well-known Tetra-corder classification map of the AVIRIS Cuprite data set employs 25 classes, along with an unknown class. While the number of classes cannot be considered

exactly the same as the NOE for several reasons, such as mixed classes, missing classification results and spectral variability, they do indicate that the proposed method obtains similar numbers.

6 Conclusions

We have presented a new method for estimating the ID in a hyperspectral image, based on comparing the IDD of the target hyperspectral image with IDDs obtained from simulated hyperspectral images. As IDD depends strongly on noise, all data sets have been denoised before estimation of ID. The proposed technique is validated on artificial data sets, and shows superior results compared to several popular alternatives, such as VD and HySime. Also, the proposed technique does not suffer from the often observed underestimation effect caused by spectral correlations between endmembers and between adjacent bands. On real data sets, realistic values seem to be obtained.

Future work involves the inclusion of more advanced denoising procedures, and the inclusion of prior knowledge on the target data set into the procedure (e.g. IDD libraries which depend on the type of data set, or inclusion of expected sparseness into the estimation procedure).

References

1. J.M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader and J. Chanussot: Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 5(2), 354–379 (2012)
2. R. Heylen, M. Parente and P. Gader: A review of nonlinear hyperspectral unmixing methods. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 7(6), 1844–1868 (2014)
3. R. Heylen and P. Scheunders: Hyperspectral intrinsic dimensionality estimation with nearest-neighbor distance ratios. *IEEE J. Sel. Top. Applied Earth Observation and Remote Sensing* 6(2), 570–579 (2013)
4. C.-I Chang and Q. Du: Estimation of the number of spectrally distinct signal sources in hyperspectral imagery. *IEEE Trans. Geosci. Remote Sens.* 42(3), 608–619 (2004)
5. J.M. Bioucas-Dias and J.P.M. Nascimento: Hyperspectral Subspace Identification. *IEEE Trans. Geosci. Remote Sens.* 46(8), 2435–2445 (2008)
6. Q. Du: Virtual dimensionality estimation for hyperspectral imagery with a fractal-based method. *IEEE Workshop on Hyperspectra Image and Signal Processing: Evolution in Remote Sensing*, 1–4 (2010)
7. M. Radovanovic, A. Nanopoulos and M. Ivanovic: Hubs in space: popular nearest neighbors in high-dimensional data. *J. Machine Learning Res.* 11, 2487–2531 (2010)
8. R. Heylen, M. Parente and P. Scheunders: Estimation of the number of endmembers via the hubness phenomenon. *IEEE Workshop on Hyperspectra Image and Signal Processing: Evolution in Remote Sensing*, 1–4 (2016)
9. R. Heylen, M. Parente and P. Scheunders: Estimation of the number of endmembers in a hyperspectral image via the hubness phenomenon. *IEEE Trans. Geosci. Remote Sens.*, to be published.