Discriminative Archetypal Self-taught Learning for Multispectral Landcover Classification

Ribana Roscher, Susanne Wenzel
Photogrammetry Lab
Institute of Geodesy and Geoinformation
University of Bonn
Germany, 53115 Bonn
ribana.roscher@uni-bonn.de
susanne.wenzel@uni-bonn.de

Björn Waske
Division of Remote Sensing and Geoinformatics
Institute of Geographical Sciences
Freie Universität Berlin
Germany, 12249 Berlin
Email: bjoern.waske@fu-berlin.de

***Abstract*—Self-taught learning (STL) has become a promising paradigm to exploit unlabeled data for classification. The most commonly used approach to self-taught learning is sparse representation, in which it is assumed that each sample can be represented by a weighted linear combination of elements of a unlabeled dictionary. This paper proposes discriminative archetypal self-taught learning for the application of landcover classification, in which unlabeled discriminative archetypal samples are selected to build a powerful dictionary. Our main contribution is to present an approach which utilizes reversible jump Markov chain Monte Carlo method to jointly determine the best set of archetypes and the number of elements to build the dictionary. Experiments are conducted using synthetic data, a multi-spectral Landsat 7 image of a study area in the Ukraine and the Zurich benchmark data set comprising 20 multispectral Quickbird images. Our results confirm that the proposed approach can learn discriminative features for classification and show better classification results compared to self-taught learning with the original feature representation and compared to randomly initialized archetypal dictionaries.**

I. Introduction

Landcover classification is one of the major topics within the remote sensing community. In this context, multiple paradigms has been introduced covering supervised, semi-supervised or transfer learning. Semi-supervised learning has become popular, because it exploits unlabeled data to improve over supervised classification models [1]. However, semisupervised learning is limited to the assumption that both labeled and unlabeled data follow the same distribution. In contrast to this, self-taught learning (STL, [2]) utilizes both labeled and unlabeled data without the requirement that both sets have to share the same distribution. This makes the approach suitable for classification tasks using satellite remote sensing data, where generally massive amounts of unlabeled data with unknown distribution exists.

The most common approach to STL is sparse representation, in which each sample is approximated by a weighted linear combination of a few elements of a dictionary. The dictionary is learned from unlabeled data and thus, the choice of the elements is crucial for the classification result. One major difference in dictionary learning is either to select representative samples from a large set such as cluster centers obtained by k-means [3] or archetypes [4], or to learn an adapted dictionary with methods like K-SVD [5]. The latter one may result in a better reconstruction of samples, however, the learned dictionary elements are no real data samples anymore and therefore, not interpretable. In many applications, such as unmixing or classification with limited user-interaction interpretable elements are required (e.g., [6], [7], [8]), so that a selection of suitable elements is preferable.

STL has become famous in the context of unsupervised feature learning for classification tasks. Specifically, the obtained sparse coefficient vector is used as higher-level feature representation, which serves as input into a classifier. In contrast to unmixing tasks, where a high reconstruction ability is necessary, dictionary elements for classification should lead to a high discrimination power of the new feature representation. For example, [4] show that STL in combination with archetypal dictionaries lead to higher classfication accuracies than using the original feature representation. However, finding the best set of archetypes is still an open research question and strongly related to topics in the deep learning community [9] or for unmixing remote sensing data [10]. Approaches such as e.g. N-FINDR [11] are commonly used, but may return a local optimum and thus, a sub-optimal result with low classification accuracies. Another possible solution would be to design overcomplete dictionaries with a large amount of archetypes. However, this is not appropriate for STL, since generally the number of training samples is low and an overcomplete dictionary would lead to a too diverse, inappropriate feature representation.

In this paper, we propose an approach which utilizes reversible jump Markov chain Monte Carlo (rjMCMC, [12]) method to identify a suitable set of discriminative archetypes. As criteria to evaluate the set of archetypes during optimization, we use the reconstruction error of sparse representation with the current archetypal dictionary and the logistic regression error function value of the new feature representation to ensure a high discrimination power. We show in our experiments that the extracted archetypal dictionaries provide discriminative features for classification.

II. Self-Taught Learning

The self-taught learning procedure uses labeled training data and unlabeled data, which can belong to arbitrary classes and need not to follow the same distribution as the labeled data. The training set is given by $\left({}^{l}\boldsymbol{x}_n, {}^{l}y_n\right)$, $n = 1, \ldots, N$ of N labeled samples with M-dimensional feature vectors ${}^{l}\boldsymbol{x}_n \in \mathbb{R}^M$ and class labels ${}^{l}y_n \in \mathcal{C} = \{1, \ldots, c, \ldots, C\}$. The unlabeled data set is denoted by ${}^{u}\boldsymbol{x}_p$, $p = 1, \ldots, P$ and the test data is given by ${}^{t}\boldsymbol{x}_q$, $q = 1, \ldots, Q$. It must be noted that the test data, used for evaluation purposes, follows the same distribution as the labeled training data.

A. Sparse Representation

In terms of sparse coding a training sample ${}^{l}\boldsymbol{x}_n$ is represented by a weighted linear combination of a few elements taken from a $(M \times T)$-dimensional dictionary $\boldsymbol{D}$, so that

$$ {}^{l}\boldsymbol{x}_n = \boldsymbol{D}\, {}^{l}\boldsymbol{\alpha}_n + \boldsymbol{\gamma} \tag{1} $$

with $\|\boldsymbol{\gamma}\|_2$ being the reconstruction error (see Fig. 1). The dictionary $\boldsymbol{D} = [\boldsymbol{d}_t]$ is embodied by unlabeled data samples such that $\{\boldsymbol{d}_t\} \in \{{}^{u}\boldsymbol{x}_p\}$, where generally $T \leq P$ with T as the number of dictionary elements. The coefficient vector comprising the weights is given by ${}^{l}\boldsymbol{\alpha}_n$. The optimization problem for the determination of optimal ${}^{l}\hat{\boldsymbol{\alpha}}$ is given by

$$ {}^{l}\hat{\boldsymbol{\alpha}}_n = \underset{\boldsymbol{\alpha}_n}{\operatorname{argmin}} \ \|\boldsymbol{D}\, {}^{l}\boldsymbol{\alpha}_n - {}^{l}\boldsymbol{x}_n\|_2, \tag{2} $$

$$ \text{subject to} \quad \|{}^{l}\boldsymbol{\alpha}_n\|_0 > Z, \ {}^{l}\boldsymbol{\alpha}_n \geq 0 \tag{3} $$

where the first term is the reconstruction error and the second term is the L_0-norm enforcing sparsity. We further introduce a non-negativity constraint. We solve the equation with orthogonal matching pursuit in combination with non-negative least squares optimization.

A classifier model is trained with $\begin{bmatrix}{}^{l}\hat{\boldsymbol{\alpha}}_n\end{bmatrix}$ being the new higher-level feature representations of $\begin{bmatrix}{}^{l}\boldsymbol{x}_n\end{bmatrix}$ with respect to the dictionary $\boldsymbol{D}$. In the same way, higher-level features are extracted for test samples ${}^{t}\boldsymbol{x}_q$, which are classified by the learned model.

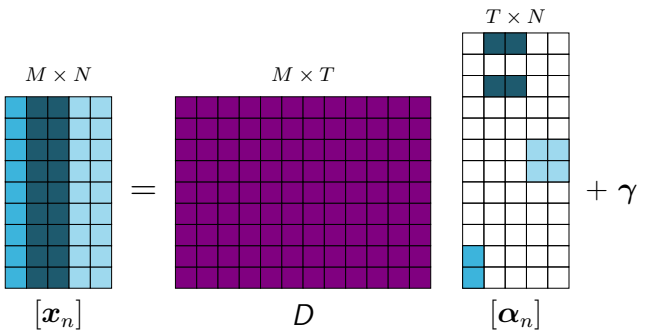


Fig. 1. Self-taught learning with sparse representation: Each sample $\boldsymbol{x}_n$ is represented by a weighted linear combination of a few elements taken from a $(M \times T)$-dimensional dictionary $\boldsymbol{D}$, which contains only unlabeled data. Learned weights are used as new data representation for classification.

III. Archetypal Dictionary Learning

The dictionary is constructed by archetypal analysis, which was introduced by [13] as a variant of principal component analysis (PCA). PCA achieves a sparse data approximation by decomposing a data matrix to $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$, where $\boldsymbol{X} = [\boldsymbol{x}_n]$, $\boldsymbol{W}$ is a set of basis vectors and $\boldsymbol{H}$ are the coefficients for data reconstruction. The signs in $\boldsymbol{W}$ and $\boldsymbol{H}$ are arbitrary and thus, the basis vectors have no interpretable physical meaning. Archetypal analysis therefore restricts the basis vectors in $\boldsymbol{W}$ to lie within the column space of $\boldsymbol{X}$, i.e.the basis vectors are real data points or positive combinations of real data points. This, on the one hand, leads to an improved interpretability of the low rank approximations. On the other hand, the algorithmic complexity to find the basis vectors scales quadratically with the size of the data. As a consequence, archetypal analysis is not suitable for typically large hyperspectral data sets. Therefore, [14] introduced a greedy approach called simplex volume maximization (SiVM), which restricts the basis vectors to data points lying on the approximation of the convex hull of the data matrix. This means that each data point can be expressed as a linear combination of the most extreme data points, the archetypes. The basic idea of this approach is to successively collect archetypes by choosing this sample as archetype, which is farthest away from all former selected ones. This is equivalent to maximizing the volume, which is spanned by the archetypes. SiVM was first applied for unsupervised classification of hyperspectral data in [8] using the coefficients as higher-level features to express similarities to archetypes. In this paper, for the construction of archetypal dictionaries only unlabeled data samples are used, which are collected in the same image which is meant to be classified. Assuming the test data follows the same distribution as the labeled data, in this way, the test data lies within the convex hull of the extracted archetypes. Generally, sparse representation with archetypal dictionaries is performed with an additional sum-to-one constraint. However, in recent and former experiments (e.g., [15]) we could observe that using this constraint led to a decrease in accuracy. Therefore, we only use non-negativity constraint.

The disadvantage of archetypal analysis is that the final set of T archetypes $\boldsymbol{D}$ depends on the starting point, and as a result, there is no unique solution to the final set. Especially, if the number of archetypes in the dictionary is low, various solutions lead to significantly different accuracies. To overcome this problem, we propose an optimization procedure to find the best set of archetypes from a large set of preselected ones, called the initial set. In more detail, our task is to find the set of archetypes $\boldsymbol{D}$ which minimizes the energy

$$ U(\boldsymbol{D}) = -\log(e) + \|\boldsymbol{\gamma}\|_2 \ , \tag{4} $$

where we couple the sparse representation reconstruction error $\boldsymbol{\gamma}$ which is obtained by using the current set of archetypes as dictionary $\boldsymbol{D}$, and the logistic regression error function value e using the feature representation obtained by this sparse

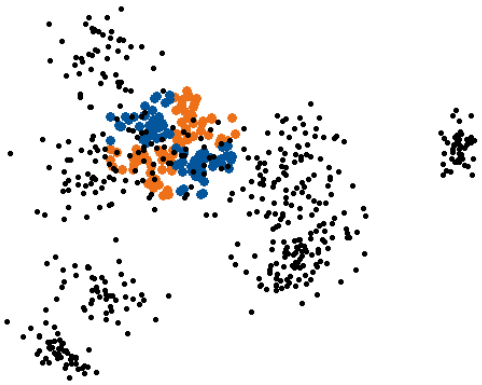


Fig. 2. Toy example data set.

Fig. 3. Map of the study area in western Ukraine.

representation. The value of e is obtained using the cross-validation error of the training data. This way we achieve a set of archetypes that fulfill both requirements, leading to a good reconstruction, which makes the resulting set interpretable, and having good discriminative properties. The energy U is a complex function with rough landscape and unknown dimensionality due to the unknown number of archetypes. Therefore, we optimize with rjMCMC coupled with simulated annealing to find the global optimum. Introducing the temperature parameter K, the optimizer is given by

$$\widehat{D} = \underset{D}{\operatorname{argmin}} \ \frac{U(D)}{K_k}, \ \lim_{k\to\infty} K_k = 0 \ . \tag{5}$$

While MCMC is dedicated to sample from probably unnormalized densities, simulated annealing allows to make a point estimate of its global optimum. Using simulated annealing we create a Markov chain, such that the samples explore the whole state space in the beginning and gradually concentrate around the global optimum of the energy function U. In this way we avoid trapping into local optima, as it is usually the case for greedy algorithms. We use the so called birth an death algorithm [16] to sample from the space of possible sets of archetypes, which turns out to be a special type of Green's rjMCMC sampler [17].

IV. Experimental Setup and Results

A. Data Sets

In our experiments we use one synthetic data set and two real world data sets to show the performance of our proposed approach.

1) Toy Example: The synthetic data set consists of non-linearly separable four corner shaped clusters of 2 classes[1] serving as training and test data, and a mixture of Gaussian distributed unlabeled data (see Fig. 2). The data set is randomly sampled in each run, where 100 runs are conducted in our experiments.

2) Ukraine Data Set: We use a Landsat 7 image acquired in November, 2010, covering a study area in western Ukraine (see Fig. 3). The thermal and panchromatic band were removed resulting in six spectral bands ranging from blue to shortwave-infrared. Training and reference data were acquired during an extensive field campaign in 2012 [18] resulting in 357 test data samples and 62808 training samples, from which we use a fraction for our experiments. We randomly sampled 10 different training sets, while keeping the test set fixed, and report the average accuracies. Both training and test set are spatially disjoint. We are aiming at five land cover classes, namely CROPLAND, PASTURE, FALLOW, FOREST and URBAN.

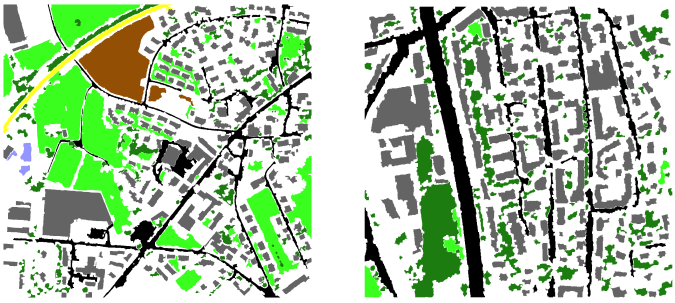


Fig. 4. Zurich data set, images # 1 and # 2.

3) Zurich Data Set: As second real world data set we use the Zurich data set [20][2], consisting of 20 multispectral VHR images acquired over the city of Zurich by the Quickbird satellite. Two of the images are illustrated in Fig. 4. The images have 4 spectral bands (R-G-B-NIR) and a spatial resolution of approximately 0.61m/pixel. The classification task aims at 8 land cover classes (ROADS, BUILDINGS, TREES, GRASS, BARE SOIL, WATER, RAILS, POOLS). For our experiments we performed a leave-one-out estimation, i.e. we trained the classifier on 19 images and tested on the remaining image. We use a subset of training samples of approximately 350 samples for each run.

[1] http://www.junuxx.net/datasets.zip

[2] https://sites.google.com/site/michelevolpiresearch/data/zurich-dataset

B. Experimental Setup

For the Ukraine and Zurich data sets, the archetypal dictionary was learned from all unlabeled data samples in the image. Since our data set contains outliers, which may be chosen as archetype, we compute the local outlier factor [19] for each point and its ten nearest neighbors and remove all samples, which value is too high. Our data is pre-processed using global contrast normalization [9]:

$$x'_b = \frac{x_b - \bar{x}_b}{\max\left(\epsilon, \sqrt{\lambda + \sigma_b}\right)}, \tag{6}$$

where x'_b is the normalized pixel, x_b is the non-normalized pixel, $\bar{x}_b$ is the mean over all pixels in image band b, σ_b is the standard deviation of all pixels in image band b and λ is a positive regularization parameter, set to $\lambda = 100$, in order to bias the standard deviation estimation. The denominator is constrained to be at least $\epsilon = 10^{-4}$.

In our parameter settings we fixed the number of non-zero elements for sparse representation to $W = 5$. For rjMCMC, we use a geometric temperature schedule, using $\alpha = 0.9999$ and a start temperature $T_0 = 0.2$. The latter was determined empirically, such that the average acceptance rate at beginning was around 70%. Please note, that the optimization procedure is guarantied to find the global optimum, as long as we cool down the system logarithmically, which is not feasible in practice. By choosing a slow geometric cooling scheme we try to find a balance between speed and stable results. The initial set of archetypes for our proposed approach is accumulated by using each training sample as initialization for SiVM, whereas finally redundant archetypes are removed. In more detail, using an arbitrary training sample as initial point, the first archetype is defined as the sample with the largest distance to the initial point. For each initialization, ten archetypes are selected. For evaluation, we compare the classification accuracy obtained by using the original features (`OriFeat`), the accuracy obtained by randomly initialized archetypes (`ATrandInit`) and the accuracy obtained by our proposed approach (`ATbest`). In all cases, we use a logistic regression classifier. For `ATrandInit` we use the same number of archetypes as estimated by our approach. We report overall accuracy, average accuracy and kappa coefficient.

C. Results

1) Toy Example: Fig. 5 shows the overall accuracy of the compared approaches. Our approach `ATbest` obtained the highest accuracy with the lowest standard deviation. The classification accuracy obtained by using the original features (`OriFeat`) yielded the worst result, since the features are not linearly separable and the global optimum of logistic regression classifier is reached if approximately all samples are assigned to one class. `ATrandInit` shows the highest standard deviation, which underlines the fact that the set of archetypes highly influences the classification accuracy. However, information from unlabeled data can help to find a better data representation.

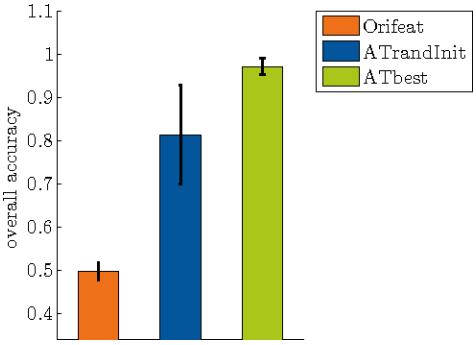
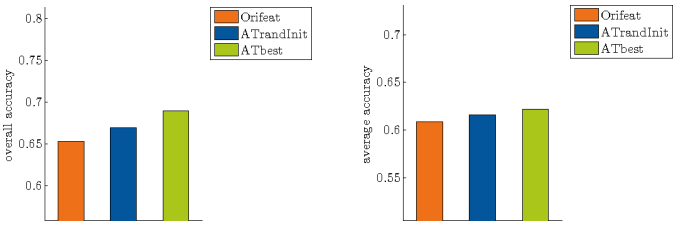


Fig. 5. Overall accuracies of toy example data set.

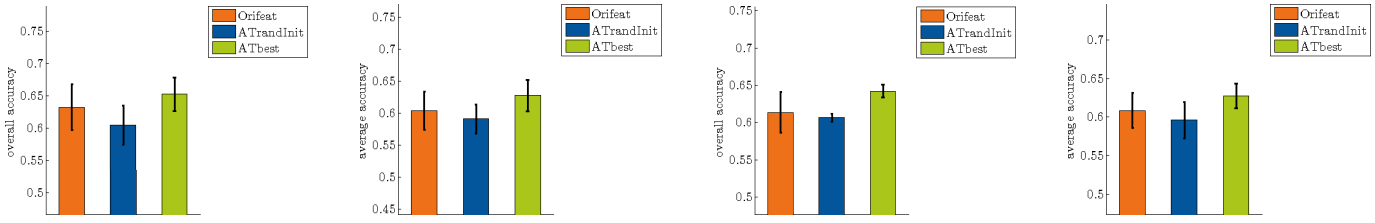
2) Ukraine Data Set: Fig. 6 shows the obtained classification results for the Ukraine data set when using 20 and 30 training samples per class, respectively. The results show that the obtained results using our approach with the new feature representation (`ATbest`) is slightly better than using logistic regression with the original feature representation (`OriFeat`). The worst result was obtained by `ATrandInit`, underlining the fact that the accuracy highly depends on the initialization. However, we observed that the training data and test data did not accurately follow the same underlying distribution, which results in a drop in accuracy. `ATbest` achieved the lowest standard deviation, especially for the overall accuracy, and thus, our proposed approach turned out to be more stable than `OriFeat` and `ATrandInit`. The average number of used dictionary elements is 10 with a standard deviation of approximately 2 elements. Although the standard deviation is low, the number of used dictionary elements obtained by our approach is in the range of 7 and 14.



(a) Overall accuracy (b) Average accuracy

Fig. 7. Accuracies for the Zurich data set.

3) Zurich Data Set: The Zurich data set underlines similar findings as the previous data set, as presented in Fig. 7. In this case, standard deviations are not reported since not all classes are present in all images and thus, the results differ vastly. In contrast to the Ukraine data set, also `ATrandInit` achieve better results than `OriFeat`. The plots show that our proposed method is able to find discriminative archetypal dictionaries by choosing suitable dictionary elements and their number. The average number of used dictionary elements is 22 with a standard deviation of approximately 6 elements, i.e. that although the images are similar the best number of used dictionary elements differ and should be determined during the optimization procedure to ensure a good classification result.



(a) Overall accuracy using 20 training samples per class

(b) Average accuracy using 20 training samples per class

(c) Overall accuracy using 30 training samples per class

(d) Average accuracy using 30 training samples per class

Fig. 6. Overall and average classification accuracy for the Ukraine data set.

V. Conclusion

We presented an approach to learn discriminative archetypal dictionaries from unlabeled data used for self-taught learning. This approach is an extension to the work presented in [4] and is able to provide a new discriminative feature representation, which can be more suitable for classification than using the original feature representation. To find the best set of archetypes we utilize reversible jump Markov Chain Monte Carlo method to jointly determine the elements and the number of elements to build the dictionary. Our results underlined a gain in accuracy, especially if the number of training samples is small. The presented approach is promising and expandable since e.g., additional structured sparsity like group sparsity priors can be introduced in order to further increase the accuracy. Our future research will focus on learning structured priors for archetypal dictionaries, which can be seen as an further extension to selected the most suitable archetypes for classification purposes. We are convinced that our findings are also useful in other research communities such as these ones focusing on unmixing.

Acknowledgment

The authors would like to thank the German Research Foundation (DFG) WA 2728/3-1 for funding. The research herein was performed in part while Ribana Roscher was working within this project. The authors would also like to thank Christoph Römer for fruitful discussions and Julian Horst, Sebastian Riedel and Kristian Kersting for provision of the SiVM software.

References

[1] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 11, pp. 3363–3373, 2006.

[2] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 759–766.

[3] A. Coates and A. Y. Ng, "Learning feature representations with k-means," in *Neural Networks: Tricks of the Trade*. Springer, 2012, pp. 561–580.

[4] R. Roscher, C. Römer, B. Waske, and L. Plümer, "Landcover classification with self-taught learning on archetypal dictionaries," in *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2015, pp. 2358–2361.

[5] M. Aharon, M. Elad, and A. Bruckstein, "K -SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[6] G. Zhao, C. Zhao, and X. Jia, "Multilayer unmixing for hyperspectral imagery with fast kernel archetypal analysis," *IEEE Geoscience and Remote Sensing Letters*, vol. PP, no. 99, pp. 1–5, 2016.

[7] M. Wahabzada, A.-K. Mahlein, C. Bauckhage, U. Steiner, E.-C. Oerke, and K. Kersting, "Plant phenotyping using probabilistic topic models: Uncovering the hyperspectral language of plants," *Scientific reports*, vol. 6, 2016.

[8] C. Römer, M. Wahabzada, A. Ballvora, F. Pinto, M. Rossini, C. Panigada, J. Behmann, J. Léon, C. Thurau, and C. Bauckhage, "Early drought stress detection in cereals: simplex volume maximisation for hyperspectral image analysis," *Functional Plant Biology*, vol. 39, no. 11, pp. 878–890, 2012.

[9] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[10] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[11] M. E. Winter, "N-findr: an algorithm for fast autonomous spectral end-member determination in hyperspectral data," in *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*. International Society for Optics and Photonics, 1999, pp. 266–275.

[12] P. J. Green, "Reversible jump markov chain monte carlo computation and bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[13] A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics*, vol. 36, pp. 338–347, 1994.

[14] C. Thurau, K. Kersting, and C. Bauckhage, "Yes we can: simplex volume maximization for descriptive web-scale matrix factorization," in *Proc. International Conference on Information and Knowledge Management*. ACM, 2010, pp. 1785–1788.

[15] R. Roscher, J. Behmann, A.-K. Mahlein, . Dupuis, H. Kuhlmann, and L. Plümer, "Detection of disease symptoms on hyperspectral 3d plant models," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, pp. 89–96, 2016.

[16] C. J. Geyer and J. Møller, "Simulation Procedures and Likelihood Inference for Spatial Point Processes," *Scandinavian Journal of Statistics*, vol. 21, no. 4, pp. pp. 359–373, 1994.

[17] P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

[18] J. Stefanski, O. Chaskovskyy, and B. Waske, "Mapping and monitoring of land use changes in post-soviet western ukraine using remote sensing data," *Applied Geography*, vol. 55, pp. 155–164, 2014.

[19] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, no. 2. ACM, 2000, pp. 93–104.

[20] M. Volpi and V. Ferrari, "Semantic segmentation of urban scenes by learning local class interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 1–9.