ALPCAH: Sample-wise Heteroscedastic PCA & Union of Subspace (UoS) Extension

Javier Salazar Cavazos, Jeffrey A. Fessler, Laura Balzano
EECS Department, University of Michigan, Ann Arbor, Michigan, United States
Email: javiersc@umich.edu, fessler@umich.edu, girasole@umich.edu

***Abstract*—Principal component analysis (PCA) is a key tool in the field of data dimensionality reduction that is useful for various data science problems. However, many applications involve heterogeneous data that varies in quality due to noise characteristics associated with different sources of the data. Methods that deal with this mixed dataset are known as heteroscedastic methods. This paper develops a PCA method that can estimate the sample-wise noise variances and use this information in the model to improve the estimate of the subspace basis associated with the low-rank structure of the data. This method is extended to the Union of Subspace (UoS) setting also known as subspace clustering which is an unsupervised machine learning problem that generally entails grouping points that lie near a union of affine subspaces. Some applications involve heterogeneous data that varies in quality due to noise characteristics associated with different sources of the data, and methods that deal with this mixed dataset are known as heteroscedastic methods. We present an ALPCAH extension based on ensemble K-Subspaces (EKSS) that incorporates the ALPCAH objective function for heteroscedastic subspace basis estimation. Simulations show the effectiveness of accounting for such heteroscedasticity in the data with comparisons against other subspace approximation and subspace clustering methods established in the literature. Code available at https://github.com/javiersc1/ALPCAH and https://github.com/javiersc1/ALPCAHUS.**

I. ALPCAH

A. Introduction

Many modern data science problems require learning an approximate subspace basis for some collection of data. For example, lesion detection [1], motion estimation [2], dynamic MRI [3], and image/video denoising [4] are practical applications involving the estimation of a subspace basis. Today, a voluminous amount of data is collected to solve problems and this data tends to have a high dimensional ambient space. However, the underlying relationships between the variables are often low dimensional so the problem becomes finding low dimensional structure in the data to achieve a certain task.

PCA methods like Robust PCA [5] and Probabilistic PCA [6] work well in the homoscedastic setting, i.e., when the data is the same quality throughout, but fail to accurately estimate the basis when the data varies in quality, i.e., in the heteroscedastic setting [7]. In this setting, the noisier data samples can wildly corrupt the basis estimate. Some examples of heteroscedastic data sets that involve subspace bases include environmental air data [8], astronomical image data [9], and biological sequencing data [10]. A natural question to ask is whether it is possible to simply discard the noisy samples to avoid this issue. This question assumes that the practitioner knows what samples are good and bad, which may be difficult to know in practice. The question also assumes that there is enough good data to estimate the basis, but it is possible that the general lack of good data requires using the noisy data if the subspace dimension is higher than the amount of good data. More optimistically, even the noisier samples can help improve the estimate of the basis if properly modeled [7], so it is preferable to use all of the data available.

Although one can consider heteroscedasticity across the feature space with methods such as HeteroPCA [11], this paper focuses on heteroscedasticity across the data samples. The weighted PCA (WPCA) [12] approach for heteroscedastic data forms a weighted sample covariance matrix and requires knowledge of the noise variances. However, data quality may not be known, e.g., unknown origin of the dataset or unavailable data sheet for physical sensors. Other heteroscedastic methods like HePPCAT [7] use factor analysis and hard rank constraints to estimate the subspace basis. Being a probabilistic PCA approach, HePPCAT makes Gaussian assumptions about the basis coefficients. Additionally, HePPCAT either assumes the subspace dimension is known or requires an estimate of the rank parameter. The proposed method in the next section allows for optional usage of rank knowledge via a unique low-rank promoting functional and makes no distributional assumptions about the low-rank component, allowing it to achieve higher accuracy than current methods even without knowing the noise variances.

B. Proposed Method

Let $y_i \in \mathbb{R}^D$ represent the data samples for index $i \in \{1, \dots, N\}$ given N total samples, and let D denote the ambient dimension. Let x_i represent the low-dimensional data sample generated by $x_i = Uz_i$ where $U \in \mathbb{R}^{D\times k}$ is an unknown subspace basis of dimension k and $z_i \in \mathbb{R}^k$ are basis coordinates. Then the heteroscedastic model is described as follows assuming Gaussian noise

$$y_i = x_i + \epsilon_i \;\text{ where }\; \epsilon_i \sim \mathcal{N}(0, \nu_i I) \tag{1}$$

for noise variances ν_i. Note that we are considering the general case where each data point has its own noise variance since it is more challenging to tackle. However, one can consider groups of data $\{\nu_1, \dots, \nu_L\}$ where L represents the number of groups and each data point belongs to one of the L groups.

Supported in part by NSF CAREER Grant 1845076.

For the measurement model $y_i \sim \mathcal{N}(x_i, \nu_i I)$, the probability density function for a single point is

$$\frac{1}{\sqrt{(2\pi)^k |\nu_i I|}} \exp\left[-\frac{1}{2}(y_i - x_i)^T (\nu_i I)^{-1}(y_i - x_i)\right]. \quad (2)$$

For uncorrelated samples, the joint log likelihood of all y_i is the following after dropping constants

$$\sum_{i=1}^{N} -\frac{1}{2}\log|\nu_i I| - \frac{1}{2}(y_i - x_i)^T(\nu_i I)^{-1}(y_i - x_i). \quad (3)$$

Let $\Pi = \mathrm{diag}(\nu_1, \ldots, \nu_N) \in \mathbb{R}^{N\times N}$ be a diagonal matrix representing the (typically unknown) noise variances. Let $Y = [y_1, \ldots, y_N] \in \mathbb{R}^{D\times N}$ represent all of the data samples. Then, the log likelihood in matrix form is

$$-\frac{D}{2}\log|\Pi| - \frac{1}{2}\mathrm{Trace}[(Y - X)^T \Pi^{-1}(Y - X)]. \quad (4)$$

Using trace properties, the optimization problem we pose for the heteroscedastic model is

$$\arg\min_{X,\Pi} \lambda f_k(X) + \frac{1}{2}\|(Y - X)\Pi^{-1/2}\|_{\mathrm{F}}^2 + \frac{D}{2}\log \underbrace{|\Pi|}_{\text{determinant}}, \quad (5)$$

where $f_k(X)$ is a relatively new functional in the literature [13] that promotes low-rank structure in X and $\lambda \in \mathbb{R}_+$ is a regularization parameter. Our algorithm for solving (5) is called ALPCAH (**A**lgorithm for **L**ow-rank regularized **PCA** for **H**eteroscedastic data). Since X represents the denoised data matrix, the subspace basis is calculated by performing an SVD on the optimal solution from (5) so that $\hat{X} = \sum_i \hat{\sigma}_i \hat{u}_i \hat{v}_i{}'$ and thus $\hat{U} = [\hat{u}_1, \ldots, \hat{u}_k]$. The low-rank promoting functional we use is the summation of the tail singular values defined as the following

$$f_k(X) \triangleq \sum_{i=k+1}^{\min(D,N)} \sigma_i(X) = \|X\|_* - \|X\|_{\mathrm{Ky-Fan}(k)} \quad (6)$$

where $\sigma_i(X)$ is the ith singular value of X, $\|\cdot\|_*$ is the nuclear norm, and $\|\cdot\|_{\mathrm{Ky-Fan}(k)}$ is the Ky-Fan norm defined as the sum of the first k singular values. For $k = 0$, $f_0(X) = \|X\|_*$. For a general $k > 0$, $f_k(X)$ is a nonconvex difference of convex functions. When $k > 0$ and $\lambda \to \infty$, then the solution of the optimization problem approaches $\hat{X} = \sum_{i=1}^{k} \sigma_i u_i v_i' \in \mathbb{R}^{D\times k}$ meaning the solution becomes identical to a singular value projection approach.

C. Algorithm & Convergence Analysis

We apply the inexact augmented Lagrangian method ADMM [14] to the cost function (5). Introducing the auxiliary variable $Z = Y - X$, the augmented penalty parameter $\mu \in \mathbb{R}$, and dual variable $\Lambda \in \mathbb{R}^{D\times N}$, the augmented Lagrangian, as defined in [15], is

$$\begin{aligned}\mathcal{L}_\mu(X, Z, \Lambda, \Pi) = \lambda_r f_k(X) + \frac{1}{2}\|Z\Pi^{-1/2}\|_{\mathrm{F}}^2 + \frac{D}{2}\log|\Pi| \\ + \langle \Lambda, Y - X - Z\rangle + \frac{\mu}{2}\|Y - X - Z\|_{\mathrm{F}}^2. \quad (7)\end{aligned}$$

Definition 1. *Let $A \in \mathbb{R}^{D\times N}$ be a rank k matrix such that its decomposition is $\mathrm{SVD}(A) = U_A D_A V_A'$ where $D_A = \mathrm{diag}(\sigma_1(A), \ldots, \sigma_{\min(D,N)}(A))$. Let the soft thresholding operation be defined as $\mathcal{S}_\tau[x] = \mathrm{sign}(x)\max(|x| - \tau, 0)$ for some threshold $\tau > 0$. Decompose D_A such that $D_A = D_{A1} + D_{A2} = \mathrm{diag}(\sigma_1(A), \ldots, \sigma_k(A), 0, \ldots, 0) + \mathrm{diag}(0, \ldots, 0, \sigma_{k+1}(A), \ldots, \sigma_N(A))$. Then, the proximal mapping solution for $f_k(X)$, as shown in [13], is denoted as the tail singular value thresholding operation and expressed as*

$$\mathrm{TSVT}(A, \tau, k) \triangleq U_A \left(D_{A1} + \mathcal{S}_\tau[D_{A2}]\right) V_A'. \quad (8)$$

Performing a block Gauss-Seidel pass for each variable results in the following closed-form updates

$$\begin{aligned} Z_{i+1} &= \arg\min_{Z_i} \mathcal{L}_\mu(X_i, Z_i, \Lambda_i, \Pi_i) \\ &= [\mu(Y - X_i) + \Lambda_i](\Pi_i^{-1} + \mu I)^{-1} \quad (9) \\ X_{i+1} &= \arg\min_{X_i} \mathcal{L}_\mu(X_i, Z_i, \Lambda_i, \Pi_i) \\ &= \mathrm{TSVT}(Y - Z_i + \frac{1}{\mu}\Lambda_i, \frac{\lambda_r}{\mu}, k) \quad (10) \\ \Lambda_{i+1} &= \Lambda_i + \mu(Y - X_i - Z_i). \quad (11) \end{aligned}$$

When each point is treated as having its own noise variance, then the variance update is

$$\Pi_{i+1} = \arg\min_{\Pi_i} \mathcal{L}_\mu(X_i, Z_i, \Lambda_i, \Pi_i) = \frac{1}{D} Z_i^T Z_i \odot I. \quad (12)$$

For the case when the data points have grouped noise variances, let $l \in \{1, \ldots, L\}$ and let n_l signify the number of points in group l out of L total groups; then the grouped noise variance update instead becomes

$$\nu_l = \frac{1}{Dn_l}\|Z^{(p_l)}\|_{\mathrm{F}}^2 = \frac{1}{Dn_l}\|Y^{(p_l)} - X^{(p_l)}\|_{\mathrm{F}}^2 \quad (13)$$

where p_l signifies the points associated with group l, meaning that $Y^{(p_l)} \subset Y$. Consider the cost function for the case when the variances are known. The formulation consists of a two-block setup written as

$$\arg\min_{X,Z} \underbrace{\lambda_r f_k(X)}_{f(X)} + \underbrace{\frac{1}{2}\|Z\Pi^{-1/2}\|_{\mathrm{F}}^2}_{g(Z)} \text{ s.t. } Y = X + Z. \quad (14)$$

Theorem 1. *Let $\Psi(X, Z) = f(X) + g(Z)$. Let $\nu_i \geq \epsilon > 0 \quad \forall i$. Assuming that μ in* (7) *satisfies $\mu > 2L_g = 2\|\Pi^{-1}\|_2$, the sequence generated by* (9), (10), (11) *converges to a KKT (Karush–Kuhn–Tucker) point of the augmented Lagrangian $\mathcal{L}_\mu(X, Z, \Lambda)$.*

Proof. ADMM convergence for nonconvex problems has been studied by [16] for two-block setups. The functional $f(X)$ is a proper, lower semi-continuous function since it is a sum of continuous functions. The function $g(Z)$ is a continuous differentiable function whose gradient is Lipschitz continuous with modulus of continuity $L_g = \|\Pi^{-1}\|_2$. Since $g(Z) = \nu_1^{-1/2} Z_{1,1} + \nu_1^{-1/2} Z_{2,1} + \ldots$ is a polynomial equation, then its graph is a semi-algebraic set.

To the best of our knowledge, there is no literature that explores semi-algebraic properties of nuclear norm based functions and so the following results are our own. Let $f_k(X) = h(X) - q(X) = \|X\|_* - \|X\|_{\mathrm{Ky-Fan}(k)}$. Let $X \in \mathbb{R}^{M\times N}$ such that $G = X'X \in \mathbb{R}^{N\times N}$. Then, by Cayley Hamilton theorem, the characteristic polynomial is expressed as $p_G(\lambda) = \lambda^n + c_{n-1}(G)\lambda^{n-1} + \ldots + c_1(G)\lambda + c_0$ for constants $c_i \in \mathbb{R}$. Let λ be eigenvalues of G which implies $p_G(\lambda) = 0$. Then, the set $\mathcal{S}_G = \{\forall\lambda \mid p_G(\lambda) = 0\}$ is semi-algebraic since it is defined by polynomial equations. Note that $\lambda_i = \sigma_i^2$ since G is the gram matrix of X. The set $\mathcal{S}_X = \{\forall\sigma \mid \sigma^2 = \lambda \in S_G, \sigma \geq 0\} = \{(\sigma_1, \ldots, \sigma_n)\}$ is semi-algebraic as it is expressed in terms of polynomial inequalities. Expressing $h(X) = \|X\|_* = h(\sigma_1, \ldots, \sigma_n)$, its graph $h = \{(\sigma, f(\sigma))\}$ is semi-algebraic and thus so is the nuclear norm. By Tarksi-Seidenburg theorem [17], defining the projection map $\Phi : \mathbb{R}^n \to \mathbb{R}^k$, the set $\Phi(\mathcal{S}_X) = \{(\sigma_1, \ldots, \sigma_k)\}$ is semi-algebraic and thus so is $q(X) = \|X\|_{\mathrm{Ky-Fan}(k)}$. A finite weighted sum of semi-algebraic functions is known to be semi-algebraic [18] and so $f(X) = h(X) - q(X)$ is semi-algebraic. Since the functions $f(X)$ and $g(Z)$ are lower, semi-continuous and definable on an o-minimal structure (such as semi-algebraic or sub-analytic as an example) [19] then it follows that $\Psi(X, Z) = f(X) + g(Z)$ is a Kurdyka-Łojasiewicz function [18] which is sufficient to proving a bounded sequence. Then the sequence $\{(X_i, Z_i)\}_{i\in\mathbb{N}}$ converges to a KKT point by applying Theorem 3.1 from [16]. □

D. Results

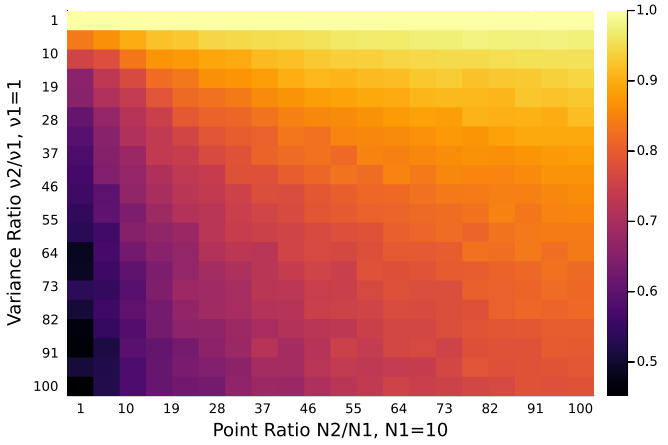
This section uses a synthetic dataset to compare ALPCAH with PCA, RPCA, and HePPCAT. We consider two groups of data, one with fixed quality (i.e., fixed size and fixed additive noise variance) and one whose parameters we vary. Let $y_i \in \mathbb{R}^{100\times N}$ where N, the total number of points, changes depending on parameter values. Let $U \in \mathbb{R}^{100\times 10}$ represent a 10 dimensional subspace generated by random uniform matrices such that $U\Sigma V^T = \mathrm{svd}(A)$, where $A_{i,j} \sim \mathcal{U}[0, 1]$. The low-rank data x_i we simulated as $x_i = Uz_i$ where the coordinates $z_i \in \mathbb{R}^{10}$ were generated from $\mathcal{U}[-100, 100]$ for each element in the vector. Then, we generated $y_i = Uz_i + \epsilon_i$ where $\epsilon_i \in \mathbb{R}^{100}$ is drawn from $\mathcal{N}(0, \nu_i I)$. The noise variance for group 1 (ν_1) was fixed to 1 and we varied group 2 noise variances (ν_2). The error metric used is subspace affinity error that compares the difference in projection matrices $\|UU' - \hat{U}\hat{U}'\|_{\mathrm{F}}/\|UU'\|_{\mathrm{F}}$ so that a low error signifies a closer estimate of the true subspace. In summary, the noisy data $Y = [y_1, \ldots, y_N]$ is generated accordingly, a solution $\hat{X}$ is generated from (5), the subspace basis is calculated by $\hat{X} = \sum_i \hat{\sigma}_i\hat{u}_i\hat{v}_i{}' \implies \hat{U} = [\hat{u_1}, \ldots, \hat{u_k}]$, and the subspace affinity error is reported.

For the heatmaps in Fig. 1a-Fig. 1d, each pixel represents a ratio of ALPCAH error divided by the error of some other method like HePPCAT. A value close to 1 implies ALPCAH did not perform much better relative to the other method, whereas a ratio closer to 0 implies ALPCAH performed relatively well. For the heatmaps in Fig. 1a-Fig. 1d, the x-axis represents the point ratio between group 1 and 2, where group 1 always has 10 points. The y-axis represents the variance ratio between group 1 and 2, where the group 1 noise variance was fixed to 1. The average of 25 trials is plotted at each pixel in the heatmap where each trial has different noise, basis coefficients, and subspace basis realizations. To summarize, we explored the effects of data quality and data quantity on the heteroscedastic subspace basis estimates in different situations.

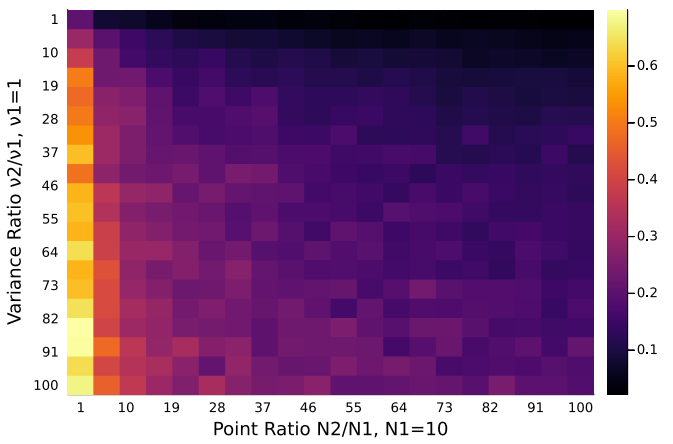
Fig. 1a and Fig. 1b compare ALPCAH against PCA in the simpler situation where noise variances are known. Fig. 1b is similar to Fig. 1a but only using the high quality points for PCA specifically, whereas ALPCAH used all of the data. This result confirms that it is useful to use very noisy data, as opposed to throwing it away and treating the remaining data as homoscedastic. Fig. 1c and Fig. 1d compare ALPCAH against two other PCA methods in the unknown variance setting. Fig. 1c compares against Robust PCA to see if an "outlier" model can capture heteroscedastic noise. Fig. 1d compares against HePPCAT.

Since Fig. 1a-Fig. 1d only show relative error, it is important to discuss absolute error of these algorithms. For Fig. 1e-Fig. 1f, we fixed the total number of points to be 500 with just enough high quality samples that have noise variance $\nu_1 = 0.25$ and the rest of the points have noise variance $\nu_2 = 100$. The regularization parameter λ is varied both when rank knowledge is known or estimated (for these results $k = 10$) and when rank knowledge is not utilized ($k = 0 \implies f_k(X) = \|\cdot\|_*$). The y-axis consists of the subspace affinity error function as shown before. The average error is plotted out of 25 trials with maximum error bounds for each λ value. Fig. 1f considers the unknown variance case but shows WPCA (a known variance method) as a bottom floor to illustrate the lowest realistic affinity error if one knew the noise variances. Because the unknown variance setting requires a tuning parameter λ, we performed cross validation to determine a different λ value for each heatmap pixel location to generate Fig. 1c and Fig. 1d. The robustness of λ for different point and variance ratios is mentioned in the discussion section. In practice, we found that one λ value works well across the entire heatmap. Experimentally, we found that a sufficiently large $\lambda \geq \|Y\|_2$ gave the lowest subspace affinity error in the known variance setting so cross-validation is not performed in the known variance setting for these experiments.

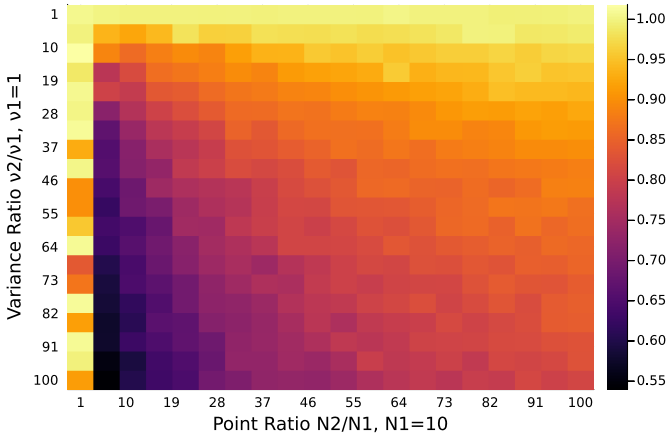
Note that the subspace basis dimension was used for these results in Fig. 1a-Fig. 1d by setting $k = 10$ for $f_k(X)$ in ALPCAH to compare against HePPCAT but one may use $k = 0$ when the subspace dimension is not known. For many applications, the subspace dimension is unknown such as non-Lambertian surfaces under non-isotropic lighting conditions [20]. For this situation, there are rank estimation methods proven to be robust in this heteroscedastic noise setting, such as randomly flipping signs in the data matrix [21]. Thus, it is possible to approximate the rank beforehand given a reasonably sized data matrix such that SVD methods are feasible.



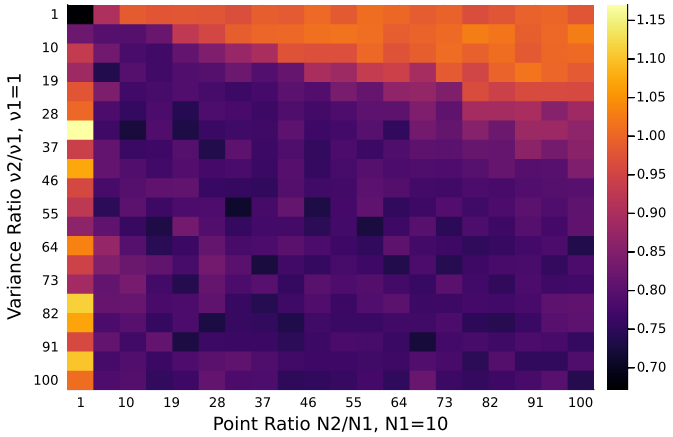
(a) Ratio of subspace affinity errors ALPCAH/PCA (known variance, no cross-validation required)



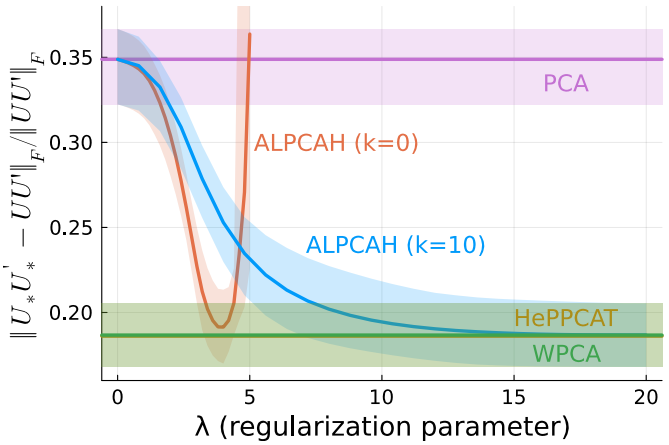
(b) Ratio of subspace affinity errors ALPCAH/PCA-GOOD (PCA using good data only, no cross-validation required)



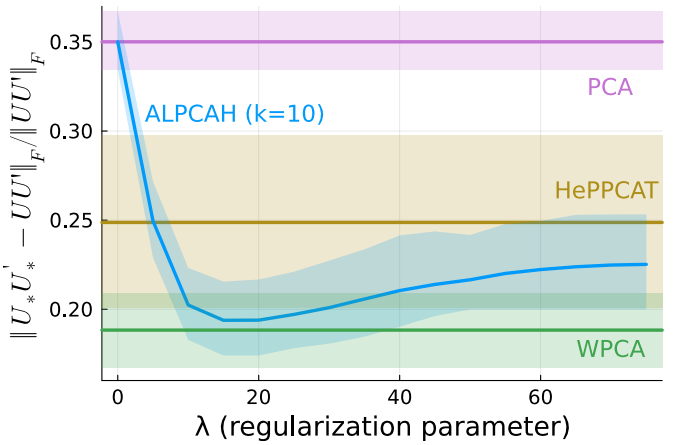
(c) Ratio of subspace affinity errors ALPCAH/RPCA (unknown variance, no group knowledge, cross-validated λ for both methods)



(d) Ratio of subspace affinity errors ALPCAH/HePPCAT (unknown variance, no group knowledge, cross-validated λ for ALPCAH)



(e) Subspace affinity error of various PCA methods as the regularization parameter is adjusted (known variance)



(f) Subspace affinity error of various PCA methods as the regularization parameter is adjusted (unknown variance, no group knowledge)

Fig. 1: ALPCAH experiments

E. Discussion

In the known variance cases, Fig. 1a shows that ALPCAH performs well relative to PCA in noisy situations and can improve estimation by up to 50% or 20% in more tame situations. From the bottom left corner and moving rightwards, there is a general decline as the estimation error worsened as the number of "bad" points increased. This means that the noisy points contributed too much to the estimation process when the good quality data should have more influence in the process. ALPCAH uses the term $\|Z\Pi^{-1/2}\|_{\mathrm{F}}^2$, but a user may well use something like $\|Z\Pi^{-1}\|_{\mathrm{F}}^2$ to further downplay the contribution of the noisy points. Finding the optimal weighing scheme for this method is a topic of future work, but using inverse standard deviations is a natural choice that arises from the Gaussian likelihood. Some work has been done in this area for the case when noise variances are known [22]. In Fig. 1b, one can see that even in a limited data situation with very noisy data (bottom left corner), there is a 30% improvement relative to applying PCA to just the good data alone. The improvement only increased as more noisy points were added. Thus it is beneficial to collect and use all of the data, since the noisy points offer meaningful information that can improve the estimate of the basis versus using good data alone.

For the unknown variance cases, we considered the situation where group information is not known in the sense that each data point is treated as having its own noise variance as opposed to belonging to known groups $\{1, 2\}$. This groupless situation is more challenging than the grouped case. Because of this, it is useful for comparison purposes in this unknown variance case. Since Robust PCA shares similarities with ALPCAH, Fig. 1c compared these two methods. As illustrated, using cross-validation to learn λ for both methods, ALPCAH was able to outperform RPCA in all situations shown in the heatmap. Thus it appears to be preferable to treat extremely noisy points with a noise model $\|\cdot\|_{\mathrm{F}}$ rather than treating them as outliers with a $\|\cdot\|_{1,1}$ regularizer. In Fig. 1d, the comparison with HePPCAT, there is one location (bright yellow) where ALPCAH gave a worse estimation of the subspace basis, but generally on average there was a 20% improvement over HePPCAT. Since HePPCAT is a hard rank constraint method, it seems beneficial to not completely shrink the $k+10$ singular values but rather to retain them as they seem to improve the estimation process. Moreover, since we make no distributional assumptions about X itself besides low-rank assumptions, then this assumption relaxation helps us achieve lower error in settings where the basis coordinates are not Gaussian, whereas HePPCAT makes Gaussian assumptions about the basis coordinates themselves.

For both Fig. 1e and Fig. 1f, Robust PCA is excluded since this method did not perform any better than PCA for this specific test setup of fixed noise variances and point ratio. In Fig. 1e, for ALPCAH ($k = 0$), the results became worse than PCA once $\lambda > 5$ so the y-axis range is fixed for better visibility. For this case, it is interesting that there is a certain λ range where ALPCAH ($k = 0$) performs similarly to HePPCAT. Recall that when $k = 0$ in the known variance setting, the optimization problem is convex. When rank knowledge is utilized, ALPCAH ($k = 10$) subspace affinity error approaches the error of the other methods as λ grows and stays fixed at that location for larger λ values. In Fig. 1f, ALPCAH ($k = 0$) performs poorly in the unknown variance case and is excluded just like Robust PCA. The results shown are in the harder groupless setting where noise variance group knowledge is not known. We observed that ALPCAH ($k = 10$) performs better than HePPCAT for a certain λ range and gets close to WPCA (known variance method) as if we really did learn the noise variances of the points.

As a final comment, the regularization parameter appears to be fairly robust against this landscape of different variance ratios and point ratios shown in Fig. 1a-Fig. 1d. Recall that each heatmap pixel involves a different random subspace basis with data points that have different noise and basis coordinate realizations. The robustness of the regularization parameter means that it will be easier to find a suitable value since, for example, the user does not need to worry about data split size between validation and test set or whether the variance ratio in that specific validation set will not generalize to the test set. Fig. 2 shows this result.

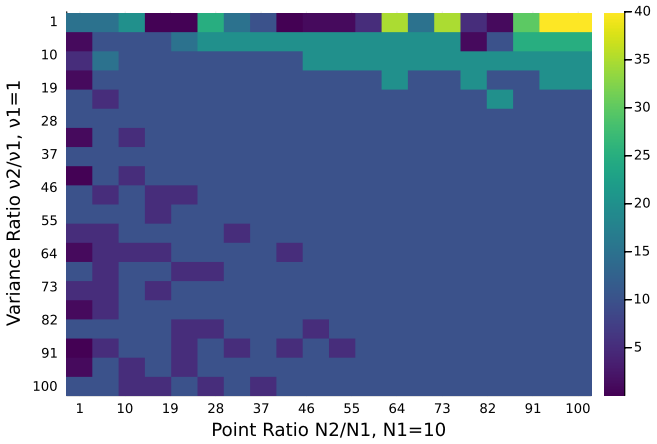


Fig. 2: ALPCAH cross validation matrix of optimized λ values

F. Conclusion

Heteroscedastic data can exist when using mixed data sets that stem from different sources to give an example. Current methods to deal with subspace models in this setting have limitations such as requiring the noise variances to be known or assuming Gaussian basis coefficients. Both of these assumptions may not be good assumptions in practice due to unknown data set group knowledge or data distribution knowledge. This work proposed a PCA method named ALPCAH that can estimate the noise variances of the collected data and use these estimates in the optimization model to not only denoise the data, but also improve the estimate of the subspace basis. ALPCAH avoids the limitations stated above and led to higher accuracy in the subspace basis estimate as shown in the results section.

II. ALPCAHUS

A. Introduction & Related Work

Subspace clustering is an unsupervised machine learning algorithm where the goal is to both cluster unlabeled data and find the underlying affine subspace associated with each cluster. If one knew one or the other, it would be relatively simple to extract the other. This chicken and egg problem becomes nontrivial when both components must be extracted. Such problem has attracted much interest in the pattern recognition and computer vision fields [23]. There are numerous applications for this problem such as image segmentation [24], motion segmentation [25], face clustering [26], image compression [27], and system identification [28]. Many algorithms fall into a general umbrella of categories such as algebraic methods [29], iterative methods [30], statistical methods [31], and spectral clustering methods [32] [33].

Lately, spectral clustering-based methods have become popular that take advantage of the self-expressiveness of data to learn linear coefficients and adopt different priors as seen with SSC [34]. These SSC-based algorithms, e.g. SSC, LLR, TRR, learn the coefficient matrix $C \in \mathbb{R}^{N\times N}$ by solving the following general optimization problem

$$\min_{C} f(Y - YC) + \lambda \mathcal{R}(C) \;\text{ s.t. }\; C \in S_C$$

where $Y \in \mathbb{R}^{D\times N}$ is the data matrix consisting of N samples and ambient dimension D. Then, spectral clustering is performed on C by performing k-means on the eigen embedding of the Laplacian matrix of C. This approach has been extended for neural networks such as Deep Subspace Clustering (DSC) [35].

Iterative methods have also recently been gaining attention that are based on alternating between cluster assignments and subspace basis approximation, i.e., the K-subspaces or Median k-flats algorithm [36]. Let $\mathcal{C}$,$\mathcal{U}$ denote the sets of estimated clusters and orthonormal subspace bases respectively such that the K-Subspaces (KSS) algorithm seeks to solve the following optimization problem

$$\min_{\mathcal{C},\mathcal{U}} \sum_{k=1}^{K} \sum_{i:\; i\in c_k} \|y_i - U_k U_k^T y_i\|_2^2. \tag{15}$$

The goal is to minimize the sum of residuals of points to their assigned subspace by alternating between performing PCA on each cluster to update $\mathcal{U}$ and using it to calculate new clusters $\mathcal{C}$ in a similar way to k-means algorithm. The quality of the solution is highly dependent on initialization for these methods. Recent work has been done on convergence guarantees and spectral initialization schemes that provably perform better than random initialization [37]. However, even with a good initialization scheme, this NP-hard problem is prone to local minima in the optimization landscape. To overcome this, consensus clustering [38] is one such tool to leverage information from many trails and then combine the results together in an ensemble process. This ensemble approach, known as Ensemble KSS (EKSS) [39], has been implemented by creating a co-association matrix whose (i, j)th entry represents the number of times the two points were clustered together. Then, spectral clustering is performed on the co-association matrix to get the final clustering from the many base clusterings.

For both KSS and EKSS, these methods implicitly assume that the data is the same quality throughout, i.e., homoscedastic data. However, with big data and crowd-sourced data, there are some data sets such as environmental air data [8], astronomical image data [9], and biological sequencing data [10] where the data varies in quality for each sample. This data is said to be heteroscedastic data. This heteroscedasticity is difficult to deal with as it may not be possible to separate good data from bad data. Moreover, it may be necessary to use lower quality data to approximate each orthonormal basis depending on subspace dimension. Thus, designing algorithms that can deal with mixed quality data is advantageous for clustering quality as it has been shown that heteroscedasticity negatively impacts results for the subspace clustering problem [40]. In the following section, we propose an algorithm that generalizes Ensemble KSS for mixed quality data sets by modeling the heteroscedastic noise in the cost function.

B. Problem Formulation & ALPCAHUS

Heteroscedastic subspace approximation, specifically when the noise variances are unknown, has been tackled by recent work [7] and [41]. Taking inspiration from [41], we modify ALPCAH (Algorithm for Low-rank PCA for Heteroscedastic data) by taking advantage of the matrix factorization literature. This is done to create a memory efficient version of ALPCAH and to avoid doing cross validation to find an ideal λ as done in [41]. Given a data matrix $Y \in \mathbb{R}^{D\times N} = \{y_1, \ldots, y_N\}$ consisting of points, assume the data model $y_i = \underbrace{Uz_i}_{x_i} + \epsilon_i$ where $z_i \in \mathbb{R}^d$ are the basis coefficients associated to the subspace $U \in \mathbb{R}^{D\times d}$. The notation $x_i \in \mathbb{R}^D = Uz_i$ means x_i is the low-rank component of the data point y_i. These points are collected into matrix $X \in \mathbb{R}^{D\times N} = \{x_1, \ldots, x_N\}$. Finally, Gaussian noise is added such that $y_i = x_i + \epsilon_i$ where $\epsilon_i \in \mathbb{R}^D \sim \mathcal{N}(0, v_i I)$. Meaning each point has its own noise variance and thus different quality in the heteroscedastic setting. We will consider each point y_i having its own noise variance ν_i in the general case but one may consider a collection of L noise variance groups where each point belongs to one of $\{\nu_1, \ldots, \nu_L\}$ instead. Assuming knowledge of subspace dimension d, let $X \in \mathbb{R}^{D\times N} = LR'$ where $L, R \in \mathbb{R}^{D\times d}$. Then, the low-memory implementation of ALPCAH we pose is the following

$$\min_{L,R,\Pi} \frac{1}{2}\|(Y - LR')\Pi^{-1/2}\|_F^2 + \frac{D}{2}\log \underbrace{|\Pi|}_{\text{determinant}} \tag{16}$$

where $\Pi = \text{Diagonal}(\nu_1, \ldots, \nu_N)$ is a diagonal matrix consisting of the noise variances for all points.

Generalizing this cost function to the Union of Subspace (UoS) setting, let the set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ represent a collection of points where each $y_i \in \mathbb{R}^D$. Let $Y \in \mathbb{R}^{D\times N}$

represent a matrix whose columns consists of these points and $Y_c \in \mathbb{R}^{D\times|c|} \subset Y$ represent a submatrix whose columns consist of points that $c \subset \mathcal{Y}$. The set $\mathcal{Y}$ is drawn from a union of K subspaces denoted as $\mathcal{U} = \{U_1, \dots, U_K\}$ where each subspace $U_i \in \mathbb{R}^{D\times d_i}$ has dimension d_i. For any point y_i, $\exists\alpha \in \{1, \dots, K\}$ such that $y_i = U_\alpha z_i + \epsilon_i$ and ideally ϵ_i is small. Let $z_i \in \mathbb{R}^{d_\alpha}$ represent the basis coefficients associated with the low-rank structure of the data and $\epsilon_i \in \mathbb{R}^D$ represents noise for that point drawn from $\epsilon_i \sim \mathcal{N}(0, \nu_i I)$. We will consider each point y_i having its own noise variance ν_i in the general case but one may consider a collection of L noise variance groups where each point belongs to one of $\{\nu_1, \dots, \nu_L\}$ instead. This means that each point may have different quality hence the data is heteroscedastic. The goal then is to identify all of the subspaces $U_1, \dots, U_K$ associated with the data and cluster assignments $\mathcal{C} = \{c_1, \dots, c_K\}$ that describe which points belong to what cluster. The optimization problem we pose is the following

$$\min_{\mathcal{C},\Pi,\mathcal{L},\mathcal{R},\mu} \sum_{k=1}^{K} \frac{1}{2}\|(Y_{c_k} - \mu_k 1' - L_k R_k')\Pi_k^{-1/2}\|_F^2 + \frac{D}{2}\log|\Pi_k| \tag{17}$$

where $\mathcal{C}, \Pi, \mathcal{L}, \mathcal{R}, \mu$ denotes the sets of estimated clusters, noise variances, factorized matrices, and means respectively. In this problem, we consider general *affine* subspaces meaning one must estimate the cluster sample mean to calculate the *linear* subspace associated with the data. Since we leverage consensus clustering over many trials to improve results, the algorithm presented in the next section is the general algorithm but one could select the base clusterings parameter $B = 1$ to reduce to one trial of the optimization problem 17.

C. Algorithm & Theoretical Results

1) Convergence analysis (K=1): For simplicity, assume a *linear* subspace model when there are $K = 1$ subspaces. The optimization problem reduces to the following form

$$\min_{L,R,\Pi} \Psi(L, R, \Pi) := \tag{18}$$

$$\underbrace{\frac{1}{2}\|(Y - LR')\Pi^{-1/2}\|_F^2}_{H(L,R,\Pi)} + \underbrace{\frac{D}{2}\log|\Pi|}_{f(\Pi)} + \underbrace{0}_{g(L)} + \underbrace{0}_{h(R)} \tag{19}$$

Observe that f, g, h are all proper, lower semi-continuous functions and H is continuously differentiable thus satisfying assumption 1 of [42]. Trivially, $\inf g > -\infty$ and $\inf h > -\infty$. By ensuring the noise variances do not go below an ϵ threshold, $\inf f > -\infty$. The partial gradients of $H(L, R, \Pi)$ with respect to each variable is globally Lipschitz with modulus of continuity $L_L = \|R\|_2^2\|\Pi^{-1}\|_2, L_R = \|L\|_2^2, L_\Pi = \frac{1}{\epsilon^3}\|Y - LR'\|_{\infty,2} + \frac{D}{2\epsilon^2}$. There exists constants λ_i^-, λ_i^+ for each variable that is independent of all variables that bounds the modulus of continuity on the upper and lower end. Since H is C^2, it follows from Mean Value Theorem that such constants exist. Thus, assumption 2 of [42] is satisfied. From this, the sequence (L, R, Π) generated by the alternating minimization algorithm (PALM) converges to a critical point of Ψ.

2) Convergence analysis (K¿1): Text here

3) Initialization scheme for ALPCAHUS (B=1): For the non-ensemble version of ALPCAHUS, it remains to be seen whether there exits an initialization scheme that performs better in expectation than random cluster assignment. In [37], the authors propose thresholding inner-product based spectral method (TIPS) for KSS that in summary, given a thresholding parameter $\tau > 0$, an adjacency matrix is generated by

$$A_{ij} = 1 \text{ if } |\langle y_i, y_j\rangle| \geq \tau \text{ and } i \neq j$$

and zero otherwise. Then, the cluster assignment is calculated by applying k-means to the eigenvector matrix associated with A. Comparing this metric $|\langle y_i, y_j\rangle|$ against others such as $\|y_i - y_j\|_2^2$, it can be seen that in expectation,

$$\mathbb{E}[\|y_i - y_j\|_2^2] = \|x_i - x_j\|_2^2 + D(\nu_i + \nu_j). \tag{20}$$

This term is inflated by the noise variances associated with each point and so is not robust towards heteroscedastic noise. However, the metric used in TIPS, in expectation

$$\mathbb{E}[|\langle y_i, y_j\rangle|] \leq \mathbb{E}[\langle x_i, x_j\rangle] \tag{21}$$

is independent of the noise variances. Because of this, this metric is more robust against heteroscedastic noise than others meaning the TIPS scheme is amenable to the heteroscedastic setting.

D. Experimental Results

Text here describing experimental results.

E. Conclusion

Text here

III. Funding Disclosure

This work is supported in part by NSF CAREER Grant 1845076.

References

[1] Y. Fu, W. Wang, and C. Wang, "Image change detection method based on rpca and low-rank decomposition," in *2016 35th Chinese Control Conference (CCC)*. IEEE, 2016, pp. 9412–9417.

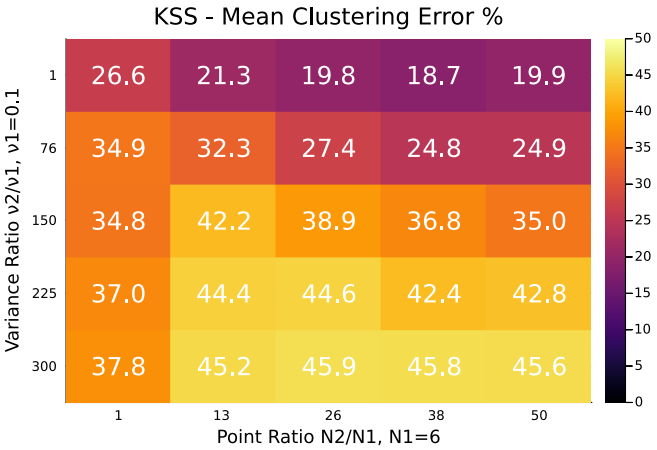
[2] Z. Zhou, X. Li, J. Wright, E. Candes, and Y. Ma, "Stable principal component pursuit," in *2010 IEEE international symposium on information theory*. IEEE, 2010, pp. 1518–1522.

[3] R. Otazo, E. Candès, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components," *Magnetic Resonance in Medicine*, vol. 73, no. 3, pp. 1125–1136, 2015.

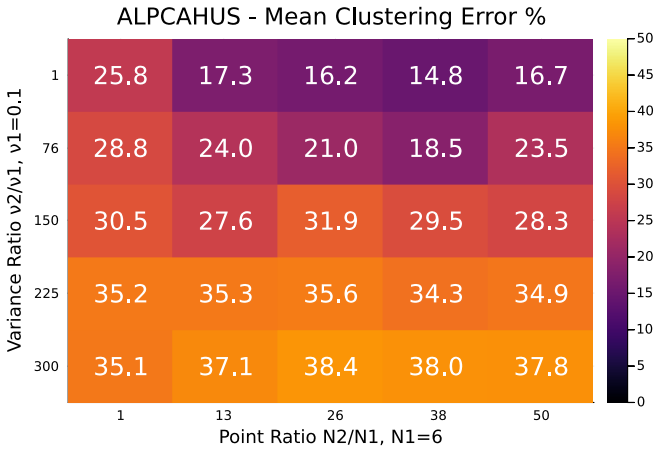
[4] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust pca, robust subspace tracking, and robust subspace recovery," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 32–55, 2018.

[5] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 1–37, 2011.

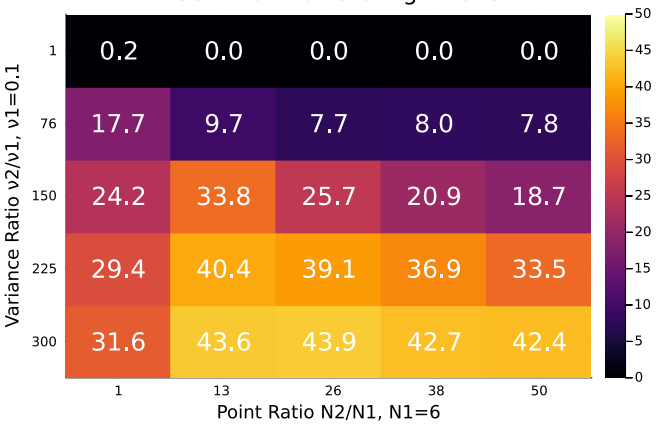
[6] M. E. Tipping and C. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society, Series B*, vol. 21, no. 3, pp. 611–622, January 1999, available from http://www.ncrg.aston.ac.uk/Papers/index.html. [Online]. Available: https://www.microsoft.com/en-us/research/publication/probabilistic-principal-component-analysis/



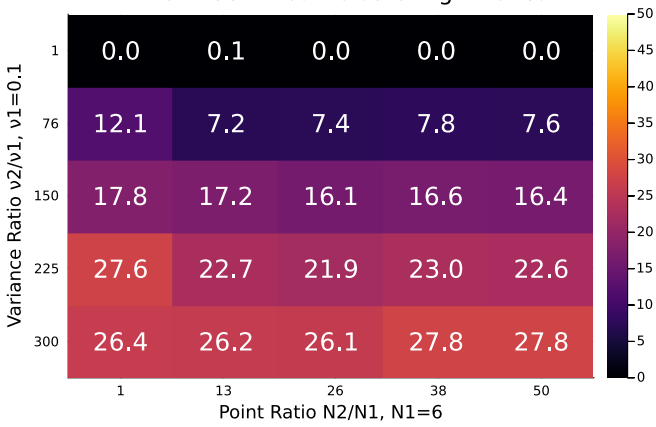
(a) Second subfigure.



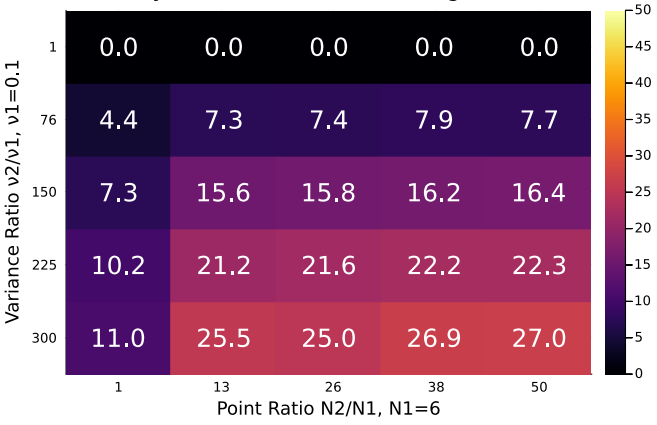
(b) Firts subfigure.



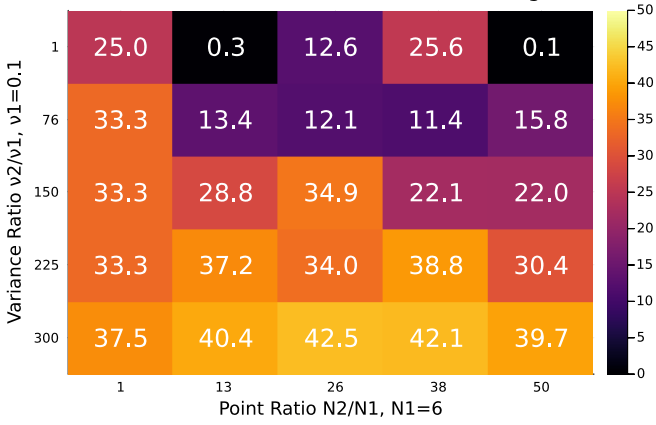
(c) Second subfigure.



(d) Firts subfigure.



(e) Third subfigure.



(f) Foruth subfigure.

Fig. 3: Creating subfigures in LaTeX.

Algorithm 1 ALPCAHUS

Input: $\mathcal{Y} = \{y_1, \ldots, y_N\} \subset \mathbb{R}^D \implies Y \in \mathbb{R}^{D \times N}$: data, $\bar{K} \in \mathbb{Z}^+$: number of candidate subspaces, $\{\bar{d}_k \in \mathbb{Z}^+ \ \forall k \in \{1, \ldots, K\}\}$: candidate dimension for all clusters, $K \in \mathbb{Z}^+$: number of output clusters, $q \in \mathbb{Z}^+$: threshold parameter, $B \in \mathbb{Z}^+$: number of base clusterings, $T \in \mathbb{Z}^+$: number of ALPCAH iterations

Output: $\mathcal{C} = \{c_1, \ldots, c_K\}$: clusters of $\mathcal{Y}$

for $b = 1, \ldots, B$ (in parallel) **do**

  $c_k \sim \{1, \ldots, N\}$ s.t. $|c_k| \approx \frac{N}{\bar{K}}$ for $k = 1, \ldots, \bar{K}$ — Initialize clusters randomly without replacement

  for $k = 1, \ldots, \bar{K}$ (in parallel) **do**

    $\mu_k \leftarrow \frac{1}{|c_k|} \sum_{i \in c_k} y_i$ — Initialize sample means

    $L_k, R_k \leftarrow \text{PCA}(Y_{c_k} - \mu_k 1', \bar{d}_k)$ — Initialize balanced-energy matrices

    $\nu_k \leftarrow \{\max(\frac{1}{D}\|(Y_{c_k} - \mu_k 1' - L_k R_k')e_j\|_2^2, \alpha) \ : \ \forall j \in \{1, \ldots, |c_k|\}\}$ — Initialize noise variances

    $\Pi_k^{-1} \leftarrow \text{Diagonal}(\{v^{-1} \ : \ v \in \nu_k\})$ — Form diagonal matrix

  end for

  for $t = 1, \ldots, T$ (in sequence) **do**

    for $k = 1, \ldots, \bar{K}$ (in parallel) **do**

      $L_k \leftarrow (Y_{c_k} - \mu_k 1')\Pi_k^{-1} R_k (R_k' \Pi_k^{-1} R_k)^{-1}$ — Argmin solution for L

      $R_k \leftarrow (Y_{c_k} - \mu_k 1')' L_k (L_k' L_k)^{-1}$ — Argmin solution for R

      $\nu_k \leftarrow \{\max(\frac{1}{D}\|(Y_{c_k} - \mu_k 1' - L_k R_k')e_j\|_2^2, \alpha) \ : \ \forall j \in \{1, \ldots, |c_k|\}\}$ — Argmin solution for v

      $\Pi_k^{-1} \leftarrow \text{Diagonal}(\{v^{-1} \ : \ v \in \nu_k\})$ — Form diagonal matrix

    end for

    $c_k \leftarrow \{y \in \mathcal{Y} : \forall j \ \ \|L_k L_k^\dagger (y - \mu_k)\|_2 \geq \|L_j L_j^\dagger (y - \mu_j)\|_2\}$ for $k = 1, \ldots, \bar{K}$ — Cluster by projection

    $\mu_k \leftarrow \frac{1}{|c_k|} \sum_{i \in c_k} y_i$ — Calculate new sample means

  end for

  $\mathcal{C}^{(b)} \leftarrow \{c_1, \ldots, c_{\bar{K}}\}$ — Collect results from all trials

end for

$A_{i,j} \leftarrow \frac{1}{B} \left|\{b : y_i, y_j \text{ are co-clustered in } \mathcal{C}^{(b)}\}\right|$ for $i, j = 1, \ldots, N$ — Form affinity matrix

for $i = 1, \ldots, N$ (in parallel) **do**

  $Z^{\text{row}}_{i,:} \leftarrow A_{i,:}$ with the smallest $N - q$ entries set to zero. — Threshold rows of affinity matrix

  $Z^{\text{col}}_{:,i} \leftarrow A_{:,i}$ with the smallest $N - q$ entries set to zero. — Threshold columns of affinity matrix

end for

$\bar{A} \leftarrow \frac{1}{2}\left(Z^{\text{row}} + Z^{\text{col}}\right)$ — Average affinity matrix

$\mathcal{C} \leftarrow \text{SpectralClustering}(\bar{A}, K)$ — Final Clustering

[7] D. Hong, K. Gilman, L. Balzano, and J. A. Fessler, "HePPCAT: Probabilistic PCA for data with heteroscedastic noise," *IEEE Transactions on Signal Processing*, vol. 69, pp. 4819–4834, 2021.

[8] [Online]. Available: https://www.epa.gov/outdoor-air-quality-data

[9] P. Tsalmantza and D. W. Hogg, "A data-driven model for spectra: Finding double redshifts in the sloan digital sky survey," *The Astrophysical Journal*, vol. 753, no. 2, p. 122, 2012.

[10] Y. Cao, A. Zhang, and H. Li, "Multisample estimation of bacterial composition matrices in metagenomics data," *Biometrika*, vol. 107, no. 1, pp. 75–92, 12 2019. [Online]. Available: https://doi.org/10.1093/biomet/asz062

[11] A. R. Zhang, T. T. Cai, and Y. Wu, "Heteroskedastic pca: Algorithm, optimality, and applications," *The Annals of Statistics*, vol. 50, no. 1, pp. 53–80, 2022.

[12] I. T. Jolliffe, *Principal component analysis for special types of data*. Springer, 2002.

[13] T.-H. Oh, Y.-W. Tai, J.-C. Bazin, H. Kim, and I. S. Kweon, "Partial sum minimization of singular values in robust pca: Algorithm and applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 4, pp. 744–758, 2015.

[14] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[15] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Coordinated Science Laboratory Report no. UILU-ENG-09-2215, DC-247*, 2009.

[16] K. Guo, D. Han, and T.-T. Wu, "Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints," *International Journal of Computer Mathematics*, vol. 94, no. 8, pp. 1653–1669, 2017.

[17] B. Mishra, *Algorithmic algebra*. Springer Science & Business Media, 2012.

[18] H. Attouch, J. Bolte, and B. F. Svaiter, "Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods," *Mathematical Programming*, vol. 137, no. 1-2, pp. 91–129, 2013.

[19] L. van den Dries and C. Miller, "Geometric categories and o-minimal structures," *Duke Mathematical Journal*, vol. 84, no. 2, pp. 497 – 540, 1996. [Online]. Available: https://doi.org/10.1215/S0012-7094-96-08416-1

[20] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[21] D. Hong, Y. Sheng, and E. Dobriban, "Selecting the number of components in pca via random signflips," 2023.

[22] D. Hong, F. Yang, J. A. Fessler, and L. Balzano, "Optimally weighted pca for high-dimensional heteroscedastic data," *SIAM Journal on Mathematics of Data Science*, vol. 5, no. 1, pp. 222–250, 2023.

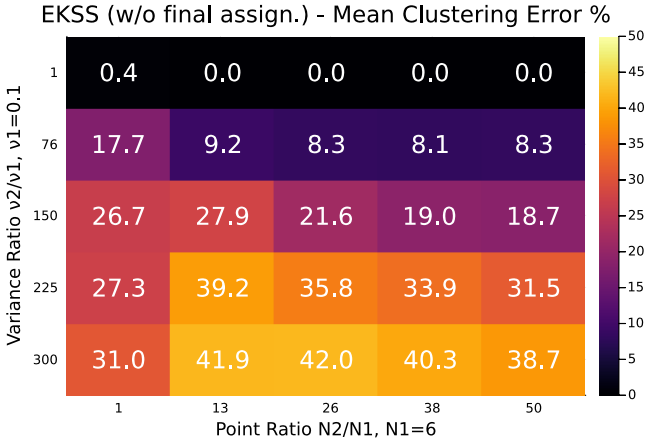
[23] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[24] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 212–225, 2008.

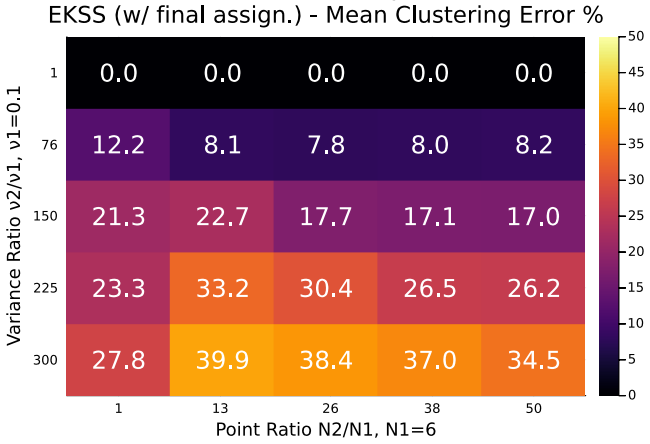
[25] R. Vidal, R. Tron, and R. Hartley, "Multiframe motion segmentation with

missing data using powerfactorization and gpca," *International Journal of Computer Vision*, vol. 79, pp. 85–105, 2008.
[26] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, vol. 1. IEEE, 2003, pp. I–I.
[27] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3655–3671, 2006.
[28] R. Vidal, S. Soatto, Y. Ma, and S. Sastry, "An algebraic geometric approach to the identification of a class of linear hybrid systems," in *42nd IEEE International Conference on Decision and Control (IEEE Cat. No. 03CH37475)*, vol. 1. IEEE, 2003, pp. 167–172.
[29] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *International Journal of Computer Vision*, vol. 29, pp. 159–179, 1998.
[30] L. Lu and R. Vidal, "Combined central and subspace clustering for computer vision applications," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 593–600.
[31] S. R. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," in *2008 IEEE conference on computer vision and pattern recognition*. IEEE, 2008, pp. 1–8.
[32] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 14, 2001.
[33] F. Nie, Z. Zeng, I. W. Tsang, D. Xu, and C. Zhang, "Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering," *IEEE Transactions on Neural Networks*, vol. 22, no. 11, pp. 1796–1808, 2011.
[34] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 11, pp. 2765–2781, 2013.
[35] X. Peng, J. Feng, J. T. Zhou, Y. Lei, and S. Yan, "Deep subspace clustering," *IEEE transactions on neural networks and learning systems*, vol. 31, no. 12, pp. 5509–5521, 2020.
[36] P. K. Agarwal and N. H. Mustafa, "K-means projective clustering," in *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2004, pp. 155–165.
[37] P. Wang, H. Liu, A. M.-C. So, and L. Balzano, "Convergence and recovery guarantees of the k-subspaces method for subspace clustering," in *International Conference on Machine Learning*. PMLR, 2022, pp. 22 884–22 918.
[38] J. Ghosh and A. Acharya, "Cluster ensembles," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, vol. 1, no. 4, pp. 305–315, 2011.
[39] J. Lipor, D. Hong, Y. S. Tan, and L. Balzano, "Subspace clustering using ensembles of k-subspaces," *Information and Inference: A Journal of the IMA*, vol. 10, no. 1, pp. 73–107, 2021.
[40] A. S. Xu, L. Balzano, and J. A. Fessler, "Hemppcat: mixtures of probabilistic principal component analysers for data with heteroscedastic noise," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
[41] J. A. S. Cavazos, J. A. Fessler, and L. Balzano, "Sample-wise heteroscedastic PCA with tail singular value regularization," in *Fourteenth International Conference on Sampling Theory and Applications*, 2023. [Online]. Available: https://openreview.net/forum?id=hocMFgyxrCD
[42] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Mathematical Programming*, vol. 146, no. 1-2, pp. 459–494, 2014.

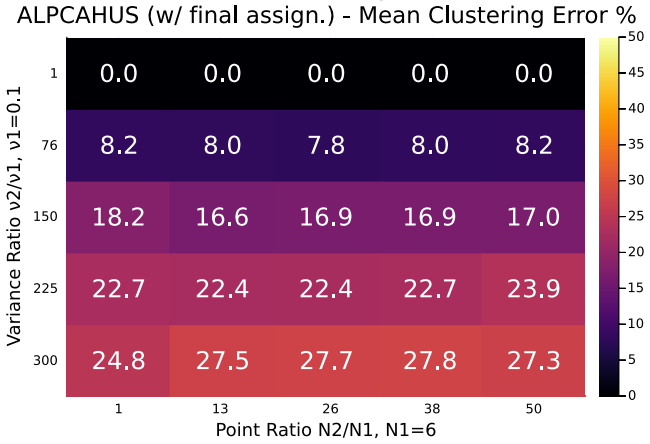
IV. Appendix



(a) Second subfigure.



(b) Firts subfigure.

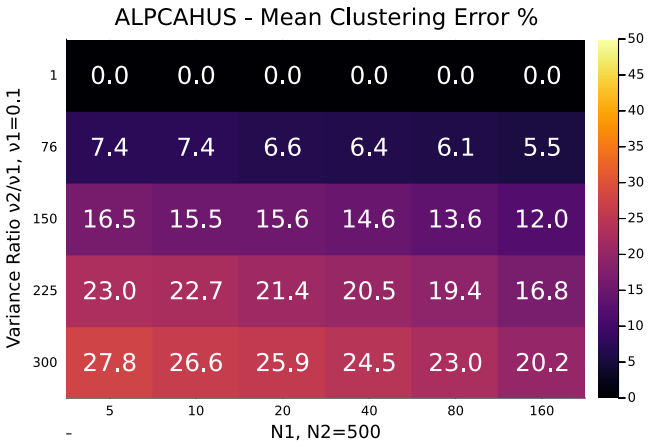


(c) Firts subfigure.

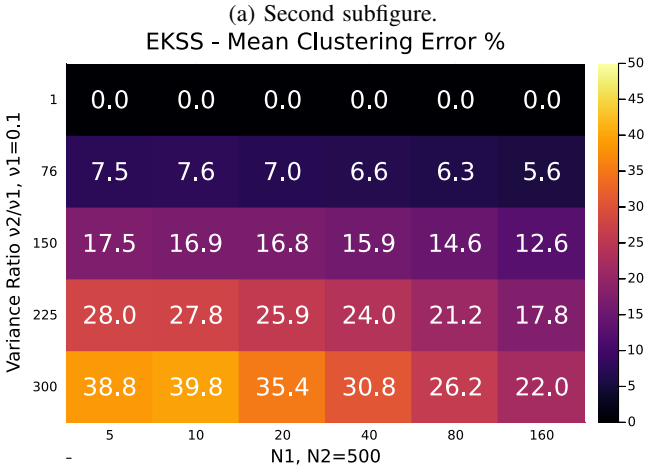
Fig. 4: ALPCAH post processing on EKSS

Note: Final clustering step means we take the final cluster from performing spectral clustering of the affinity matrix, use ALPCAH to find the subspaces of each cluster, then use those subspaces to reassign points to their clusters. ALPCAH (alpcahIter=100?1000?) is used for final clustering step for both EKSS and ALPCAHUS.

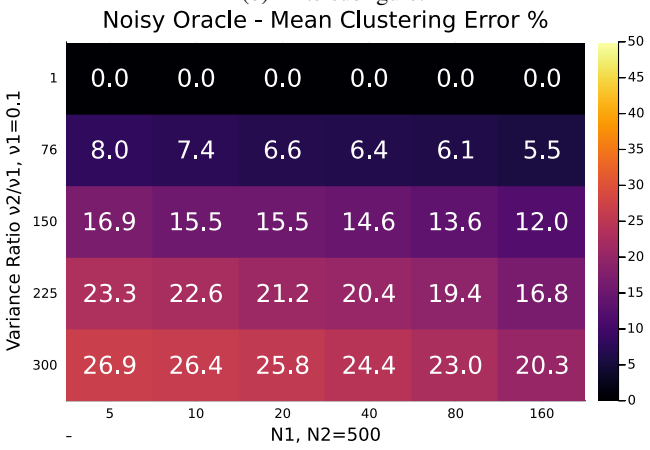
The purpose of this experiment is to show that even if you used EKSS to perform the clustering and then applied a heteroscedastic method like HePPCAT or ALPCAH as the final step you still do not get close to ALPCAHUS. As illustrated, there is a good improvement over not using it at all as a final step, but clearly ALPCAHUS should be used as we are able to more correctly learn the true subspaces bases.



(a) Second subfigure.



(b) Firts subfigure.



(c) Second subfigure.

Fig. 5: Good data experiment

For this experiment, we instead vary N1 and fix N2=500 points to understand how ALPCAHUS and EKSS perform as the data shifts to more good data (v1=0.1). From the figure, we see large differences in clustering error between EKSS and ALPCAHUS when the percentage of good data is very low relative to noisy data. As the good data increases to 160 points (30% of 500 points), we see that EKSS is able to more accuratly determine the correct clustering and achieve similar performance to ALPCAHUS. Of course, in real-world applications, its possible that one does not know the sources of the data and unable to determine whether the data is "effectively" homoscedastic or whether it is heteroscedastic enough to cause trouble in the clustering task. Thus, ALPCAHUS can be useful when both the noise variances and the mixture proportions are unknown to further refine clustering estimates and subspace basis estimates.