ALPCAH : Sample-wise Heteroscedastic PCA with Tail Singular Value Regularization

Javier Salazar Cavazos, Jeffrey A. Fessler, Laura Balzano
EECS Department, University of Michigan
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. Current methods like HePPCAT make Gaussian assumptions of the basis coefficients that may not hold in practice. Other methods such as Weighted PCA (WPCA) assume the noise variances are known, which may be difficult to know in practice. This paper develops a PCA method named ALPCAH 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 is done without distributional assumptions of the low-rank component and without assuming the noise variances are known. Simulations show the effectiveness of accounting for such heteroscedasticity in the data, the benefits of using such a method with all of the data versus retaining only good data, and comparisons against other PCA methods established in the literature like PCA, Robust PCA (RPCA), and HePPCAT. Code available at Github.**

I. 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 estimate of the basis. Some examples of heteroscedastic data sets that involve the subspace basis 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 remove 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 (depending on subspace dimension). 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. Such questions motivate research into heteroscedastic algorithms in general.

Note that while one can consider heteroscedasticity across the feature space with methods such as HeteroPCA [11], this paper will only discuss 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, this method 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 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.

II. 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 D represent the ambient dimension. Let x_i represent the low-dimensional data sample generated by $x_i = U z_i$ 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. Observing the conditional likelihood $y_i | x_i \sim \mathcal{N}(x_i, \nu_i I)$, the probability density function for a single point is

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

Supported in part by NSF Grants ???

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

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

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

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

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

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

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. This algorithm that solves (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

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

where $\sigma_i(X)$ is the ith singular value of X, $\|\cdot\|_*$ is the nuclear norm, and $\|\cdot\|_{\text{Ky-Fan}(k)}$ is the Ky-Fan norm defined as the sum of the first k singular values. Observe that for $k = 0$, $f_k(X) = \|X\|_*$. 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 is $\hat{X} = \sum_{i=1}^{k} \sigma_i u_i v_i' \in \mathbb{R}^{D\times k}$ meaning the solution is identical to a singular value projection approach.

III. 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

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

Definition 1. *Let $A \in \mathbb{R}^{D\times N}$ be a rank k matrix such that SVD$(A) = U_A D_A V_A'$ where $D_A = diag(\sigma_1(A), \ldots, \sigma_{\min(D,N)}(A))$. Let the soft thresholding operation be defined as $\mathcal{S}_\tau[x] = sign(x)\max(|x| - \tau, 0)$ for some threshold $\tau > 0$. Decompose D_A such that $D_A = D_{A1} + D_{A2} = diag(\sigma_1(A), \ldots, \sigma_k(A), 0, \ldots, 0) + diag(0, \ldots, 0, \sigma_{k+1}(A), \ldots, \sigma_N(A))$. 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:*

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

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

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

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

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

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

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

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

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

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 $\mathcal{L}_\mu(X, Z, \Lambda)$. JEFF, MOST OF THE PROOF HERE IS NEW!!!*

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 $f_k(X) = h(X) - q(X) = \|X\|_* - \|X\|_{\text{Ky-Fan}(k)}$. 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 $p_G(\lambda) = 0$. 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 $q(X) = \|X\|_{\text{Ky-Fan}(k)}$. 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 essential 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]. The unknown variance case involves a challenging nonconvex three-block non-separable setup because of the $g(Z, \Pi)$ term that, to the best of our knowledge, has not been explored in the ADMM literature and thus is a topic of future work. □

IV. 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 $U\Sigma V^T = \text{svd}(A)$, 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 Gaussian noise $\mathcal{N}(0, \nu_i I)$. 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 $\|UU' - \hat{U}\hat{U}'\|_{\text{F}}/\|UU'\|_{\text{F}}$ 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. 1-Fig. 4, 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. 1-Fig. 4, 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 coefficient, 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. 1 and Fig. 2 compare ALPCAH against PCA in the simpler situation where noise variances are known. Fig. 2 is similar to Fig. 1 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. 3 and Fig. 4 compare ALPCAH against two other PCA methods in the unknown variance setting. Fig. 3 compares against Robust PCA to see if an "outlier" model can capture heteroscedastic noise. Fig. 4 compares against HePPCAT.

Since Fig. 1-Fig. 4 only show relative error, it is important to discuss absolute error of these algorithms. For Fig. 5-Fig. 6, 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. Note that in Fig. 6, we consider the unknown variance case but show 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. 3 and Fig. 4. 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. 1-Fig. 4 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 case, 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.

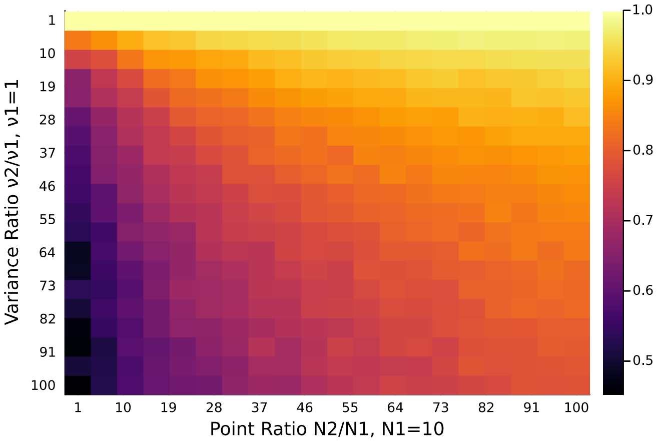

Fig. 1: Ratio of subspace affinity errors ALPCAH/PCA (known variance, no cross-validation required)

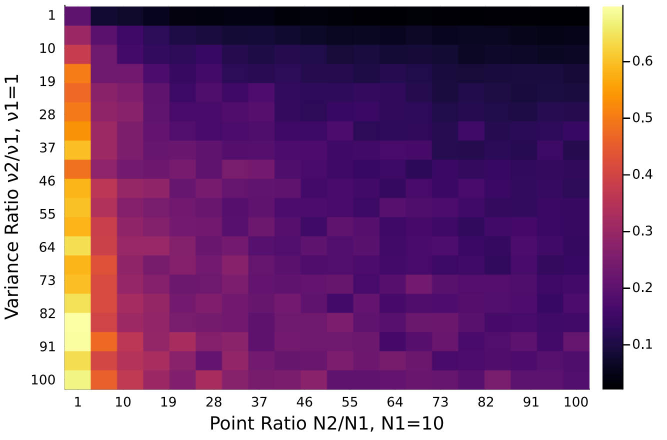

Fig. 2: Ratio of subspace affinity errors ALPCAH/PCA-GOOD (PCA using good data only and ALPCAH using all of the data, no cross-validation required)

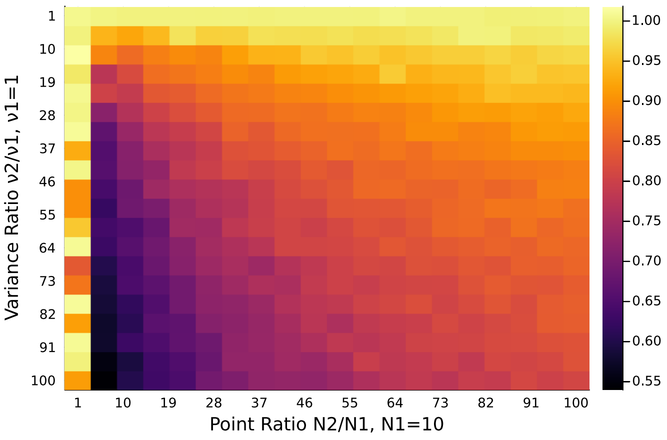

Fig. 3: Ratio of subspace affinity errors ALPCAH/RPCA (unknown variance, no group knowledge, cross-validated λ for both methods)

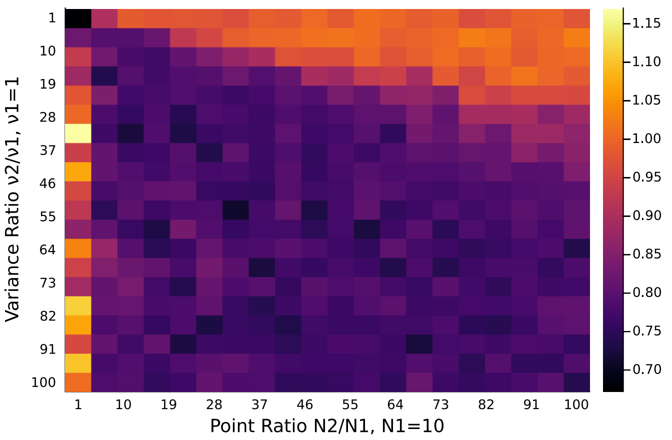

Fig. 4: Ratio of subspace affinity errors ALPCAH/HePPCAT (unknown variance, no group knowledge, cross-validated λ for ALPCAH)

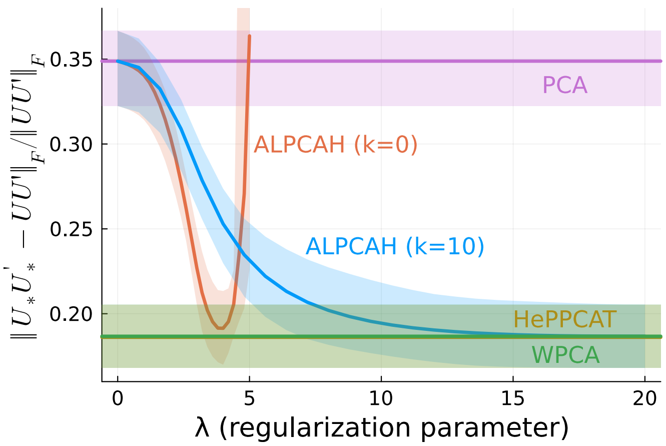

Fig. 5: Subspace affinity error of various PCA methods as the regularization parameter is adjusted (known variance)

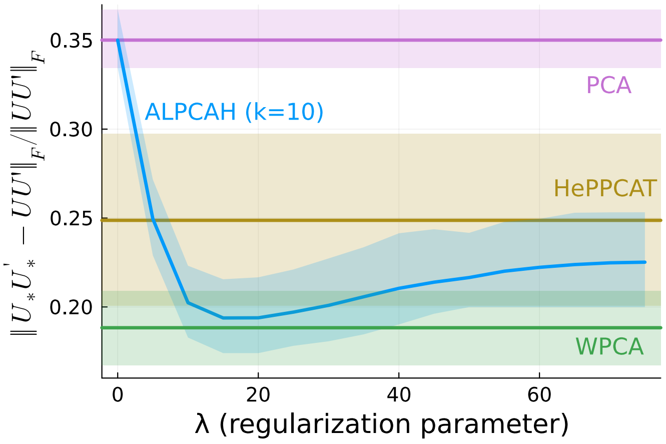

Fig. 6: Subspace affinity error of various PCA methods as the regularization parameter is adjusted (unknown variance, no group knowledge)

V. Discussion

In the known variance cases, Fig. 1 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. 2, 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 was 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 and as such is useful for comparison purposes in this unknown variance case. Since Robust PCA shares similarities with ALPCAH, Fig. 3 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. 4, 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 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. 5 and Fig. 6, 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. 5, 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. 6, 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. 1-Fig. 4. Recall that each heatmap pixel involves a different random subspace basis with data points that have different noise realizations and different 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. A graph showing this result is shown below in Fig. 7.

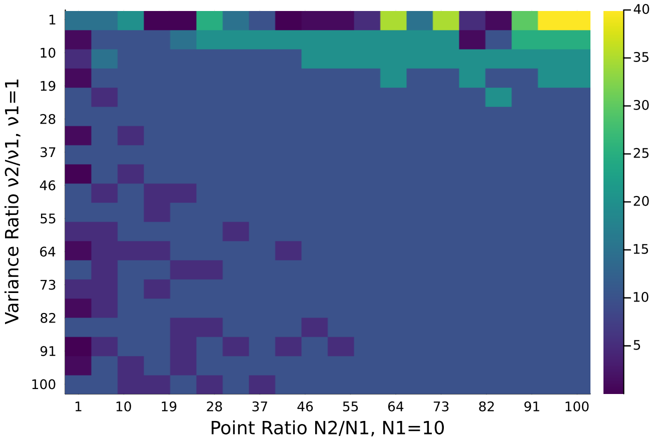

Fig. 7: ALPCAH cross validation matrix of optimized λ values

VI. Conclusion

Heteroscedastic data exists when using mixed data sets that stem from different sources. 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 through the denoised data. ALPCAH avoids the limitations stated above and leads to higher accuracy in the subspace basis estimate as shown in the results section.

VII. Funding Disclosure

This work is supported in part by NSF Grants ???.

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?” *CoRR*, vol. abs/0912.3599, 2009. [Online]. Available: http://arxiv.org/abs/0912.3599

[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/

[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, and Y. Ma, “The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices,” 2010.

[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,” 2020. [Online]. Available: https://arxiv.org/abs/2012.02985

[22] D. Hong, F. Yang, J. A. Fessler, and L. Balzano, “Optimally weighted pca for high-dimensional heteroscedastic data,” 2018. [Online]. Available: https://arxiv.org/abs/1810.12862