Subspace Recovery in Winsorized PCA: Insights into Accuracy and Robustness

Sangil Han
Seoul National University

Kyoowon Kim
Seoul National University

Sungkyu Jung
Seoul National University

Abstract

In this paper, we explore the theoretical properties of subspace recovery using Winsorized Principal Component Analysis (WPCA), utilizing a common data transformation technique that caps extreme values to mitigate the impact of outliers. Despite the widespread use of winsorization in various tasks of multivariate analysis, its theoretical properties, particularly for subspace recovery, have received limited attention. We provide a detailed analysis of the accuracy of WPCA, showing that increasing the number of samples while decreasing the proportion of outliers guarantees the consistency of the sample subspaces from WPCA with respect to the true population subspace. Furthermore, we establish perturbation bounds that ensure the WPCA subspace obtained from contaminated data remains close to the subspace recovered from pure data. Additionally, we extend the classical notion of breakdown points to subspace-valued statistics and derive lower bounds for the breakdown points of WPCA. Our analysis demonstrates that WPCA exhibits strong robustness to outliers while maintaining consistency under mild assumptions. A toy example is provided to numerically illustrate the behavior of the upper bounds for perturbation bounds and breakdown points, emphasizing winsorization's utility in subspace recovery.

1 INTRODUCTION

Winsorization, often referred to as "clipping," has long been recognized as a common and effective tool for handling extreme values in data analysis. Winsorization involves capping extreme values by projecting data points lying outside a specified boundary onto that boundary, ensuring that the support of the transformed data becomes bounded within a specified radius. This transformation is widely used in various fields, including differential privacy, where bounded data support is required to set scales for privacy-preserving noise (Karwa and Vadhan, 2017; Abadi et al., 2016; Kamath et al., 2019; Biswas et al., 2020). Winsorization also effectively handles outliers, particularly from heavy-tailed distributions or corrupted data, by reducing their influence on subsequent analyses without excluding data points (Bickel, 1965; Yale and Forsythe, 1976). This ability to mitigate the impact of anomalies while preserving the overall dataset makes winsorization a frequently adopted technique in multivariate analysis, robust statistics, and other applications (Jose and Winkler, 2008; Beaumont and Rivest, 2009).

In the context of high-dimensional data, where dimension reduction and subspace recovery are crucial, winsorization has been incorporated as a preprocessing step to enhance robustness against anomalies and outliers. Dimension reduction is essential for summarizing high-dimensional datasets by identifying a lower-dimensional subspace that retains the significant variance of the data. Principal Component Analysis (PCA) is the most commonly used method for subspace recovery, but its sensitivity to outliers has led researchers to explore robust alternatives. Various approaches, including optimization-based methods, robust covariance estimation, and subsampling techniques, often involve complex optimization or filtering processes with heavy computational burdens (Candès et al., 2011; Brahma et al., 2018; Zhang et al., 2013). In contrast, data transformations such as winsorization offer a convenient and scalable solution. Winsorization can be applied universally before performing analyses, ensuring that subsequent analyses operate on transformed data with reduced outlier influence. This versatility has made winsorization a valuable tool in a wide range of high-dimensional applications, from dimension reduction to other forms of multivariate analysis.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

Despite the widespread use of winsorization in practice, the theoretical foundations of its impact on subspace recovery have not been fully established. While empirical results have demonstrated its effectiveness in controlling the influence of outliers, there remains a significant gap in understanding how winsorization affects the accuracy and robustness of PCA from a theoretical standpoint. While Raymaekers and Rousseeuw (2019); Leyder et al. (2024) demonstrated the robustness of the covariance matrix of a winsorized random vector in terms of the influence function of eigenvectors and the breakdown of eigenvalues, their results do not address the case where the number of variables p increases, nor do they explore how the winsorization radius (clipping threshold) interacts with p in the high-dimensional model. Furthermore, the effects of winsorization on subspace recovery, particularly in terms of consistency and breakdown points, have yet to be rigorously quantified.

We contribute to the theoretical understanding and robustness of Winsorized PCA (WPCA) in subspace recovery, offering new insights into its consistency and breakdown points.

Accuracy and Consistency in Subspace Recovery. We derive concentration bounds (in Theorem 1) for the PC subspace obtained through WPCA under a broad class of elliptical distributions (Cambanis et al., 1981; Kelker, 1970; Kollo and von Rosen, 2006), which generalize multivariate Gaussian distributions and account for both heavy and light-tailed behavior. The derived concentration bounds for the principal angles between the sample WPCA subspace and the population subspace demonstrate that the sample subspace converges as the sample size increases and the proportion of contamination decreases. Additionally, we demonstrate that WPCA maintains consistency even with extremely large winsorization radius in the subgaussian case, where the distribution has light tails. We further validate the performance of our concentration bounds through a simulation study in high-dimensional settings. The results show that while the concentration bounds perform well in practice, they are not fully optimized, suggesting potential for further improvement.

Strong and Weak breakdown. We introduce a new notion of strong breakdown (Definitions 1 and 2), which offers a more sensitive measure of breakdown compared to the traditional notion. In subspace recovery, while traditional breakdown implies partial orthogonality between corrupted and uncorrupted subspaces (Han et al., 2024), strong breakdown implies full orthogonality. This provides a more refined understanding of estimator behavior in extreme scenarios. We apply both strong and weak breakdown concepts to WPCA, providing a detailed analysis of its robustness.

Breakdown Point Analysis for WPCA and traditional PCA. We show in Theorem 4 that the (strong) breakdown point of the d-dimensional subspace from WPCA has a lower bound proportional to the ratio of the (averaged) eigenvalue gap of the sample covariance of the winsorized data to the square of the winsorization radius, indicating WPCA's resistance to contamination. In contrast, the breakdown points for traditional PCA are much smaller than those of WPCA. This demonstrates WPCA's superior robustness in subspace recovery. We confirm, in a simulated data example, our lower bound is indeed effective.

Robustness through Perturbation Bounds. We demonstrate that the PC subspaces obtained through WPCA not only resist breakdown under contamination but also experience minor perturbation when comparing subspaces from uncontaminated and contaminated data (Theorem 5). We derive perturbation bounds for WPCA, showing that the deviation in the recovered subspace scales linearly with the level of contamination. These bounds confirm WPCA's robustness, indicating that it can tolerate small amounts of corruption without significant deviation in the subspace recovery.

2 WINSORIZED PCA

We implement WPCA as follows. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]' \in \mathbb{R}^{n\times p}$ represent a centered, potentially contaminated data matrix consisting of n samples with p variables. The winsorized dataset is denoted by $\mathbf{X}^{(r)} = [\mathbf{x}_1^{(r)}, \ldots, \mathbf{x}_n^{(r)}]'$, where each winsorized observation is defined as:

$$\mathbf{x}_i^{(r)} := \begin{cases} \mathbf{x}_i & \text{if } \|\mathbf{x}_i\|_2 \le r, \\ \frac{r\mathbf{x}_i}{\|\mathbf{x}_i\|_2} & \text{if } \|\mathbf{x}_i\|_2 > r, \end{cases} \tag{1}$$

where $r > 0$ is the winsorization radius. The winsorization radius r defines the boundary beyond which data points are projected onto the surface of a radius-r ball.

Let $\mathcal{V}_d^{(r)}(\mathbf{X})$ denote the d-dimensional PC subspace spanned by the eigenvectors corresponding to the largest d eigenvalues of the winsorized sample covariance matrix, $\frac{1}{n}(\mathbf{X}^{(r)})'(\mathbf{X}^{(r)})$. We call this subspace d-dimensional winsorized (sample) PC subspace.

Infinitesimally small radius r corresponds to a limiting case of winsorization, where all observations are normalized, which is equivalent to the transformation used in Spherical PCA (SPCA) by (Locantore et al., 1999), and other methods using normalization (Marden, 1999; Visuri et al., 2001; Taskinen et al., 2012; Han et al., 2024). On the other hand, when the radius r is sufficiently large such that $r \ge \max_i\{\|\mathbf{x}_{i,\epsilon}\|_2\}$, no data points are winsorized, and WPCA coincides with traditional PCA.

When performing WPCA on a given dataset, any efficient Singular Value Decomposition (SVD) algorithm can be applied to the winsorized data (1). Numerous studies and implementations of SVD have been developed to handle high-dimensional or large-sample datasets effectively. Factors such as the gap between singular values, the number of rows or columns, and the sparsity (the number of nonzero elements) of the data matrix can influence the choice of the SVD algorithm. Under various scenarios, fast and accurate SVD implementations, such as algorithms proposed by Allen-Zhu and Li (2016); Musco and Musco (2015); Bhojanapalli et al. (2016) can be utilized.

Our analysis focuses on how closely this estimated subspace $\mathcal{V}_d^{(r)}(\mathbf{X})$ approximates the (population) true subspace (in Section 3) and the target subspace derived from the uncontaminated dataset $\mathbf{X}_0$ (in Section 4).

3 ACCURACY OF WINSORIZED PCA

A zero-mean random vector $\mathbf{x} \in \mathbb{R}^p$ is said to follow an *elliptical distribution* with covariance matrix $\mathbf{\Sigma}$, if, for any orthogonal matrix $\mathbf{R} \in \mathbb{R}^{p\times p}$,

$$\mathbf{R}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x} \overset{\mathrm{d}}{=} \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}, \tag{2}$$

where $\mathbf{x} \overset{\mathrm{d}}{=} \mathbf{y}$ means $\mathbf{x}$ and $\mathbf{y}$ have the same distribution. Elliptical distributions, which include multivariate normal and t-distributions, generalize multivariate normal distributions by preserving elliptical symmetry (2). Elliptical distributions characterized by a covariance matrix $\mathbf{\Sigma}$ form a family of distributions defined solely by their elliptical symmetry. A distribution belonging to the family of elliptical distributions cannot be fully specified by its covariance matrix alone, as it may exhibit either heavy or light tails. We use the notation $\mathbf{x} \sim \mathcal{F}_{\mathbf{\Sigma}}$ to indicate that $\mathbf{x}$ follows an elliptical distribution with covariance matrix $\mathbf{\Sigma}$, allowing us to encompass a wide range of random vectors with various tail behaviors. One notable property is that for any $\mathbf{x} \sim \mathcal{F}_{\mathbf{\Sigma}}$ with population covariance matrix $\mathbf{\Sigma}$, winsorization of $\mathbf{x}$ preserves the eigenvectors and the order of eigenvalues in the covariance matrix (Raymaekers and Rousseeuw, 2019), allowing us to infer the eigenstructure of the population covariance matrix even after winsorizing the data points.

To model contamination in the data, we introduce a contamination parameter $\epsilon \in [0, 0.5)$, representing the proportion of corrupted data points among n data points. We assume that the uncontaminated $(1-\epsilon)n$ data points in $\mathbf{X}$, are i.i.d. realizations of a random vector $\mathbf{x} \sim \mathcal{F}_{\mathbf{\Sigma}}$, and the contaminated ϵn data points in $\mathbf{X}$ follow an arbitrary distribution. We denote the ϵ-contaminated dataset by $\mathbf{X} = \mathbf{X}_\epsilon = [\mathbf{x}_{1,\epsilon}, \ldots, \mathbf{x}_{n,\epsilon}]'$, and the set of indices corresponding to the contaminated data points by $\mathcal{I}_\epsilon$, with $|\mathcal{I}_\epsilon| = \epsilon n$. When $\epsilon = 0$, the contaminated dataset becomes the uncontaminated dataset $\mathbf{X}_0$ with n realizations of $\mathbf{x} \sim \mathcal{F}_{\mathbf{\Sigma}}$.

Note that in Section 4, we will remove the distributional assumption. In this case, ϵ-contamination will be allowed to occur at arbitrary positions in the pure dataset $\mathbf{X}_0$ with arbitrary values.

3.1 PC Subspace Concentration

In this section, we provide concentration inequalities for the winsorized PC subspace $\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon)$. This subspace is obtained by applying the traditional PCA on the winsorized contaminated dataset $\mathbf{X}_\epsilon^{(r)}$ as described in Section 2. We demonstrate how the subspace $\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon)$ concentrates around the target subspace. Let the population covariance matrix have eigendecomposition: $\mathrm{Cov}(\mathbf{x}) = \mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$ where $\mathbf{V}'\mathbf{V} = \mathbf{I}_p$, and $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p > 0$. Here $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ contains the eigenvectors of $\mathbf{\Sigma}$ and λ_j are the corresponding eigenvalues. Our target population PC subspace is the d-dimensional subspace spanned by the first d eigenvectors:

$$\mathcal{V}_d = \mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_d).$$

We adopt the largest principal angle as a metric to measure the difference between subspaces (Wedin, 1983; Knyazev and Argentati, 2007; Qiu et al., 2005). For two given d-dimensional subspaces $\mathcal{U}$ and $\mathcal{W}$ of $\mathbb{R}^p$, the d principal angles $0 \leq \theta_1 \leq \cdots \leq \theta_d$ between $\mathcal{U}$ and $\mathcal{W}$ are defined as follows. The smallest principal angle θ_1 between $\mathcal{U}$ and $\mathcal{W}$ is

$$\cos(\theta_1) = \max_{\mathbf{u}\in\mathcal{U}} \max_{\mathbf{w}\in\mathcal{W}} \frac{|\mathbf{u}'\mathbf{w}|}{\|\mathbf{u}\|_2\|\mathbf{w}\|_2} = \frac{|\mathbf{u}_1'\mathbf{w}_1|}{\|\mathbf{u}_1\|_2\|\mathbf{w}_1\|_2} \tag{3}$$

where $\mathbf{u}_1 \in \mathcal{U}$ and $\mathbf{w}_1 \in \mathcal{W}$ are the vectors satisfying $\cos(\theta_1) = \frac{|\mathbf{u}_1'\mathbf{w}_1|}{\|\mathbf{u}_1\|_2\|\mathbf{w}_1\|_2}$. The subsequent principal angles θ_j $(j = 1, \ldots, d)$ are defined recursively by:

$$\cos(\theta_j) = \max_{\mathbf{u}\in\mathcal{U}} \max_{\mathbf{w}\in\mathcal{W}} \frac{|\mathbf{u}'\mathbf{w}|}{\|\mathbf{u}\|_2\|\mathbf{w}\|_2} = \frac{|\mathbf{u}_j'\mathbf{w}_j|}{\|\mathbf{u}_j\|_2\|\mathbf{w}_j\|_2} \tag{4}$$

subject to $\mathbf{u}'\mathbf{u}_k = 0$ and $\mathbf{w}'\mathbf{w}_k = 0$ for $k = 1, \ldots, j-1$. The largest principal angle θ_d provides an upper bound on the deviation between the subspaces. Since $\theta_1 \leq \cdots \leq \theta_d$, if $\theta_d = 0$, then all principal angles are zero, which implies that the two subspaces $\mathcal{U}$ and $\mathcal{W}$ coincide. In this context, we denote $\Theta(\mathcal{U}, \mathcal{W}) = \theta_d$ for the largest principal angle.

Let $\Theta_\epsilon^{(r)} = \Theta(\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon), \mathcal{V}_d)$ be the largest principal angle between $\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon)$ and $\mathcal{V}_d$. We present the following theorem to establish the consistency of the winsorized PC subspace.

Theorem 1. *Assume* $\mathbf{x}_i|_{i\notin\mathcal{I}_\epsilon} \overset{\text{i.i.d}}{\sim} \mathcal{F}_{\mathbf{\Sigma}}$ *follow an elliptical distribution and* $\lambda_d > \lambda_{d+1}$*. Let* $\lambda_j^{(r)}$ *denote the* j*th largest eigenvalue of* $Cov(\mathbf{x}^{(r)})$*, where* $\mathbf{x}^{(r)}$ *is the winsorized random vector of* $\mathbf{x} \sim \mathcal{F}_{\mathbf{\Sigma}}$*. For any* n *and* p*,*

$$E[\sin\Theta_\epsilon^{(r)}] \leq \frac{2r^2\epsilon}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}} + \frac{2^8(\frac{r^2\lambda_1}{p\lambda_p})(\sqrt{\frac{p}{n}} \vee \frac{p}{n})}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}}. \quad (5)$$

Moreover, if

$$\sup_{\mathbf{v}\in S^{p-1}} E[(\mathbf{v}'\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x})^{2k}] \leq \frac{(2k)!}{2^k k!}\sigma^{2k} \quad (6)$$

for all $k = 1, 2, \ldots$ *with some* $\sigma > 0$*, then*

$$E[\sin\Theta_\epsilon^{(r)}] \leq \frac{2r^2\epsilon}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}} + \frac{2^8\lambda_1(\frac{r^2}{p\lambda_p} \wedge \sigma^2)(\sqrt{\frac{p}{n}} \vee \frac{p}{n})}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}}. \quad (7)$$

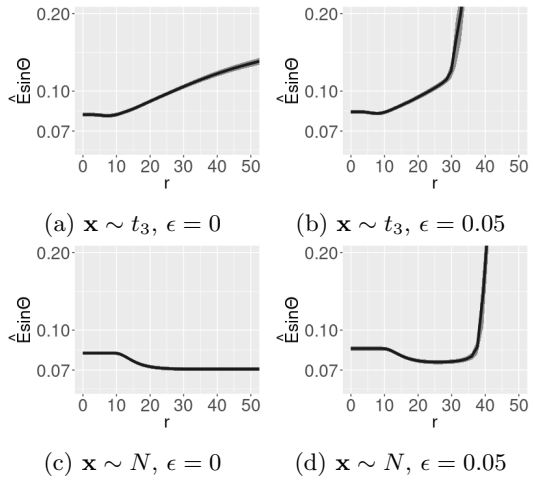
The assumption (6) states that $\mathbf{y} := \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}$ is σ-subgaussian (Wainwright, 2019; Vershynin, 2018). This subgaussian assumption implies that each component of the random vector $\mathbf{y}$ exhibits tail behavior similar to that of a Gaussian distribution, meaning its tails decay exponentially.

Note that the winsorized eigenvalues $\lambda_j^{(r)}$ depend on the winsorization radius r, the eigenvalues of $\mathbf{\Sigma}$, and the number of variables p. To analyze the consistency of the winsorized PC subspace—in terms of convergence in mean—we consider how the parameters r, p, and n interact.

3.1.1 Effect of Winsorization Radius r

We begin with a remark on SPCA: One might conjecture that decreasing r, leading to SPCA, would cause the upper bound in (5) and (7) to converge to 0. However, since the ratio $\lambda_j^{(r)}/r^2$ converges to the jth eigenvalue of the covariance matrix of $\mathbf{x}/\|\mathbf{x}\|_2$, the upper bounds do not vanish as r decreases.

Fixing the number of variables p, we define $g(n, r) = \Omega(h(n, r))$ if there exist constants $a \leq b$ such that $a \leq g(n, r)/h(n, r) \leq b$. First, consider the case without outliers ($\epsilon = 0$). As both the winsorization radius r and the sample size n increase, the upper bound becomes $\Omega\left(\frac{r^2}{\sqrt{n}}\right)$, since the eigenvalue gap $\lambda_d^{(r)} - \lambda_{d+1}^{(r)}$ converges to $\lambda_d - \lambda_{d+1}$ as r grows. Consistency is guaranteed if $\frac{r^2}{\sqrt{n}}$ converges to zero. Without any assumptions on tail behavior, however, increasing r^2 too rapidly relative to n may negatively impact estimation due to potential extreme values from heavy tails. In the presence of outliers ($\epsilon > 0$), the deviation term $2r^2\epsilon/(\lambda_d^{(r)} - \lambda_{d+1}^{(r)})$ grows at the rate $\Omega(r^2)$ resulting in an upper bound of $\Omega(r^2\epsilon + r^2/\sqrt{n})$ in (5).



(a) $\mathbf{x} \sim t_3$, $\epsilon = 0$ (b) $\mathbf{x} \sim t_3$, $\epsilon = 0.05$

(c) $\mathbf{x} \sim N$, $\epsilon = 0$ (d) $\mathbf{x} \sim N$, $\epsilon = 0.05$

Figure 1: Empirical expectation $\widehat{E}[\sin\Theta_\epsilon^{(r)}]$ for different tail behavior and contamination levels. Panels (a) and (b) show the results when $\mathbf{x}_i$ follows a multivariate t_3-distribution, while (c) and (d) represent the case where $\mathbf{x}_i$ follows a multivariate Gaussian distribution. In each figure, ϵ denotes the proportion of contaminated data.

For light-tailed distributions and no outliers, a large r with increasing n ensures consistency, as WPCA approaches traditional PCA, which reliably captures PC directions as n grows. The upper bound (7) reflects this scenario. When $\epsilon = 0$, and $\mathbf{x}$ is subgaussian, the upper bound becomes $\Omega(\frac{1}{\sqrt{n}})$ as n and r increase. In the presence of outliers, the upper bound simplifies to $\Omega(r^2\epsilon + 1/\sqrt{n}) = \Omega(r^2)$.

Figure 1 illustrates the effect of the winsorization radius r on the empirical expectation $\widehat{E}[\sin\Theta_\epsilon^{(r)}]$ (or, simply the 'loss') in both heavy-tailed (t_3) and light-tailed (Gaussian) distributions. The data generation details are provided in the supplementary material. The upper figures correspond to the heavy-tailed t_3-distribution. In Figure 1a, even in the absence of outliers, we observe that the loss increases, reaching approximately 0.19 as r grows. When outliers are present, as shown in Figure 1b, the loss rises significantly and approaches 1 as r increases. The results suggest that the radius does not need to be infinitesimally small; there exists a non-zero radius r where the loss is minimized in both cases.

In contrast, the lower figures depict the results for a multivariate Gaussian distribution, which has light tails. As shown in Figure 1a and 1b, the behavior differs from that of the t_3 distribution. When there are no outliers, increasing r slightly improves the loss. However, when outliers are present, as shown in Figure 1d, the loss

decreases slightly for small r, then increases sharply as r continues to grow.

3.1.2 Effect of Winsorization Radius r in High Dimension

In this section, we examine the high-dimensional setting where the number of variables p increases. We assume that for $j \geq p_0$, the eigenvalues of $\mathbf{\Sigma}$ remain constant at $\lambda_j = \lambda$ for some $p_0 > d$. To analyze the scenario where n, p, and r increase together, we assume $r = p^{1/2+\beta}$, where $\beta \in (-\infty, \infty)$. When $\beta = 0$, the radius is proportional to $\sqrt{p}$. Since the expected norm of the random vector is $E[\mathbf{x}'\mathbf{x}] = \sum_{j=1}^{p} \lambda_j = \Omega(p)$, setting $r = \sqrt{p}$ results in many data points being projected (by the winsorization), while a sufficient number remain un-projected. Positive β implies fewer projected points, while negative β means more projected points.

Corollary 2. *Assume* $\mathbf{x}_i|_{i \notin \mathcal{I}_\epsilon} \overset{\text{i.i.d}}{\sim} \mathcal{F}_{\mathbf{\Sigma}}$ *follow an elliptical distribution and* $\lambda_d > \lambda_{d+1}$. *Let* $r = p^{1/2+\beta}$ *with* $\beta \in (-\infty, \infty)$, *and* C_1, C_2, *and* C_3 *be positive absolute constants.*

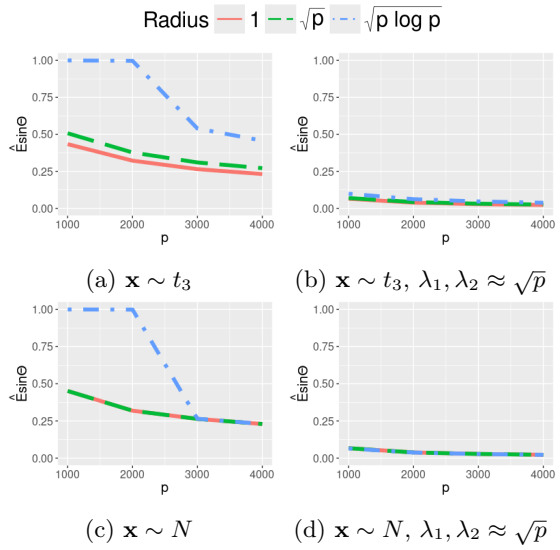
$$E[\sin \Theta_\epsilon^{(r)}] \leq C_1 p^{1+2(\beta \vee 0)} \epsilon + C_2 p^{2(\beta \vee 0)} \left(\sqrt{\frac{p}{n}} \vee \frac{p}{n}\right). \tag{8}$$

Moreover, if $\mathbf{\Sigma}^{-1/2}\mathbf{x}$ *is* σ*-subgaussian, then*

$$E[\sin \Theta_\epsilon^{(r)}] \leq C_1 p^{1+2(\beta \vee 0)} \epsilon + C_3 \left(\sqrt{\frac{p}{n}} \vee \frac{p}{n}\right). \tag{9}$$

In both upper bounds (8) and (9), the first term, related to the contamination proportion ϵ, is dominant, growing at the rate $p^{1+2(\beta \vee 0)}$. Therefore, an excessively large radius $r = p^{1/2+\beta}$ with $\beta > 0$ may cause significant distortion in the winsorized PC subspaces. On the other hand, when there are no outliers ($\epsilon = 0$), a large sample size with p/n converging to 0 guarantees consistency, provided that $\Sigma^{-1/2}\mathbf{x}$ is subgaussian or the data is heavily winsorized with $\beta \leq 0$. We numerically demonstrate the consistency of winsorized PC subspaces in the scenario with increasing p in Section 3.2.

It is known that the estimation of PC subspace has the minimax rate of $\sqrt{p/n}$ (Duchi et al., 2022; Cai et al., 2015; Zhang et al., 2022; Cai et al., 2024). Our asymptotic upper bound can be compared with this rate. For a careful comparison between our rate involving the contamination rate ϵ and the minimax rate, we will assume that the number of contaminated observations is fixed. This simplification gives the rates $(\sqrt{p/n} + p/n)p^{2(\beta \vee 0)}$ for elliptical distributions, and $(\sqrt{p/n} + p^{1+2(\beta \vee 0)}/n)$ under additional sub-Gaussian assumption, for our error bounds. When $n > p$ and



(a) $\mathbf{x} \sim t_3$ (b) $\mathbf{x} \sim t_3$, $\lambda_1, \lambda_2 \approx \sqrt{p}$

(c) $\mathbf{x} \sim N$ (d) $\mathbf{x} \sim N$, $\lambda_1, \lambda_2 \approx \sqrt{p}$

Figure 2: Empirical expectation $\widehat{E}[\sin \Theta_\epsilon^{(r)}]$ for different tail behaviors. Panels (a) and (c) show the results under non-spiked model with the t_3 and Gaussian distributions, respectively. Panels (b) and (d) represent the spiked model.

if we choose $\beta \leq 0$, our rates above become $O(\sqrt{p/n})$. Thus, our method achieves the minimax rate of $\sqrt{p/n}$, demonstrating strong performance even in the presence of contamination. While our rate $O(p/n)$ is sub-optimal when compared to the minimax lower bound for $p > n$, WPCA maintains both robustness and accuracy, even in the challenging scenarios where outliers are heavily contaminated.

3.2 Numerical Study

We simulate the concentration bounds of winsorized PC subspaces in a high-dimensional setting, where the number of variables p increases. For $k = 1, \ldots, 4$, we set the dimension $p_k = 1000k$, and the sample size n_k to satisfy $p_k/n_k = 1/(2k)$. This ensures p/n converges to 0 as p increases. We generate data from two distributions: a heavy-tailed multivariate t_3 and a light-tailed Gaussian. The target subspace dimension is $d = 2$, and we set two outliers with magnitudes proportional to np, positioned orthogonally to the target subspace; that is, $\epsilon = 2/n$.

We test three winsorization radii: $r_{1k} = 1$, $r_{2k} = p_k^{1/2}$, and $r_{3k} = (p_k \log p_k)^{1/2}$. With respect to the parameterization of the radius in Corollary 2, $r_{lk} = p_k^{1/2+\beta_l}$ with $\beta_1 = -1/2$ and $\beta_2 = 0$. The third radius r_{3k} grows slightly faster than r_{2k}. These choices correspond to the effects of small, moderate, and large winsorization radii in high-dimensional settings.

Figure 2a displays the loss (empirical expectation $\widehat{E}[\sin\Theta_\epsilon^{(r)}]$) from the heavy-tailed t_3 distribution for which the eigenvalues of covariance matrix $\mathbf{\Sigma}$ are constant. As one can expect from (8), the losses decrease as p grows. For this heavy-tailed distribution, smaller winsorization radius provides better accuracy.

We next use a spiked covariance model for $\mathbf{\Sigma}$ where the first d eigenvalues scale with $\sqrt{p}$. As shown in Figures 2b, the losses for all radii are smaller than those in the non-spiked model, and tend to zero. This is due to the higher signal-to-noise ratio inherent in the spiked model.

The bottom panels of Figure 2 correspond to the Gaussian distribution. In these light-tailed cases, the winsorization with moderate radius works as good as that with small radius.

Note that when the larger radius (r_{3k}) is used, and for lower p, the outlier adversely affects the subspace estimates. For high p, since the magnitude of the outlier becomes larger than the winsorization radius, winsorization is effective, as can be inspected from Figures 2(a) and (c).

4 ROBUSTNESS OF WINSORIZED PCA

Recall that $\mathbf{X}_0$ represents the uncontaminated data, and $\mathbf{X}_\epsilon$ is the contaminated dataset. In this section, we investigate two aspects of robustness: subspace breakdown points and perturbation bounds.

4.1 Breakdown Point Analysis

4.1.1 Breakdown Points for Real-valued Statistics

We focus on the concept of breakdown points (Hampel, 1968; Bickel et al., 1982; Huber, 1984; Huber and Ronchetti, 2011), which measure the robustness of a statistic against corrupted data. The breakdown point of a statistic is the minimum proportion of corrupted data required to make the statistic "break down." For instance, the sample mean has a breakdown point of $1/n$, meaning a single outlier can drastically affect it, while the sample median, with a breakdown point of $1/2$, is more robust. Formally, the breakdown point of a real-valued statistic $f : \mathcal{X}^n \to \mathbb{R}$ at an n-sample $\mathbf{X}_0 \in \mathcal{X}^n$ is defined as

$$\mathrm{bp}(f; \mathbf{X}_0) := \min_{1\le l\le n} \{\frac{l}{n} : \sup_{\mathbf{Z}_l} |f(\mathbf{Z}_l) - f(\mathbf{X}_0)| = \infty\}, \tag{10}$$

where the supremum is taken over the collection of all possible corrupted data $\mathbf{Z}_l$, obtained by replacing l data points in $\mathbf{X}_0$ with arbitrary values. A breakdown point $\mathrm{bp}(\mathbf{X}_0; f) = m/n$ represents a threshold of resistance, meaning that the statistic f will not break down as long as the proportion of corruption does not exceed m/n.

For the cases where $f(\mathbf{X}_0) \in \mathbb{R}^d$, many researchers have used a global dissimilarity measure in determining the breakdown of f (Hubert et al., 2008; Lopuhaa and Rousseeuw, 1991; Becker and Gather, 1999; He and Simpson, 1992). Typically, $|f(\mathbf{Z}_l) - f(\mathbf{X}_0)|$ in (10) is replaced with $D(f(\mathbf{Z}_l) - f(\mathbf{X}_0))$, where D is the metric that quantifies the dissimilarity between vector-valued estimates. However, the breakdown of a multivariate estimator does not imply that all components break down simultaneously. As an instance, consider the five-number summary, $f(\mathbf{X}_0) = [Q_0, Q_1, Q_2, Q_3, Q_4]$, where Q_0 through Q_4 are the minimum, quartiles, and maximum. The breakdown point of f is $1/n$ when using $D_{\mathbb{R}^5}(f(\mathbf{X}_0), f(\mathbf{Z}_l)) = \|f(\mathbf{X}_0) - f(\mathbf{Z}_l)\|_2$, because the minimum and maximum are sensitive to a single outlier. However, the median Q_2 has a breakdown point of $1/2$. To focus on Q_2 alone, we can use a modified dissimilarity function $\overline{D}_{\mathbb{R}^5}$ on $\mathbb{R}^5 \times \mathbb{R}^5$, defined as $\overline{D}_{\mathbb{R}^5}(f_1, f_2) = |f_{13} - f_{23}|$, where $f_i = (f_{i1}, \ldots, f_{i5})'$, with f_{i3} representing the median component. By replacing $D_{\mathbb{R}^5}$ with $\overline{D}_{\mathbb{R}^5}$, the breakdown point increases to approximately $1/2$. In the next section, we introduce a new notion of *strong breakdown* to explain these different types of breakdown.

4.1.2 Strong Breakdown

Consider a space $\mathcal{D}$ and a statistic $f : \mathcal{X}^n \to \mathcal{D}$. We measure the dissimilarity between $f(\mathbf{X}_0)$ and $f(\mathbf{Z}_l)$ using a dissimilarity function $D : \mathcal{D} \times \mathcal{D} \to [0, \infty)$ satisfying $D(f, f) = 0$ for all $f \in \mathcal{D}$, and $\sup D > 0$. Note that the dissimilarity function D may have two distinct elements $f_1 \neq f_2$ satisfying $D(f_1, f_2) = 0$. In the five-number summary example, $D_{\mathbb{R}^5}$ and $\overline{D}_{\mathbb{R}^5}$ are different dissimilarity functions on $\mathbb{R}^5$. The breakdown point of f with respect to the dissimilarity function D at $\mathbf{X}_0 \in \mathcal{X}^n$ is defined as

$$\mathrm{bp}(f, D; \mathbf{X}_0) := \min_{1\le l\le n} \{\frac{l}{n} : \sup_{\mathbf{Z}_l} D(f(\mathbf{Z}_l), f(\mathbf{X}_0)) = \infty_D\}. \tag{11}$$

where $\infty_D := \sup_{f_1, f_2 \in \mathcal{D}} D(f_1, f_2)$ represents the maximal possible dissimilarity.

For two dissimilarity functions D and $\overline{D}$ on $\mathcal{D}$, we say that $\overline{D}$ is weaker than D (denoted $\overline{D} \preceq D$) if $\lim_{k\to\infty} D(f_{1k}, f_{2k}) = \infty_D$ for any two sequences $f_{1k}, f_{2k} \in \mathcal{D}$ satisfying $\lim_{k\to\infty} \overline{D}(f_{1k}, f_{2k}) = \infty_{\overline{D}}$. Simply put, if $\overline{D}(f, g) = \infty_{\overline{D}}$ gives $D(f, g) = \infty_D$, then $\overline{D} \preceq D$. This relation implies that $\overline{D}$ is less sensitive to breakdown (reaching the maximal difference)

than D. For any $\mathbf{X}_0 \in \mathcal{X}^n$ and statistic $f : \mathcal{X}^n \to \mathcal{D}$, if $\overline{D} \preceq D$, then:

$$\mathrm{bp}(f, \overline{D}; \mathbf{X}_0) \geq \mathrm{bp}(f, D; \mathbf{X}_0).$$

This result indicates that using a weaker dissimilarity function leads to a higher (stronger) breakdown point. In practice, this means that an estimator may appear more robust when assessed with respect to a weaker dissimilarity function, focusing on specific components or aspects of the estimator. Given two dissimilarity functions $\overline{D} \preceq D$, we define strong breakdown point as follows:

Definition 1. *For two given breakdown points,* $\mathrm{bp}(f, \overline{D}; \cdot)$ *and* $\mathrm{bp}(f, D; \cdot)$*, with* $\overline{D} \preceq D$*, we say that* $\mathrm{bp}(f, \overline{D}; \cdot)$ *is the strong breakdown point and* $\mathrm{bp}(f, D; \cdot)$ *is the (weaker) breakdown point.*

4.1.3 Breakdown Points for Subspace-valued statistics

Our interest lies in PC subspaces, thus we extend the notion of breakdown points to subspace-valued statistics using the largest and the smallest principal angles, as dissimilarity functions, on the Grassmannian manifold $\mathrm{Gr}(d, p)$, the set of all d-dimensional linear subspaces in $\mathbb{R}^p$. Let $\mathcal{V} : \mathbb{R}^{n\times p} \to \mathrm{Gr}(d, p)$ be the subspace-valued statistic of interest. An example is the d-dimensional PC subspace derived from data $\mathbf{X}_0$. The smallest and the largest principal angles defined in (3) and (4) are denoted by $\theta(\cdot, \cdot)$ and $\Theta(\cdot, \cdot)$, respectively.

Definition 2. *For* $\mathcal{V} \in \mathrm{Gr}(d, p)$ *and* $\mathbf{X}_0 \in \mathbb{R}^{n\times p}$*, the breakdown point of* $\mathcal{V} \in \mathrm{Gr}(d, p)$ *at* $\mathbf{X}_0$ *is*

$$\begin{aligned}\mathrm{bp}(\mathcal{V}; \mathbf{X}_0) &:= \mathrm{bp}(\mathcal{V}, \Theta; \mathbf{X}_0) \\ &= \min_{1\leq l\leq n} \{\frac{l}{n} : \sup_{\mathbf{Z}_l} \Theta(\mathcal{V}(\mathbf{Z}_l), \mathcal{V}(\mathbf{X}_0)) = \frac{\pi}{2}\}\end{aligned} \tag{12}$$

and the strong breakdown point of $\mathcal{V} : \mathbb{R}^{n\times p} \to \mathrm{Gr}(d, p)$ *at* $\mathbf{X}_0$ *is*

$$\begin{aligned}\overline{\mathrm{bp}}(\mathcal{V}; \mathbf{X}_0) &:= \mathrm{bp}(\mathcal{V}, \theta; \mathbf{X}_0) \\ &= \min_{1\leq l\leq n} \{\frac{l}{n} : \sup_{\mathbf{Z}_l} \theta(\mathcal{V}(\mathbf{Z}_l), \mathcal{V}(\mathbf{X}_0)) = \frac{\pi}{2}\}.\end{aligned} \tag{13}$$

The breakdown point (12) was proposed in Han et al. (2024). For $d \leq p/2$, the strong breakdown point $\overline{\mathrm{bp}}(\mathcal{V}; \mathbf{X}_0)$ is always greater than or equal to the breakdown point $\mathrm{bp}(\mathcal{V}; \mathbf{X}_0)$, since $\theta \preceq \Theta$. When $d > p/2$, any two d-dimensional subspaces must intersect, and strong breakdown never occurs. (In fact, (13) is ill-defined for this case.) Strong breakdown implies that the subspace derived from contaminated data becomes fully orthogonal to the subspace obtained from uncontaminated data. In contrast, weak breakdown occurs when contaminated subspace is only partially orthogonal to its uncontaminated counterpart.

4.2 Breakdown Points of Winsorized PCA

In this section, we begin by examining the lack of robustness of traditional PCA, clarifying the breakdown and strong breakdown points of d-dimensional PC subspaces.

Theorem 3. *Let* $\mathcal{V}_d(\mathbf{X}_0)$ *be given by the* d*-dimensional PC subspace obtained from traditional PCA applied to the data* $\mathbf{X}_0$*, and* $\widehat{\lambda}_j$ *be the* j*th largest eigenvalue of* $\mathbf{X}_0'\mathbf{X}_0/n$*. Assume that* $\widehat{\lambda}_d > \widehat{\lambda}_{d+1}$*. Then,*

$$\mathrm{bp}(\mathcal{V}_d; \mathbf{X}_0) = \frac{1}{n}, \text{ and } \overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}_0) = \frac{d}{n}.$$

This theorem highlights that traditional PCA is highly sensitive to outliers. A single outlier can significantly impact the estimation of the PC subspace, as reflected by the low breakdown point $\mathrm{bp}(\mathcal{V}_d; \mathbf{X}_0) = 1/n$. The strong breakdown point $\overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}_0)$ increases with the dimension d of the subspace. However, since the dimension d is typically much smaller than the number of samples n, even a small fraction of contaminated data can cause substantial distortion.

We provide lower bounds for the breakdown points of winsorized PC subspaces, indicating the robustness of WPCA compared to traditional PCA.

Theorem 4. *Let* $\mathcal{V}_d^{(r)}$ *be a* d*-dimensional PC subspace from WPCA. Then,*

$$\begin{aligned}&\mathrm{bp}(\mathcal{V}_d^{(r)}; \mathbf{X}_0) \geq \frac{1}{2r^2}(\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}), \\ &\overline{\mathrm{bp}}(\mathcal{V}_d^{(r)}; \mathbf{X}_0) \geq \sup_{d_0\leq d} \frac{\sum_{j=1}^{d_0} \widehat{\lambda}_j^{(r)} - \sum_{j=1}^{d_0} \widehat{\lambda}_{d+j}^{(r)}}{2r^2 d_0}.\end{aligned} \tag{14}$$

The lower bounds in (14) are less than or equal to $\frac{1}{2}$, as $\widehat{\lambda}_j^{(r)} \leq r^2$ for all $j = 1, \ldots, p$.

We empirically observe that the smaller the radius r, the more robust the winsorized PC subspace becomes in terms of both strong and weak breakdown points. Figure 3 illustrates how the breakdown points vary as the winsorization radius r varies. The lower bounds of each breakdown point decrease as r increases. We observe that, for every r, the lower bound for the strong breakdown point is larger than that for the (weak) breakdown point. Moreover, the gap between the lower bounds becomes larger as r decreases. All in all, WPCA appears to be less robust to contamination when using a larger winsorization radius, breaking down at lower contamination levels.

4.3 Subspace Perturbation Bound

The notion of breakdown examines only the extreme cases in which the dissimilarity is maximized. Thus,

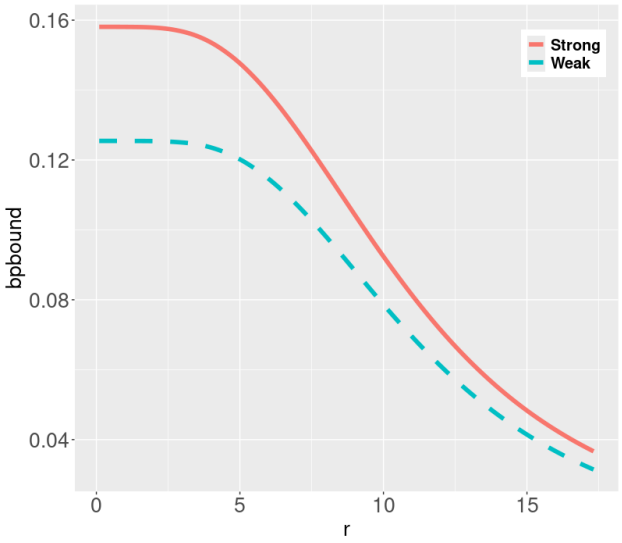


Figure 3: Estimated lower bounds for the breakdown points in (14).

there may be cases under which breakdown does not occur but the quality of the statistics from contaminated data is low. To inspect the amount of deviation of $\mathcal{V}_d(\mathbf{X}_\epsilon)$ from $\mathcal{V}_d(\mathbf{X}_0)$, we establish the subspace perturbation bound using the largest principal angle.

Theorem 5. *Let $\widehat{\lambda}_j^{(r)}$ be the jth largest eigenvalue of $\frac{1}{n}\mathbf{X}_0^{(r)\prime}\mathbf{X}_0^{(r)}$, and $\widehat{\Theta}_\epsilon^{(r)} = \Theta\left(\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon), \mathcal{V}_d^{(r)}(\mathbf{X}_0)\right)$ be the largest principal angle between $\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon)$ and $\mathcal{V}_d^{(r)}(\mathbf{X}_0)$. If $\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)} > 0$, then*

$$\sin\widehat{\Theta}_\epsilon^{(r)} \leq \frac{2r^2\epsilon}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}}. \tag{15}$$

Additionally, if $\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)} > 4r^2\epsilon$, then

$$\sin\widehat{\Theta}_\epsilon^{(r)} \leq \frac{r^2\epsilon}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)} - 2r^2\epsilon}. \tag{16}$$

The theorem establishes that for small values of ϵ, the sine of the largest principal angle $\widehat{\Theta}_\epsilon^{(r)}$ can be bounded by either the linear bound (15) or the rational bound (16). This implies that the winsorized PC subspace remains stable under minor contamination. However, it is important to note that the upper bound (15) does not appear tight, as it grows linearly with the fraction of outliers. Similar statement can be made for SPCA. In particular, the bounds for SPCA are obtained by replacing $\widehat{\lambda}_j^{(r)}/r^2$ with the jth largest eigenvalue of $\sum_{i=1}^n \mathbf{x}_{i,0}\mathbf{x}_{i,0}'/n\|\mathbf{x}_{i,0}\|_2^2$.

To compare the perturbation bounds (15) and (16) in Theorem 5 and the lower bound of the breakdown point

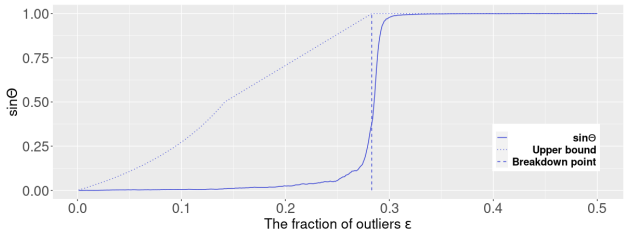


Figure 4: The largest principal angle $\Theta(\mathcal{V}_1^{(r)}(\mathbf{X}_\epsilon), \mathcal{V}_1^{(r)}(\mathbf{X}_0))$ and the perturbation bound versus contamination level ϵ. The solid line represents the observed largest principal angle, the dotted line represents the perturbation bound from Theorem 5, and the vertical dashed line indicates the lower bound of the (weak) breakdown point from Theorem 4.

(14) in Theorem 4, we use a data example. We fix the winsorization radius r to be the median of the norms of the data points, i.e., $r = \text{med}_i\{\|\mathbf{x}_i\|\}$.

For small contamination levels ϵ, the perturbation bound closely follows the observed largest principal angle. This indicates that minor contamination leads to minor perturbation in the WPCA subspace, confirming the robustness of WPCA under small contamination. As ϵ increases, the perturbation bound becomes less sharp, overestimating the actual perturbation. The perturbation bound is conservative for larger contamination levels, suggesting that WPCA performs better in practice than the bound predicts.

In terms of the breakdown point, the principal angle remains relatively small until ϵ approaches the lower bound of the breakdown point in Theorem 4. Once ϵ surpasses this breakdown point (indicated by the vertical dashed line), the principal angle rapidly increases towards $\pi/2$. The lower bound effectively predicts the actual breakdown point beyond which WPCA fails to recover the target subspace.

Consequently, WPCA demonstrates strong robustness to minor contamination, with the perturbation bound effectively predicting subspace deviation for small ϵ. The breakdown point serves as a reliable threshold for subspace stability, effectively indicating when WPCA may fail to recover the target subspace. However, it is important to note that increasing the winsorization radius r can reduce the robustness of WPCA, leading to higher perturbation bounds and lower breakdown points.

5 CONCLUSION

In terms of subspace recovery, our study demonstrates the accuracy of WPCA through concentration inequalities. We show that WPCA maintains consistency

across a wide range of winsorization radii and performs well even in heavy-tailed distributions. Additionally, we demonstrate its consistency and scalability in high-dimensional settings with numerical examples.

Importantly, we introduce the concept of "strong breakdown." Based on this concept, we reveal that WPCA exhibits higher resistance to contamination when compared to traditional PCA. However, we find that an excessively large winsorization radius negatively impacts subspace recovery, causing subspaces to diverge from the target, similar to the case of traditional PCA.

WPCA extends the applicability of traditional PCA by introducing robustness against anomalies. WPCA is most suited for the application areas that involve contaminated data. Examples include analyzing fMRI data for assessing brain connectivity (Lindquist, 2008), brain imaging visualization (Han and Liu, 2018), socio-economic studies for constructing indices like socio-economic status (Vyas and Kumaranayake, 2006). In system dynamics modeling, eigenvalue decomposition—closely related to PCA—serves as a key tool for extracting single-rate system dynamics from multi-rate sampled-data systems Han et al. (2024), in which robustness appears a valuable feature of the modeling.

Future research could explore a spike model where population eigenvalues increase with the number of variables p. As observed in the numerical study in Section 3.2, we anticipate that the spiked eigenvalue model can be used to tighten the upper bounds in Theorem 1 as r and p grow. Investigating the relationship between r, p, and eigenvalues in subspace recovery will provide valuable insights into the theoretical and practical applications of WPCA in high-dimensional settings.

Acknowledgements

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-202400453397). This work was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA2002-03.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016), "Deep Learning with Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, New York, NY, USA: Association for Computing Machinery, CCS '16, pp. 308–318.

Allen-Zhu, Z. and Li, Y. (2016), "LazySVD: Even Faster SVD Decomposition Yet Without Agonizing Pain," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 29.

Beaumont, J.-F. and Rivest, L.-P. (2009), "Chapter 11 - Dealing with Outliers in Survey Data," in *Handbook of Statistics*, ed. Rao, C. R., Elsevier, vol. 29 of *Handbook of Statistics*, pp. 247–279.

Becker, C. and Gather, U. (1999), "The Masking Breakdown Point of Multivariate Outlier Identification Rules," *Journal of the American Statistical Association*, 94, 947–955.

Bhojanapalli, S., Neyshabur, B., and Srebro, N. (2016), "Global Optimality of Local Search for Low Rank Matrix Recovery," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 29.

Bickel, P. J. (1965), "On Some Robust Estimates of Location," *The Annals of Mathematical Statistics*, 36, 847–858.

Bickel, P. J., Doksum, K., and Hodges, J. L. (1982), *A Festschrift For Erich L. Lehmann*, CRC Press.

Biswas, S., Dong, Y., Kamath, G., and Ullman, J. (2020), "CoinPress: Practical Private Mean and Covariance Estimation," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 33, pp. 14475–14485.

Brahma, P. P., She, Y., Li, S., Li, J., and Wu, D. (2018), "Reinforced Robust Principal Component Pursuit," *IEEE Transactions on Neural Networks and Learning Systems*, 29, 1525–1538.

Cai, T., Ma, Z., and Wu, Y. (2015), "Optimal Estimation and Rank Detection for Sparse Spiked Covariance Matrices," *Probability Theory and Related Fields*, 161, 781–815.

Cai, T. T., Xia, D., and Zha, M. (2024), "Optimal Differentially Private PCA and Estimation for Spiked Covariance Matrices," *arXiv preprint arXiv:2401.03820*, 2401.03820.

Cai, T. T. and Zhang, A. (2018), "Rate-Optimal Perturbation Bounds for Singular Subspaces with Applications to High-Dimensional Statistics," *The Annals of Statistics*, 46, 60–89.

Cambanis, S., Huang, S., and Simons, G. (1981), "On the Theory of Elliptically Contoured Distributions," *Journal of Multivariate Analysis*, 11, 368–385.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011), "Robust Principal Component Analysis?" *J. ACM*, 58, 11:1–11:37.

Duchi, J. C., Feldman, V., Hu, L., and Talwar, K. (2022), "Subspace Recovery from Heterogeneous Data with Non-isotropic Noise," *Advances in Neural Information Processing Systems*, 35, 5854–5866.

Hampel, F. R. (1968), *Contributions to the Theory of Robust Estimation*, University of California, Berkeley.

Han, F. and Liu, H. (2018), "ECA: High-Dimensional Elliptical Component Analysis in Non-Gaussian Distributions," *Journal of the American Statistical Association*, 113, 252–268.

Han, S., Jung, S., and Kim, K. (2024), "Robust SVD Made Easy: A Fast and Reliable Algorithm for Large-Scale Data Analysis," in *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, PMLR, pp. 1765–1773.

He, X. and Simpson, D. G. (1992), "Robust Direction Estimation," *The Annals of Statistics*, 20, 351–369.

Huber, P. J. (1984), "Finite Sample Breakdown of M- and P-Estimators," *The Annals of Statistics*, 12, 119–126.

Huber, P. J. and Ronchetti, E. M. (2011), *Robust Statistics*, John Wiley & Sons.

Hubert, M., Rousseeuw, P. J., and Aelst, S. V. (2008), "High-Breakdown Robust Multivariate Methods," *Statistical Science*, 23, 92–119.

Jose, V. R. R. and Winkler, R. L. (2008), "Simple Robust Averages of Forecasts: Some Empirical Results," *International Journal of Forecasting*, 24, 163–169.

Kamath, G., Sheffet, O., Singhal, V., and Ullman, J. (2019), "Differentially Private Algorithms for Learning Mixtures of Separated Gaussians," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 32.

Karwa, V. and Vadhan, S. (2017), "Finite Sample Differentially Private Confidence Intervals," *arXiv preprint arXiv:1711.03908*, 1711.03908.

Kelker, D. (1970), "Distribution Theory of Spherical Distributions and a Location-Scale Parameter Generalization," *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 32, 419–430.

KINGMAN, J. F. C. (1972), "On Random Sequences with Spherical Symmetry," *Biometrika*, 59, 492–494.

Knyazev, A. V. and Argentati, M. E. (2007), "Majorization for Changes in Angles Between Subspaces, Ritz Values, and Graph Laplacian Spectra," *SIAM Journal on Matrix Analysis and Applications*, 29, 15–32.

Kollo, T. and von Rosen, D. (2006), *Advanced Multivariate Statistics with Matrices*, Springer Science & Business Media.

Leyder, S., Raymaekers, J., and Verdonck, T. (2024), "Generalized Spherical Principal Component Analysis," *Statistics and Computing*, 34, 104.

Lindquist, M. A. (2008), "The Statistical Analysis of fMRI Data," *Statistical Science*, 23, 439–464.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., Cohen, K. L., Boente, G., Fraiman, R., Brumback, B., Croux, C., Fan, J., Kneip, A., Marden, J. I., Peña, D., Prieto, J., Ramsay, J. O., Valderrama, M. J., Aguilera, A. M., Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T., and Cohen, K. L. (1999), "Robust Principal Component Analysis for Functional Data," *Test*, 8, 1–73.

Lopuhaa, H. P. and Rousseeuw, P. J. (1991), "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices," *The Annals of Statistics*, 19, 229–248.

Marden, J. I. (1999), "Some Robust Estimates of Principal Components," *Statistics & Probability Letters*, 43, 349–359.

Musco, C. and Musco, C. (2015), "Randomized Block Krylov Methods for Stronger and Faster Approximate Singular Value Decomposition," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol. 28.

Qiu, L., Zhang, Y., and Li, C.-K. (2005), "Unitarily Invariant Metrics on the Grassmann Space," *SIAM Journal on Matrix Analysis and Applications*, 27, 507–531.

Raymaekers, J. and Rousseeuw, P. (2019), "A Generalized Spatial Sign Covariance Matrix," *Journal of Multivariate Analysis*, 171, 94–111.

Taskinen, S., Koch, I., and Oja, H. (2012), "Robustifying Principal Component Analysis with Spatial Sign Vectors," *Statistics & Probability Letters*, 82, 765–774.

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press.

Visuri, S., Oja, H., and Koivunen, V. (2001), "Subspace-Based Direction-of-Arrival Estimation Using Nonparametric Statistics," *IEEE Transactions on Signal Processing*, 49, 2060–2073.

Vyas, S. and Kumaranayake, L. (2006), "Constructing Socio-Economic Status Indices: How to Use Principal Components Analysis," *Health Policy and Planning*, 21, 459–468.

Wainwright, M. J. (2019), *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge University Press.

Wedin, P. Å. (1983), "On Angles between Subspaces of a Finite Dimensional Inner Product Space," in *Matrix Pencils*, eds. Kågström, B. and Ruhe, A., Berlin, Heidelberg: Springer, pp. 263–285.

Yale, C. and Forsythe, A. B. (1976), "Winsorized Regression," *Technometrics*, 18, 291–300.

Yu, Y., Wang, T., and Samworth, R. J. (2015), "A Useful Variant of the Davis–Kahan Theorem for Statisticians," *Biometrika*, 102, 315–323.

Zhang, A. R., Cai, T. T., and Wu, Y. (2022), "Heteroskedastic PCA: Algorithm, Optimality, and Applications," *The Annals of Statistics*, 50, 53–80.

Zhang, L., Shen, H., and Huang, J. Z. (2013), "Robust Regularized Singular Value Decomposition with Application to Mortality Data," *The Annals of Applied Statistics*, 7, 1540–1561.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [No]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A SUPPLEMENTARY MATERIAL

A.1 Technical details in Section 3

A.1.1 Covariance concentration and subguassian parametrization

We provide concentration inequality for the sample covariance of a subgaussian random vector. We say that a random vector with zero mean $\mathbf{x} \in \mathbb{R}^p$ is σ-subgaussian random vector if

$$E[(\mathbf{v}'\mathbf{x})^{2k}] \leq \frac{(2k)!}{2^k k!}\sigma^{2k}.$$

for all $\mathbf{v} \in S^{p-1}$ and $k = 1, 2, \ldots$.

Theorem 6 (Wainwright (2019))**.** *Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]' \in \mathbb{R}^{n\times p}$ be a data matrix whose rows are i.i.d σ-subgaussian random vectors with zero mean and covariance matrix $\mathbf{\Sigma}$. Then the sample covariance $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ satisfies the bound*

$$E[e^{\lambda\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|}] \leq e^{4p+2^5\frac{\lambda^2\sigma^4}{n}}$$

for all $|\lambda| < \frac{n}{2^3\sigma^2}$.

Here $\|\cdot\|$ implies the largest singular value of a matrix. Using this theorem, we can have tail behavior and expectation bound of the sample covariance matrix as follows.

Corollary 7. *Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]' \in \mathbb{R}^{n\times p}$ be a data matrix whose rows are i.i.d σ-subgaussian random vectors with zero mean and covariance matrix $\mathbf{\Sigma}$. Then,*

$$P(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \geq t) \leq \exp\left(4p - \frac{n}{2}\left((\frac{t}{2^3\sigma^2})^2 \wedge \frac{t}{2^3\sigma^2}\right)\right).$$

In other word,

$$P(\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\| \leq 2^3\sigma^2(\sqrt{\frac{2u+8p}{n}} \vee \frac{2u+8p}{n})) \geq 1 - e^{-u}.$$

The expectation is bounded as

$$\begin{aligned} E[\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|] &= 2^3\sigma^2 E[\frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{2^3\sigma^2}] \\ &\leq 2^4\sigma^2(\frac{8p}{n} \vee \sqrt{\frac{8p}{n}}). \end{aligned}$$

Proof. For any $t > 0$, for any $0 < \lambda < n$ by Chernoff bound, we have

$$\begin{aligned} P(\frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{2^3\sigma^2} \geq t) &\leq E[e^{\frac{\lambda}{2^3\sigma^2}\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|}]/e^{\lambda t} \\ &\leq \exp(4p + \frac{1}{2n}\lambda^2 - \lambda t). \end{aligned}$$

By substitute the minimizer $\lambda = n(t \wedge 1)$, we have

$$\begin{aligned} P(\frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{2^3\sigma^2} \geq t) &\leq \exp(4p + \frac{1}{2n}n^2(t\wedge 1)^2 - n(t\wedge 1)t) \\ &\leq \exp(4p - \frac{n}{2}(t^2 \wedge t)). \end{aligned}$$

Additionally, if we set $t = \sqrt{\frac{2u+8p}{n}} \vee \frac{2u+8p}{n}$, we have $t^2 \wedge t = \frac{2u+8p}{n}$ and

$$P(\frac{\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{2^3\sigma^2} \leq \sqrt{\frac{2u+8p}{n}} \vee \frac{2u+8p}{n}) \geq 1 - e^{-u}$$

For the expectation bound,

$$\begin{aligned}
E[\frac{\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|}{2^3\sigma^2}] &= \int_0^\infty \exp\left(-\frac{n}{2}\left(t^2\wedge t\right)+4p\right)\wedge 1dt \\
&= \int_0^1 \exp\left(-\frac{n}{2}t^2+4p\right)\wedge 1dt + \int_1^\infty \exp\left(-\frac{n}{2}t+4p\right)\wedge 1dt \\
&\leq \int_0^{\sqrt{\frac{8p}{n}}\wedge 1} 1dt + \int_{\sqrt{\frac{8p}{n}}\wedge 1}^1 \exp\left(-\frac{n}{2}t^2+4p\right)dt \\
&\quad + \int_1^{\frac{8p}{n}\vee 1} 1dt + \int_{\frac{8p}{n}\vee 1}^\infty \exp\left(-\frac{n}{2}t+4p\right)dt \\
&\leq (\frac{8p}{n}\vee\sqrt{\frac{8p}{n}}) + \int_{\sqrt{\frac{8p}{n}}}^1 \exp\left(-\frac{n}{2}t^2+4p\right)dt\cdot I(\frac{8p}{n}<1) \\
&\quad + \frac{2}{n}\exp(-\frac{n}{2}(\frac{8p}{n}\vee 1)+4p) \\
&\leq (\frac{8p}{n}\vee\sqrt{\frac{8p}{n}}) + \int_0^{1-\sqrt{\frac{8p}{n}}} \exp\left(-\frac{n}{2}(t+\sqrt{\frac{8p}{n}})^2+4p\right)dt\cdot I(\frac{8p}{n}<1)+\frac{2}{n} \\
&\leq (\frac{8p}{n}\vee\sqrt{\frac{8p}{n}}) + \int_0^{1-\sqrt{\frac{8p}{n}}} \exp\left(-t\sqrt{8np}\right)dt\cdot I(\frac{8p}{n}<1)+\frac{2}{n} \\
&\leq (\frac{8p}{n}\vee\sqrt{\frac{8p}{n}}) + \frac{1}{\sqrt{8np}}\cdot I(\frac{8p}{n}<1)+\frac{2}{n} \\
&\leq 2(\frac{8p}{n}\vee\sqrt{\frac{8p}{n}}).
\end{aligned}$$

Consequently, We have

$$P(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|\geq t) = P(\frac{\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|}{2^3\sigma^2}\geq\frac{t}{2^3\sigma^2}) \leq \exp\left(4p-\frac{n}{2}\left((\frac{t}{2^3\sigma^2})^2\wedge\frac{t}{2^3\sigma^2}\right)\right),$$

$$P(\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|\leq 2^3\sigma^2(\sqrt{\frac{2u+8p}{n}}\vee\frac{2u+8p}{n})) = P(\frac{\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|}{2^3\sigma^2}\leq(\sqrt{\frac{2u+8p}{n}}\vee\frac{2u+8p}{n})) \leq 1-e^{-u}$$

and

$$\begin{aligned}
E[\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|] &= 2^3\sigma^2 E[\frac{\|\widehat{\mathbf{\Sigma}}-\mathbf{\Sigma}\|}{2^3\sigma^2}] \\
&\leq 2^4\sigma^2(\frac{8p}{n}\vee\sqrt{\frac{8p}{n}}).
\end{aligned}$$

□

Here, we characterizes the subgaussian parameter σ of $\mathbf{x}^{(r)}$.

Lemma 8. *Assume that* $\mathbf{y} := \mathbf{\Sigma}^{-1/2}\mathbf{x}$ *be a* σ*-subgaussian. If* $\mathbf{y} = (y_1,\ldots,y_p)'$ *is not subgaussian, we denote* $\sigma=\infty$*. Then* $\mathbf{x}^{(r)}$ *is* $(\sqrt{\lambda_1}\sigma\wedge\sqrt{\frac{\lambda_1 r^2}{\lambda_p p}})$*-subgaussian.*

Proof. When $\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}'$ as described in Section 3.1, Without loss of generality, we assume that $\mathbf{V} = \mathbf{I}_p$. Let

$s^2 = \|\mathbf{x}\|_2^2 = \sum_{l=1}^p \lambda_l y_l^2$. For any $\mathbf{w} \in S^{p-1}$, we have

$$\begin{aligned}
E[(\mathbf{w}'\mathbf{x}^{(r)})^{2k}] &= E\left[\left(\mathbf{w}'\mathbf{x}\left(1 \wedge \sqrt{\frac{r^2}{s^2}}\right)\right)^{2k}\right] \\
&= E\left[\left((\mathbf{w}'\sqrt{\mathbf{\Lambda}}\mathbf{y})^2\left(1 \wedge \frac{r^2}{s^2}\right)\right)^{k}\right] \\
&\leq E\left[\left(\left(\|\sqrt{\mathbf{\Lambda}}\mathbf{w}\|_2 \frac{(\sqrt{\mathbf{\Lambda}}\mathbf{w})'}{\|\sqrt{\mathbf{\Lambda}}\mathbf{w}\|_2}\mathbf{y}\right)^2\left(1 \wedge \frac{r^2}{\lambda_p \sum_{l=1}^p y_l^2}\right)\right)^{k}\right] \\
&= E\left[\|\sqrt{\mathbf{\Lambda}}\mathbf{w}\|_2^{2k}\left(\left(\frac{(\sqrt{\mathbf{\Lambda}}\mathbf{w})'}{\|\sqrt{\mathbf{\Lambda}}\mathbf{w}\|_2}\mathbf{y}\right)^2\left(1 \wedge \frac{r^2}{\lambda_p \mathbf{y}'\mathbf{y}}\right)\right)^{k}\right] \\
&\leq \lambda_1^k E\left[\left(\left(\frac{(\sqrt{\mathbf{\Lambda}}\mathbf{w})'}{\|\sqrt{\mathbf{\Lambda}}\mathbf{w}\|_2}\mathbf{y}\right)^2\left(1 \wedge \frac{r^2}{\lambda_p \mathbf{y}'\mathbf{y}}\right)\right)^{k}\right] \\
&= \lambda_1^k E\left[\left(y_1^2\left(1 \wedge \frac{r^2}{\lambda_p \mathbf{y}'\mathbf{y}}\right)\right)^{k}\right],
\end{aligned}$$

since for rotation $\mathbf{R}$ satisfying $\mathbf{R}\frac{\sqrt{\mathbf{\Lambda}}\mathbf{w}}{\|\sqrt{\mathbf{\Lambda}}\mathbf{w}\|_2} = (1, 0, \ldots, 0)'$, $\mathbf{R}\mathbf{y} \overset{\mathrm{d}}{=} \mathbf{y}$ and $\mathbf{y}'\mathbf{y} \overset{\mathrm{d}}{=} (\mathbf{R}\mathbf{y})'(\mathbf{R}\mathbf{y})$. Note that $\frac{y_1^2}{\mathbf{y}'\mathbf{y}} \overset{\mathrm{d}}{=} \frac{z_1^2}{\mathbf{z}'\mathbf{z}}$ for standard gaussian random vector $\mathbf{z} = (z_1, \ldots, z_p)'$, thus the distribution of $\frac{y_1^2}{\mathbf{y}'\mathbf{y}}$ is the beta distribution with parameters $(\frac{1}{2}, \frac{p}{2})$. Using this, we obtain

$$\begin{aligned}
E[(\mathbf{w}'\mathbf{x}^{(r)})^{2k}] &\leq \lambda_1^k \left(E\left[y_1^{2k}\right] \wedge E\left[\left(\frac{r^2 y_1^2}{\lambda_p \mathbf{y}'\mathbf{y}}\right)^k\right]\right) \\
&\leq \lambda_1^k \left(\frac{(2k)!}{2^k k!}\sigma^{2k} \wedge \frac{r^{2k}}{\lambda_p^k p^k}\frac{(2k)!}{2^k k!}\right) \\
&= \left((\sqrt{\lambda_1}\sigma)^{2k} \wedge \left(\sqrt{\frac{\lambda_1 r^2}{\lambda_p p}}\right)^{2k}\right)\frac{(2k)!}{2^k k!}.
\end{aligned}$$

Thus, $\mathbf{x}^{(r)}$ is $(\sqrt{\lambda_1}\sigma \wedge \sqrt{\frac{\lambda_1 r^2}{\lambda_p p}})$-subgaussian. □

By applying Lemma 8 and Corollary 7 to $\mathbf{x}^{(r)}$, we have the following theorem.

Theorem 9 (Covariance concentration)**.** *Let $\widehat{\mathbf{\Sigma}}_\epsilon^{(r)} = \frac{1}{n}\mathbf{X}_\epsilon'\mathbf{X}_\epsilon$ be the winsorized sample covariance matrix of contaminated data $\mathbf{X}_\epsilon$. Then for any $r > 0$, $p > 1$, and $\epsilon \in [0, \frac{1}{2})$,*

$$E[\|\widehat{\mathbf{\Sigma}}_\epsilon^{(r)} - \mathbf{\Sigma}^{(r)}\|] \leq \epsilon r^2 + 2^4\sigma_r^2\left(\frac{8p}{n} \vee \sqrt{\frac{8p}{n}}\right). \tag{17}$$

Proof. Note that $\mathcal{I}_\epsilon$ represents the indices of corrupted data. Let $\widehat{\mathbf{\Sigma}}_{\text{in}} = \frac{1}{(1-\epsilon)n}\sum_{i \notin \mathcal{I}_\epsilon} \mathbf{x}_{i,\epsilon}^{(r)}\mathbf{x}_{i,\epsilon}^{(r)\prime}$ be the sample covariance of pure samples, and $\widehat{\mathbf{\Sigma}}_{\text{out}} = \frac{1}{n\epsilon}\sum_{i \in \mathcal{I}_\epsilon} \mathbf{x}_{i,\epsilon}^{(r)}\mathbf{x}_{i,\epsilon}^{(r)\prime}$ be the sample covariance of outliers. Then, we have $\widehat{\mathbf{\Sigma}}_\epsilon^{(r)} = \epsilon\widehat{\mathbf{\Sigma}}_{\text{out}} + (1-\epsilon)\widehat{\mathbf{\Sigma}}_{\text{in}}$. Note that $\|(\widehat{\mathbf{\Sigma}}_{\text{out}} - \mathbf{\Sigma}^{(r)})\| = \eta_1(\widehat{\mathbf{\Sigma}}_{\text{out}} - \mathbf{\Sigma}^{(r)}) \vee \eta_1(-\widehat{\mathbf{\Sigma}}_{\text{out}} + \mathbf{\Sigma}^{(r)})$ where $\eta_1(\cdot)$ is the

largest eigenvalue. Thus we have $\|(\widehat{\mathbf{\Sigma}}_{\text{out}} - \mathbf{\Sigma}^{(r)})\| \le r^2$

$$\begin{aligned}
E[\|\widehat{\mathbf{\Sigma}}_{\epsilon}^{(r)} - \mathbf{\Sigma}^{(r)}\|] &= E[\|\epsilon\widehat{\mathbf{\Sigma}}_{\text{out}} + (1-\epsilon)\widehat{\mathbf{\Sigma}}_{\text{in}} - \mathbf{\Sigma}^{(r)}\|] \\
&\le E[\|\epsilon(\widehat{\mathbf{\Sigma}}_{\text{out}} - \mathbf{\Sigma}^{(r)})\|] + E[\|(1-\epsilon)(\widehat{\mathbf{\Sigma}}_{\text{in}} - \mathbf{\Sigma}^{(r)})\|] \\
&\le \epsilon r^2 + (1-\epsilon)E[\|\widehat{\mathbf{\Sigma}}_{\text{in}} - \mathbf{\Sigma}^{(r)}\|] \\
&\le \epsilon r^2 + 2^4(1-\epsilon)\sigma_r^2\left(\frac{8p}{(1-\epsilon)n} \vee \sqrt{\frac{8p}{(1-\epsilon)n}}\right) \\
&\le \epsilon r^2 + 2^4\sigma_r^2\left(\frac{8p}{n} \vee \sqrt{\frac{8p}{n}}\right).
\end{aligned}$$

□

The concentration of a sample covariance is highly related to concentration of subspaces spanned by eigenvectors of the sample covariance matrix. We provide a lemma based on the variant of Davis-Kahan theorem by Yu et al. (2015).

Lemma 10. *Let $\mathbf{\Sigma}, \widehat{\mathbf{\Sigma}} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_p$ and $\widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_p$, respectively. Assume that $\lambda_d > \lambda_{d+1}$. Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d) \in \mathbb{R}^{p\times d}$ and $\widehat{\mathbf{V}} = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_d) \in \mathbb{R}^{p\times d}$ have orthonormal eigenvector columns satisfying $\mathbf{\Sigma}\mathbf{v}_j = \lambda_j\mathbf{v}_j$ and $\widehat{\mathbf{\Sigma}}\widehat{\mathbf{v}}_j = \widehat{\lambda}_j\widehat{\mathbf{v}}_j$ for $j = 1, \ldots, d$. Let $\mathcal{V}$ and $\widehat{\mathcal{V}}$ be the subspaces in $\mathbb{R}^p$ spanned by the columns of $\mathbf{V}$ and $\widehat{\mathbf{V}}$, respectively. Let $\Theta \in [0, \frac{\pi}{2}]$ be the largest principal angle between $\mathcal{V}$ and $\widehat{\mathcal{V}}$. Then,*

$$\sin\Theta \le \frac{2\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|}{\lambda_d - \lambda_{d+1}}$$

Proof. The details of this proof is given by replacing Frobenius norm with operator 2-norm in Yu et al. (2015). Let $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_d)$, $\mathbf{\Lambda}_\perp = \text{diag}(\lambda_{d+1}, \ldots, \lambda_p)$ and $\widehat{\mathbf{\Lambda}} = \text{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_d)$. Then,

$$\begin{aligned}
\mathbf{O} &= \widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}} - \widehat{\mathbf{V}}\widehat{\mathbf{\Lambda}} \\
&= (\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma} + \mathbf{\Sigma})\widehat{\mathbf{V}} - \widehat{\mathbf{V}}(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda} + \mathbf{\Lambda}) \\
&= \mathbf{\Sigma}\widehat{\mathbf{V}} - \widehat{\mathbf{V}}\mathbf{\Lambda} + (\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\widehat{\mathbf{V}} - \widehat{\mathbf{V}}(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda})
\end{aligned}$$

where $\mathbf{O}$ is a matrix with a proper size whose elements are zero. Thus, we have

$$\begin{aligned}
\|\mathbf{\Sigma}\widehat{\mathbf{V}} - \widehat{\mathbf{V}}\mathbf{\Lambda}\| &= \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\widehat{\mathbf{V}} - \widehat{\mathbf{V}}(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda})\| \\
&\le \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\widehat{\mathbf{V}}\| + \|\widehat{\mathbf{V}}(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda})\| \\
&\le \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\|\|\widehat{\mathbf{V}}\| + \|\widehat{\mathbf{V}}\|\|(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda})\| \\
&\le \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\| + \|(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda})\| \\
&\le 2\|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\|,
\end{aligned}$$

since $\|\widehat{\mathbf{V}}\| \le 1$, and $\|(\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda})\| \le \|(\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma})\|$ by Weyl's inequality. Meanwhile, since $\|\mathbf{V}_\perp\| = 1$, we have

$$\begin{aligned}
\|\mathbf{\Sigma}\widehat{\mathbf{V}} - \widehat{\mathbf{V}}\mathbf{\Lambda}\| &= \|\mathbf{V}_\perp'\|\|\mathbf{\Sigma}\widehat{\mathbf{V}} - \widehat{\mathbf{V}}\mathbf{\Lambda}\| \\
&\ge \|\mathbf{V}_\perp'(\mathbf{\Sigma}\widehat{\mathbf{V}} - \widehat{\mathbf{V}}\mathbf{\Lambda})\| \\
&= \|\mathbf{V}_\perp'\mathbf{\Sigma}\widehat{\mathbf{V}} - \mathbf{V}_\perp'\widehat{\mathbf{V}}\mathbf{\Lambda}\| \\
&= \|\mathbf{\Lambda}_\perp\mathbf{V}_\perp'\widehat{\mathbf{V}} - \mathbf{V}_\perp'\widehat{\mathbf{V}}\mathbf{\Lambda}\| \\
&\ge \|\mathbf{V}_\perp'\widehat{\mathbf{V}}\mathbf{\Lambda}\| - \|\mathbf{\Lambda}_\perp\mathbf{V}_\perp'\widehat{\mathbf{V}}\| \\
&\ge \lambda_d\|\mathbf{V}_\perp'\widehat{\mathbf{V}}\| - \lambda_{d+1}\|\mathbf{V}_\perp'\widehat{\mathbf{V}}\| \\
&= (\lambda_d - \lambda_{d+1})\|\mathbf{V}_\perp'\widehat{\mathbf{V}}\|.
\end{aligned}$$

Here, the last inequality holds since

$$\begin{aligned}
\|\mathbf{V}_\perp'\widehat{\mathbf{V}}\mathbf{\Lambda}\| &= \sup_{\|\mathbf{u}\|_2=1} \|\mathbf{\Lambda}\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}\|_2 \\
&= \sup_{\|\mathbf{u}\|_2=1} \|\mathbf{\Lambda}\frac{\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}}{\|\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}\|_2}\|\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}\|_2\|_2 \\
&= \sup_{\|\mathbf{u}\|_2=1} \|\mathbf{\Lambda}\frac{\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}}{\|\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}\|_2}\|_2\|\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}\|_2 \\
&\geq \sup_{\|\mathbf{u}\|_2=1} \lambda_d\|\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{u}\|_2 \\
&= \lambda_d\|\mathbf{V}_\perp'\widehat{\mathbf{V}}\|.
\end{aligned}$$

Note that

$$\begin{aligned}
\|\mathbf{V}_\perp'\widehat{\mathbf{V}}\|^2 &= \|\widehat{\mathbf{V}}'\mathbf{V}_\perp\mathbf{V}_\perp'\widehat{\mathbf{V}}\| \\
&= \|\widehat{\mathbf{V}}'(\mathbf{I}_p - \mathbf{V}\mathbf{V}')\widehat{\mathbf{V}}\| \\
&= \|\widehat{\mathbf{V}}'\widehat{\mathbf{V}} - \widehat{\mathbf{V}}'\mathbf{V}\mathbf{V}'\widehat{\mathbf{V}}\| \\
&= \|\mathbf{I}_d - \widehat{\mathbf{V}}'\mathbf{V}\mathbf{V}'\widehat{\mathbf{V}}\| \\
&= \|\mathbf{P}\mathbf{P}' - \mathbf{P}\cos^2\mathbf{P}'\| \\
&= \sin^2\Theta.
\end{aligned}$$

where $\widehat{\mathbf{V}}'\mathbf{V}\mathbf{V}'\widehat{\mathbf{V}} = \mathbf{P}\text{diag}(\cos^2(\theta_1),\ldots,\cos^2(\theta_d))\mathbf{P}'$ is the eigen decomposition and the entries of the diagonal matrix $\text{diag}(\cos^2(\theta_1),\ldots,\cos^2(\theta_d))$ represent the squares of the cosines of the principal angles between $\mathcal{V}$ and $\widehat{\mathcal{V}}$. □

A.1.2 Proof of Theorem 1

Proof. By combining Theorem 9 and Lemma 10, we have

$$\begin{aligned}
E\left[\sin\Theta_\epsilon^{(r)}\right] &\leq E\left[\frac{2\|\widehat{\mathbf{\Sigma}}_\epsilon^{(r)} - \mathbf{\Sigma}^{(r)}\|}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}}\right] \\
&\leq \frac{2r^2\epsilon}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}} + \frac{2^5\sigma_r^2(\frac{8p}{n} \vee \sqrt{\frac{8p}{n}})}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}}.
\end{aligned}$$

Using Lemma 8, we conclude the results of Theorem 1. □

A.1.3 Proof of Corollary 2

Proof. We first provide asymptotic properties about winsorized eigenvalue $\lambda_j(r)$ and the radius r.

Lemma 11. *For increasing p and $r = p^{\frac{1}{2}+\beta}$, Let $\lambda_j^{(r)}$ denote the jth largest eigenvalue of Cov$(\mathbf{x}^{(r)})$. Then,*

$$\lambda_j^{(r)} = \Omega(p^{2(\beta\wedge 0)}), \text{ and } \frac{r^2}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}} = \Omega(p^{1+2(\beta\vee 0)}).$$

proof of lemma. Let $\mathbf{\Sigma}^{-1/2}\mathbf{x} = \mathbf{y} = [y_1,\ldots,y_p]'$. By KINGMAN (1972), there exists a non-negative random variable w such that $y_j = \sqrt{w}z_j$ where $z_1, z_2, \ldots$ are i.i.d. standard normal random variables. Without loss of generality, we assume that $\lambda_j^{(r)} = E[\lambda_j y_j^2(1 \wedge \frac{r^2}{s^2})]$ with $s^2 = \sum_{l=1}^p \lambda_l y_l^2$. Note that $\lambda_j y_j^2(1 \wedge \frac{r^2}{s^2}) \leq \lambda_j y_j^2$ with $E[\lambda_j y_j^2] = \lambda_j$. If $\beta > 0$, by Dominant Convergence Theorem, we have

$$\begin{aligned}
\lim_p \lambda_j^{(r)} &= \lambda_j E[y_j^2 \wedge \lim_p \frac{p^{1+2\beta}y_j^2}{\sum_{l=1}^p \lambda_l y_l^2})] \\
&= \lambda_j E[y_j^2] = \lambda_j.
\end{aligned}$$

Similarly, if $\beta = 0$,

$$\begin{aligned}\lim_p \lambda_j^{(r)} &= \lambda_j E[y_j^2 \wedge \lim_p \frac{py_j^2}{\sum_{l=1}^p \lambda_l y_l^2})] \\ &= \lambda_j E[w \wedge \frac{1}{\lambda})].\end{aligned}$$

In case of $\beta < 0$,

$$\begin{aligned}\frac{p}{r^2}\lambda_j^{(r)} &= \frac{p}{r^2} E[\lambda_j y_j^2 (1 \wedge \frac{r^2}{s^2})] \\ &\leq E[\frac{p\lambda_j z_j^2}{\lambda \sum_{l=1}^p z_l^2})] \\ &= \lambda_j/\lambda.\end{aligned}$$

Conversely, by Fatou's lemma, we have

$$\begin{aligned}\liminf_p \frac{p}{r^2}\lambda_j^{(r)} &\geq E[\liminf_p \frac{p}{r^2}\lambda_j y_j^2 (1 \wedge \frac{r^2}{s^2})] \\ &= E[\frac{\lambda_j z_j^2}{\lambda})] = \frac{\lambda_j}{\lambda}.\end{aligned}$$

Thus we have $\lim_p \frac{p}{r^2}\lambda_j^{(r)} = \lambda_j/\lambda$. Consequently,

$$\begin{aligned}\lambda_j^{(r)} &= \Omega(p^{2(\beta \wedge 0)}), \\ \frac{r^2}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}} &= \Omega(p^{1+2(\beta \vee 0)}).\end{aligned}$$

□

By applying this lemma to Theorem 1, we have the conclusion. □

A.2 Technical details in Section 4

A.2.1 Proof of Theorem 3

Proof. By Propositioin 3 in Han et al. (2024), $\mathrm{bp}(\mathcal{V}_d; \mathbf{X}) = \frac{1}{n}$ holds. We show $\overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}) \leq \frac{d}{n}$. Since $2d \leq p$, we can find d orthonormal vectors $\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_d$ belonging to $\widehat{\mathcal{W}}_d := (\mathcal{V}_d(\mathbf{X}))^\perp$, where $(\mathcal{V}_d(\mathbf{X}))^\perp$ is the complemented subspace of $\mathcal{V}_d(\mathbf{X})$. For $c > 0$, let $\mathbf{Z}_c = (c\widehat{\mathbf{w}}_1, \ldots, c\widehat{\mathbf{w}}_d, \mathbf{x}_{d+1}, \ldots, \mathbf{x}_n)'$ be the contaminated data, and $\widehat{\mathbf{\Sigma}}(c) = \frac{1}{n}\mathbf{Z}_c'\mathbf{Z}_c = \frac{1}{n}\sum_{j=1}^p \widehat{\lambda}_{j,c}\widehat{\mathbf{v}}_{j,c}\widehat{\mathbf{v}}_{j,c}'$ be the contaminated sample covariance matrix satisfying $\widehat{\lambda}_{1,c} \geq \ldots, \geq \widehat{\lambda}_{p,c}$. Let $\widehat{\mathcal{V}}_d(c) = \mathcal{V}_d(\mathbf{Z}_c)$ be the PC subspace with $\mathbf{Z}_c$, $\widetilde{\mathcal{V}}_d(c)$ be the d-dimensional subspace spanned by $\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_d$, and $\widetilde{\mathbf{\Sigma}}_d(c) = \frac{1}{n}\sum_{j=1}^d c^2\widehat{\mathbf{w}}_j\widehat{\mathbf{w}}_j'$. By the variant of Davis-Kahan theorem by Yu et al. (2015), we have

$$\begin{aligned}\|\sin\Theta(\widehat{\mathcal{V}}_d(c), \widetilde{\mathcal{V}}_d(c))\|_{\mathrm{F}} &\leq \frac{2\|\widehat{\mathbf{\Sigma}}(c) - \widetilde{\mathbf{\Sigma}}_d(c)\|_{\mathrm{F}}}{c^2} \\ &\leq \frac{2\|\frac{1}{n}\sum_{i=d+1}^n \mathbf{x}_i\mathbf{x}_i'\|_{\mathrm{F}}}{c^2} \\ &\leq \frac{2M}{c^2},\end{aligned}$$

for some M which does not depend on c. For two d-dimensional subspace $\mathcal{V}$ and $\mathcal{U}$, let $\sin\overrightarrow{\Theta}(\mathcal{V},\mathcal{U}) := \mathrm{diag}(\sin\theta_1(\mathcal{V},\mathcal{U}), \ldots, \sin\theta_d(\mathcal{V},\mathcal{U}))$ be the diagonal matrix whose entries are sine of d principal angles. Note that $\widetilde{\mathcal{V}}_d(c)$ is orthogonal to $\widehat{\mathcal{V}}_d := \mathcal{V}_d(\mathbf{X})$, indicating that $\|\sin\overrightarrow{\Theta}(\widehat{\mathcal{V}}_d, \widetilde{\mathcal{V}}_d(c))\|_{\mathrm{F}} = d$. Thus we obtain

$$\begin{aligned}\|\sin\overrightarrow{\Theta}(\widehat{\mathcal{V}}_d, \widehat{\mathcal{V}}_d(c))\|_{\mathrm{F}} &\geq \|\sin\overrightarrow{\Theta}(\widehat{\mathcal{V}}_d, \widetilde{\mathcal{V}}_d(c))\|_{\mathrm{F}} - \|\sin\overrightarrow{\Theta}(\widehat{\mathcal{V}}_d(c), \widetilde{\mathcal{V}}_d(c))\|_{\mathrm{F}} \\ &\geq d - \frac{2M}{c^2}.\end{aligned}$$

The triangle inequality holds since $\|\sin\overrightarrow{\Theta}(\mathcal{V},\mathcal{U})\|_{\mathrm{F}} = \|\Pi_{\mathcal{V}} - \Pi_{\mathcal{U}}\|_{\mathrm{F}}/\sqrt{2}$ where $\Pi_{\mathcal{V}}$, and $\Pi_{\mathcal{U}}$ are the projection matrices of $\mathcal{V}$ and $\mathcal{U}$, respectively. Thus, for any $\epsilon > 0$, we can find a contaminated data $\mathbf{Z}_c$ such that $\|\sin\overrightarrow{\Theta}(\widehat{\mathcal{V}}_d, \widehat{\mathcal{V}}_d(c))\|_{\mathrm{F}} \geq d - \epsilon$ by taking a sufficiently large c. It means $\overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}) \leq \frac{d}{n}$.

$\|\Pi\mathbf{u}_0\| \leq \epsilon$, and $\|\mathbf{v}_0 - \mathbf{w}_0\| \leq \epsilon$ for some unit vector $\mathbf{w}_0 \in \widetilde{\mathcal{V}}_d^{\perp}$.

On the other hand, we show $\overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}) > \frac{d-1}{n}$. For a d-dimensional subspace $\mathcal{V}_d$, let $\Pi_{\mathcal{V}_d}$ be the projection matrix onto $\mathcal{V}_d$. Assume that, for any $\epsilon \in [0,1)$, there exists the contaminated data $\mathbf{Z}$ satisfying $\|\Pi_{\mathcal{V}_d(\mathbf{Z})}\Pi_{\mathcal{V}_d(\mathbf{X})}\| \leq \|\Pi_{\mathcal{V}_d(\mathbf{Z})}\Pi_{\mathcal{V}_d(\mathbf{X})}\|_{\mathrm{F}} < \epsilon$ with only $d-1$ contaminated data points. Note that this assumption is equivalent to $\overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}) \leq \frac{d-1}{n}$. Without loss of generality, let $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{d-1}, \mathbf{x}_d, \ldots, \mathbf{x}_n)'$. Let $\widetilde{\mathcal{V}}_d := \mathcal{V}_d(\mathbf{Z})$, $\widehat{\mathcal{V}}_d := \mathcal{V}_d(\mathbf{X})$, $\widetilde{\Pi} = \Pi_{\widetilde{\mathcal{V}}_d}$, and $\widehat{\Pi} = \Pi_{\widehat{\mathcal{V}}_d}$. We can find unit vectors $\mathbf{v}_0 \in \widehat{\mathcal{V}}_d$ and $\mathbf{u}_0 \in \widetilde{\mathcal{V}}_d$ such that $\mathbf{x}_i'\mathbf{v}_0 = 0$ and $\mathbf{z}_i'\mathbf{u}_0 = 0$ for all $i = 1, \ldots, d-1$ since $\widehat{\mathcal{V}}_d$ and $\widetilde{\mathcal{V}}_d$ are d-dimension and spanning subspace of $(d-1)$ data points are at most $(d-1)$-dimension. Then,

$$\widehat{\lambda}_d = \min_{\mathbf{v}\in\widehat{\mathcal{V}}_d, \|\mathbf{v}\|=1} \mathbf{v}'\mathbf{X}'\mathbf{X}\mathbf{v}/n \leq \mathbf{v}_0'\mathbf{X}'\mathbf{X}\mathbf{v}_0 = \mathbf{v}_0'\mathbf{X}_-'\mathbf{X}_-\mathbf{v}_0,$$

where $\mathbf{X}_- = [\mathbf{x}_{d+1}, \ldots, \mathbf{x}_n]' \in \mathbb{R}^{(n-d)\times p}$. Since $\|\widehat{\Pi}\widetilde{\Pi}\|_{\mathrm{F}} < \epsilon$, we know that

$$\begin{aligned}\|\widehat{\Pi}\mathbf{u}_0\| &= \|\widehat{\Pi}\widetilde{\Pi}\mathbf{u}_0\| \leq \epsilon,\\ \|\widetilde{\Pi}\mathbf{v}_0\| &= \|\widetilde{\Pi}\widehat{\Pi}\mathbf{v}_0\| \leq \epsilon,\end{aligned}$$

and for $\mathbf{w}_0 := \frac{(\mathbf{I}_p - \widetilde{\Pi})\mathbf{v}_0}{\|(\mathbf{I}_p - \widetilde{\Pi})\mathbf{v}_0\|} \in \widetilde{\mathcal{V}}_d^{\perp}$,

$$\begin{aligned}\|\mathbf{v}_0 - \mathbf{w}_0\|^2 &= 2 - 2\|(\mathbf{I}_p - \widetilde{\Pi})\mathbf{v}_0\|\\ &\leq 2 - 2(\|\mathbf{v}_0\| - \|\widetilde{\Pi}\mathbf{v}_0\|)\\ &\leq 2\epsilon.\end{aligned}$$

Let $\widetilde{\lambda}_j$ be the jth largest eigenvalue of $\frac{1}{n}\mathbf{Z}'\mathbf{Z}$. Then, we have

$$\begin{aligned}\widetilde{\lambda}_d &= \min_{\mathbf{v}\in\widetilde{\mathcal{V}}_d, \|\mathbf{v}\|=1} \mathbf{v}'\mathbf{Z}'\mathbf{Z}\mathbf{v}/n\\ &\leq \mathbf{u}_0'\mathbf{Z}'\mathbf{Z}\mathbf{u}_0/n\\ &= \mathbf{u}_0'\mathbf{X}_-'\mathbf{X}_-\mathbf{u}_0/n\\ &\leq \mathbf{u}_0'\mathbf{X}'\mathbf{X}\mathbf{u}_0/n\\ &= \mathbf{u}_0'(\widehat{\Pi} + \widehat{\Pi}^{\perp})\mathbf{X}'\mathbf{X}(\widehat{\Pi} + \widehat{\Pi}^{\perp})\mathbf{u}_0/n\\ &= \mathbf{u}_0'\widehat{\Pi}^{\perp}\mathbf{X}'\mathbf{X}\widehat{\Pi}^{\perp}\mathbf{u}_0/n + 2\mathbf{u}_0\widehat{\Pi}\mathbf{X}'\mathbf{X}\widehat{\Pi}^{\perp}\mathbf{u}_0/n + \mathbf{u}_0\widehat{\Pi}\mathbf{X}'\mathbf{X}\widehat{\Pi}\mathbf{u}_0/n\\ &\leq \|\widehat{\Pi}^{\perp}\mathbf{X}'\mathbf{X}\widehat{\Pi}^{\perp}/n\| + 2\|\widehat{\Pi}\mathbf{u}_0\| \cdot \|\mathbf{X}'\mathbf{X}/n\| \cdot \|\widehat{\Pi}^{\perp}\mathbf{u}_0\| + \|\widehat{\Pi}\mathbf{u}_0'\|^2 \cdot \|\mathbf{X}'\mathbf{X}/n\|\\ &\leq \widehat{\lambda}_{d+1} + 3\widehat{\lambda}_1\epsilon,\end{aligned}$$

where $\widehat{\Pi}^{\perp} = \mathbf{I}_p - \widehat{\Pi}$. Meanwhile,

$$\begin{aligned}\widetilde{\lambda}_{d+1} &= \max_{\mathbf{v}\in\widetilde{\mathcal{V}}_d^{\perp}, \|\mathbf{v}\|=1} \mathbf{v}'\mathbf{Z}'\mathbf{Z}\mathbf{v}/n\\ &\geq \max_{\mathbf{v}\in\widetilde{\mathcal{V}}_d^{\perp}, \|\mathbf{v}\|=1} \mathbf{v}'\mathbf{X}_-'\mathbf{X}_-\mathbf{v}/n\\ &\geq \mathbf{w}_0'\mathbf{X}_-'\mathbf{X}_-\mathbf{w}_0/n\\ &= (\mathbf{w}_0 - \mathbf{v}_0 + \mathbf{v}_0)'\mathbf{X}_-'\mathbf{X}_-(\mathbf{w}_0 - \mathbf{v}_0 + \mathbf{v}_0)/n\\ &= \mathbf{v}_0'\mathbf{X}_-'\mathbf{X}_-\mathbf{v}_0/n + 2(\mathbf{w}_0 - \mathbf{v}_0)'\mathbf{X}_-'\mathbf{X}_-\mathbf{v}_0/n + (\mathbf{w}_0 - \mathbf{v}_0)'\mathbf{X}_-'\mathbf{X}_-(\mathbf{w}_0 - \mathbf{v}_0)/n\\ &\geq \widehat{\lambda}_d - 4\widehat{\lambda}_1\epsilon - 4\widehat{\lambda}_1\epsilon^2\\ &\geq \widehat{\lambda}_d - 8\widehat{\lambda}_1\epsilon.\end{aligned}$$

It means, $\widetilde{\lambda}_d - \widetilde{\lambda}_{d+1} \leq \widehat{\lambda}_{d+1} - \widehat{\lambda}_d + 11\widehat{\lambda}_1\epsilon$. Since ϵ is arbitrary, we can find a $\mathbf{Z}$ such that $\widetilde{\lambda}_d - \widetilde{\lambda}_{d+1} \leq \widehat{\lambda}_{d+1} - \widehat{\lambda}_d + 11\widehat{\lambda}_1\epsilon < 0$ by taking sufficiently small ϵ. It is a contradiction. Thus $\overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}) = \frac{d}{n}$. □

A.2.2 Proof of Theorem 4

Proof. Denote $\mathrm{bp}(\mathcal{V}_d^{(r)}; \mathbf{X}) = \epsilon$. It implies that for any small $\epsilon_0 \in (0,1)$, there exists $\mathbf{X}_\epsilon$ such that $|I_\epsilon| := |\{1 \le i \le n : \mathbf{x}_i = \mathbf{x}_{i,\epsilon}\}| = (1-\epsilon)n$, and $\sin\widehat{\Theta} := \sin\Theta\left(\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon), \mathcal{V}_d^{(r)}(\mathbf{X})\right) \ge 1-\epsilon_0$. By Theorem 5,

$$1-\epsilon_0 \le \sin\widehat{\Theta} \le \frac{2r^2\epsilon}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}}.$$

Since ϵ_0 is arbitrarily small, we have

$$\mathrm{bp}(\mathcal{V}_d^{(r)}; \mathbf{X}) = \epsilon \ge \frac{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}}{2r^2}.$$

For strong breakdown point, let $\Pi \in \mathbb{R}^{p\times p}$ be the projection matrix of $\mathcal{V}_d(\mathbf{X}^{(r)})$ and $\Pi^\perp = (\mathbf{I}_p - \Pi)$. Denote $\overline{\mathrm{bp}}(\mathcal{V}_d^{(r)}; \mathbf{X}) = \epsilon$. By definition of $\overline{\mathrm{bp}}(\mathcal{V}_d^{(r)}; \mathbf{X})$, for any $\epsilon_0 \in (0,1)$, there exists $\mathbf{X}_\epsilon = [\mathbf{x}_{1,\epsilon}, \ldots, \mathbf{x}_{n,\epsilon}]' \in \mathbb{R}^{n\times p}$ such that $|I_\epsilon| := |\{1 \le i \le n : \mathbf{x}_{i,\epsilon} = \mathbf{x}_i\}| = (1-\epsilon)n$, and eigen bases $\mathbf{v}_{j,\epsilon}$ corresponding to the jth largest eigenvalue of $\frac{1}{n}\left(\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\right)$ and $\|\Pi\mathbf{v}_{j,\epsilon}\|_2 \le \epsilon_0$ for $j = 1, \ldots, d$. Then, for each $j = 1, \ldots, d$ we have

$$\begin{aligned}
\mathbf{v}_{j,\epsilon}^{\perp\prime}\left(\frac{1}{n}\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\right)\mathbf{v}_{j,\epsilon} &= \mathbf{v}_{j,\epsilon}'\left(\frac{1}{n}\sum_{i\notin I_\epsilon}\mathbf{x}_{i,\epsilon}^{(r)}\mathbf{x}_{i,\epsilon}^{(r)\prime}\right)\mathbf{v}_{j,\epsilon} + \mathbf{v}_{j,\epsilon}'\left(\frac{1}{n}\sum_{i\in I_\epsilon}\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}\right)\mathbf{v}_{j,\epsilon} \\
&\le r^2\epsilon + \mathbf{v}_{j,\epsilon}'(\Pi^\perp + \Pi)\left(\frac{1}{n}\sum_{i\in I_\epsilon}\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}\right)(\Pi^\perp + \Pi)\mathbf{v}_{j,\epsilon} \\
&\le r^2\epsilon + 3r^2(1-\epsilon)\epsilon_0 + \mathbf{v}_{j,\epsilon}'\left(\Pi^\perp\left(\frac{1}{n}\sum_{i\in I_\epsilon}\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}\right)\Pi^\perp\right)\mathbf{v}_{j,\epsilon} \\
&\le r^2\epsilon + 3r^2(1-\epsilon)\epsilon_0 + \mathbf{v}_{j,\epsilon}'\left(\Pi^\perp\left(\frac{1}{n}\mathbf{X}^{(r)\prime}\mathbf{X}^{(r)}\right)\Pi^\perp\right)\mathbf{v}_{j,\epsilon}.
\end{aligned}$$

Meanwhile,

$$\begin{aligned}
\mathbf{v}_{j,\epsilon}'\left(\frac{1}{n}\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\right)\mathbf{v}_{j,\epsilon} &= \eta_j\left(\frac{1}{n}\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\right) \\
&= \eta_j\left(\frac{1}{n}\left(\sum_{i=1}^n \mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}\right) + \frac{1}{n}\left(\sum_{i\notin I_\epsilon}\mathbf{x}_{i,\epsilon}^{(r)}\mathbf{x}_{i,\epsilon}^{(r)\prime} - \mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}\right)\right) \\
&\ge \eta_j\left(\frac{1}{n}\mathbf{X}^{(r)\prime}\mathbf{X}^{(r)}\right) - r^2\epsilon,
\end{aligned}$$

where $\eta_j(\mathbf{A})$ is the jth largest eigenvalue of $\mathbf{A} \in \mathbb{R}^{p\times p}$. Thus we have

$$\begin{aligned}
2r^2\epsilon + 3r^2(1-\epsilon)\epsilon_0 &\ge \eta_j\left(\frac{1}{n}\mathbf{X}^{(r)\prime}\mathbf{X}^{(r)}\right) - \mathbf{v}_{j,\epsilon}'\left(\Pi^\perp\left(\frac{1}{n}\mathbf{X}^{(r)\prime}\mathbf{X}^{(r)}\right)\Pi^\perp\right)\mathbf{v}_{j,\epsilon} \\
&= \widehat{\lambda}_j^{(r)} - \mathbf{v}_{j,\epsilon}'\left(\Pi^\perp\left(\frac{1}{n}\mathbf{X}^{(r)\prime}\mathbf{X}^{(r)}\right)\Pi^\perp\right)\mathbf{v}_{j,\epsilon} \\
&= \widehat{\lambda}_j^{(r)} - \mathrm{Tr}\left(\Pi^\perp\left(\frac{1}{n}\mathbf{X}^{(r)\prime}\mathbf{X}^{(r)}\right)\Pi^\perp\mathbf{v}_{j,\epsilon}\mathbf{v}_{j,\epsilon}'\right),
\end{aligned}$$

for $j = 1, \ldots, d$. Since $(\mathbf{v}_{1,\epsilon}, \ldots, \mathbf{v}_{d,\epsilon})'(\mathbf{v}_{1,\epsilon}, \ldots, \mathbf{v}_{d,\epsilon}) = \mathbf{I}_d$, for every $d_0 = 1, \ldots, d$,

$$\begin{aligned} d_0(2r^2\epsilon + 3r^2(1-\epsilon)\epsilon_0) &\geq \sum_{j=1}^{d_0} \left(\widehat{\lambda}_j^{(r)} - \mathrm{Tr}\left(\Pi^{\perp} \left(\frac{1}{n} \mathbf{X}^{(r)\prime} \mathbf{X}^{(r)} \right) \Pi^{\perp} \mathbf{v}_{j,\epsilon} \mathbf{v}_{j,\epsilon}' \right) \right) \\ &\geq \sum_{j=1}^{d_0} \widehat{\lambda}_j^{(r)} - \sup_{u_1, \ldots, u_{d_0}} \sum_{j=1}^{d_0} \mathrm{Tr}\left(\Pi^{\perp} \left(\frac{1}{n} \mathbf{X}^{(r)\prime} \mathbf{X}^{(r)} \right) \Pi^{\perp} \mathbf{u}_j \mathbf{u}_j' \right) \\ &= \sum_{j=1}^{d_0} \widehat{\lambda}_j^{(r)} - \sup_{u_1, \ldots, u_{d_0}} \mathrm{Tr}\left(\Pi^{\perp} \left(\frac{1}{n} \mathbf{X}^{(r)\prime} \mathbf{X}^{(r)} \right) \Pi^{\perp} \sum_{j=1}^{d_0} \mathbf{u}_j \mathbf{u}_j' \right) \\ &\geq \sum_{j=1}^{d_0} \widehat{\lambda}_j^{(r)} - \sum_{j=1}^{d_0} \widehat{\lambda}_{d+j}^{(r)}. \end{aligned}$$

where the supremum is taken over all possible $\mathbf{u}_1, \ldots, \mathbf{u}_{d_0}$ satisfying $(\mathbf{u}_1, \ldots, \mathbf{u}_{d_0})'(\mathbf{u}_1, \ldots, \mathbf{u}_{d_0}) = \mathbf{I}_{d_0}$. Here, the last inequality is obtained from Von Neumann's trace inequality. Since ϵ_0 is arbitrary, we have

$$\epsilon \geq \frac{\sum_{j=1}^{d_0} \widehat{\lambda}_j^{(r)} - \sum_{j=1}^{d_0} \widehat{\lambda}_{d+j}^{(r)}}{2r^2 d_0}.$$

where $\widehat{\lambda}_j^{(r)} = 0$ for $j > p$.

□

A.2.3 Proof of Theorem 5

Proof. We adopted two inequality in Yu et al. (2015); Cai and Zhang (2018). Let $\mathcal{I}_\epsilon = \{i : \mathbf{x}_i = \mathbf{x}_{i,\epsilon}\}$ and $\mathbf{X}_\epsilon^{(r)\prime} = [\mathbf{x}_{1,\epsilon}^{(r)}, \ldots, \mathbf{x}_{n,\epsilon}^{(r)}]$. By Lemma 10, we have

$$\begin{aligned} \sin \widehat{\Theta}_\epsilon^{(r)} &\leq \frac{2\| \frac{1}{n} \mathbf{X}^{(r)\prime} \mathbf{X}^{(r)\prime} - \frac{1}{n} \mathbf{X}_\epsilon^{(r)\prime} \mathbf{X}_\epsilon^{(r)\prime} \|}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}} \\ &= \frac{2\| \frac{1}{n} (\sum_{i \notin I_\epsilon} \mathbf{x}_i^{(r)} \mathbf{x}_i^{(r)\prime} - \mathbf{x}_{i,\epsilon}^{(r)} \mathbf{x}_{i,\epsilon}^{(r)\prime}) \|}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}} \\ &\leq \frac{2r^2\epsilon}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}}. \end{aligned}$$

For the second upper bound, $\sin \widehat{\Theta}_\epsilon^{(r)} \leq \frac{r^2\epsilon}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)} - 2r^2\epsilon}$, assume that $\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)} > 4r^2\epsilon$. Let $\widehat{\mathbf{W}} = [\widehat{\mathbf{W}}_d, \widehat{\mathbf{W}}_\perp]$ be the orthogonal matrix where $\widehat{\mathbf{W}}_d$ is the right singular vector corresponding to the d largest singular values of $\mathbf{X}^{(r)}$. Using the proposition in Cai and Zhang (2018), we have

$$\sin \widehat{\Theta}_\epsilon^{(r)} \leq \frac{\sigma_d(\mathbf{X}_\epsilon^{(r)} \widehat{\mathbf{W}}_d) \| \Pi_{(\mathbf{X}_\epsilon^{(r)} \widehat{\mathbf{W}}_d)} \mathbf{X}_\epsilon^{(r)} \widehat{\mathbf{W}}_\perp \|}{\sigma_d^2(\mathbf{X}_\epsilon^{(r)} \widehat{\mathbf{W}}_d) - \sigma_{d+1}^2(\mathbf{X}_\epsilon^{(r)})}$$

where $\sigma_j(\mathbf{A})$ is the jth largest singular value of $\mathbf{A}$, $\Pi_{(\mathbf{X}_\epsilon^{(r)} \widehat{\mathbf{W}}_d)}$ is the projection operator onto the column space of

$\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d$. The inequality holds when $\sigma_d^2(\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d) - \sigma_{d+1}^2(\mathbf{X}_\epsilon^{(r)}) > 0$ Then,

$$\begin{aligned}
\sigma_d^2(\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d) =& \sigma_d(\widehat{\mathbf{W}}_d'\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d)\\
\geq& \sigma_d(\widehat{\mathbf{W}}_d'\sum_{i\in I_\epsilon}\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}\widehat{\mathbf{W}}_d)\\
\geq& \sigma_d(\widehat{\mathbf{W}}_d'\sum_{i=1}^{n}\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}\widehat{\mathbf{W}}_d) - r^2m_n\\
=& n\widehat{\lambda}_d^{(r)} - r^2m_n,\\
\sigma_{d+1}^2(\mathbf{X}_\epsilon^{(r)}) =& \sigma_{d+1}(\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)})\\
\leq& \sigma_{d+1}(\sum_{i\in I_\epsilon}\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}) + r^2m_n\\
\leq& \sigma_{d+1}(\sum_{i=1}^{n}\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime}) + r^2m_n\\
=& n\widehat{\lambda}_{d+1}^{(r)} + r^2m_n,\\
\|\Pi_{(\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d)}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_\perp\|^2 =& \sigma_1(\widehat{\mathbf{W}}_\perp'\mathbf{X}_\epsilon^{(r)\prime}\Pi_{(\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d)}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_\perp)\\
=& \sigma_1(\widehat{\mathbf{W}}_\perp'\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d \times (\widehat{\mathbf{W}}_d'\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d)^{-1} \times \widehat{\mathbf{W}}_d'\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_\perp)\\
\leq& \frac{\sigma_1^2(\widehat{\mathbf{W}}_\perp'\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d)}{\sigma_d(\widehat{\mathbf{W}}_d'\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d)},\\
\sigma_1(\widehat{\mathbf{W}}_\perp'\mathbf{X}_\epsilon^{(r)\prime}\mathbf{X}_\epsilon^{(r)}\widehat{\mathbf{W}}_d) =& \sigma_1(\widehat{\mathbf{W}}_\perp'\mathbf{X}_n^{(r)\prime}\mathbf{X}_n^{(r)}\widehat{\mathbf{W}}_d - \widehat{\mathbf{W}}_\perp'\sum_{i\notin I_\epsilon}(\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime} - \mathbf{x}_{i,\epsilon}^{(r)}\mathbf{x}_{i,\epsilon}^{(r)\prime})\widehat{\mathbf{W}}_d)\\
\leq& \sigma_1(\widehat{\mathbf{W}}_\perp'\sum_{i\notin I_\epsilon}(\mathbf{x}_i^{(r)}\mathbf{x}_i^{(r)\prime} - \mathbf{x}_{i,\epsilon}^{(r)}\mathbf{x}_{i,\epsilon}^{(r)\prime})\widehat{\mathbf{W}}_d)\\
\leq& r^2m_n.
\end{aligned}$$

Thus we have

$$\sin\widehat{\Theta}_\epsilon^{(r)} \leq \frac{r^2\epsilon}{\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)} - 2r^2\epsilon}.$$

□

A.3 Empirical Expectation example generation in Section 3.1.1

Let $n = 200$, $p = 100$, $d = 1$, and $\mathbf{\Sigma} = \mathrm{diag}(100, 1, \ldots, 1)$. The data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently generated from the multivariate Gaussian distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$ or $t_3(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{0} = (0, \ldots, 0)' \in \mathbb{R}^p$. Here $\mathbf{\Sigma}$ in $t_3(\mathbf{0}, \mathbf{\Sigma})$ implies the covariance matrix of $t_3(\mathbf{0}, \mathbf{\Sigma})$.

For the outliers, we replace the first $m := 0.05n = 10$ data points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ with $\mathbf{z}_i = (0, 100np, 0, \ldots, 0)$ for $i = 1, \ldots, 0.05n$. Consequently, the contaminated data with m outliers is denoted by $\mathbf{X}_{m/n} = [\mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{x}_{m+1}, \ldots, \mathbf{x}_n]'$. We replicate the experiment 100 times.

A.4 Empirical Expectation example generation in Section 3.2

For $k = 1, \ldots, 4$, we set the dimension $p_k = 1000k$, and the sample size n_k to satisfy $p_k/n_k = 1/2k$. For the non-spiked model where the eigenvalues of the covariance matrix $\mathbf{\Sigma}$ are constant, $\mathbf{\Sigma} = \mathrm{diag}(2^2, 3^2)$. For the spiked model, we set $\mathbf{\Sigma} = \mathrm{diag}(2^2\sqrt{p_k}, 3^2\sqrt{p_k})$. The radii we concern are $r_{1k} = 1$, $r_{2k} = \sqrt{p_k}$, and $r_{3k} = \sqrt{p_k \log p_k}$. The data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently generated from the multivariate Gaussian distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$ or $t_3(\mathbf{0}, \mathbf{\Sigma})$. Here $\mathbf{\Sigma}$ in $t_3(\mathbf{0}, \mathbf{\Sigma})$ implies the covariance matrix of $t_3(\mathbf{0}, \mathbf{\Sigma})$.

For the outliers, we replace the first $m := 2$ data points $\mathbf{x}_1, \mathbf{x}_2$ with $\mathbf{z}_i = (\mathbf{0}_d', n_k p_k, \mathbf{0}_{p_k-d-1}')$ for $i = 1, 2$. Here $\mathbf{0}_l$ is the zero vector with the length l. Consequently, the contaminated data with m outliers is denoted by $\mathbf{X}_{m/n} = [\mathbf{z}_1, \mathbf{z}_2, \mathbf{x}_3, \ldots, \mathbf{x}_n]'$. We replicate the experiment 10 times.

A.5 Data Generation for Estimated Lower Bounds in Section 4.2

Let $n = 1000$, $p = 4$, $d = 2$, and $\mathbf{\Sigma} = \text{diag}(5^2, 5^2, 5, 1)$. The data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently generated from the multivariate Gaussian distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{0} = (0, 0)'$. The data matrix is represented as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]' \in \mathbb{R}^{n \times p}$. We replicate the experiment 1000 times.

A.6 Toy example generation in Section 4.3

Let $n = 1000$, $p = 2$, $d = 1$, and $\mathbf{\Sigma} = \text{diag}(5^2, 1)$. The data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independently generated from the bivariate Gaussian distribution $N_p(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{0} = (0, 0)'$. The data matrix is represented as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]' \in \mathbb{R}^{n \times p}$. Define $r = \text{med}_i \|\mathbf{x}_i\|_2$. To introduce outliers, we replace the first m data points $\mathbf{x}_1, \ldots, \mathbf{x}_m$ with $\mathbf{z}_i = (0, \max_i \|\mathbf{x}_i\|_2^2 + 100)$ for $i = 1, \ldots, m$. Note that the outliers need not be excessively large, as we are only concerned with the median radius. Consequently, the contaminated data with m outliers is denoted by $\mathbf{X}_{m/n} = [\mathbf{z}_1, \ldots, \mathbf{z}_m, \mathbf{x}_{m+1}, \ldots, \mathbf{x}_n]'$.