Subspace Recovery in Winsorized PCA: Insights into Accuracy and Robustness

Sangil Han [1] Kyoowon Kim [1] Sungkyu Jung [1]

[1]Seoul National University

Challenge and Motivation for WPCA (Winsorized PCA)

- Challenge
  Real-world data often contain noise, outliers, and other anomalies.
  The standard SVD (Singular Value Decompostion) and PCA (Principal Component Analysis) estimates are highly sensitive to large scale outliers.
- Motivation
  Winsorization has long been recognized as a common and effective tool for handling extreme values in data analysis. It can be applied universally before performing analysis, ensuring that subsequent analysis 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.
- *Winsorized PCA* (WPCA): a robust approximation of the d dimensional PC subspace, spanned by the first d eigenvectors of the population covariance matrix.

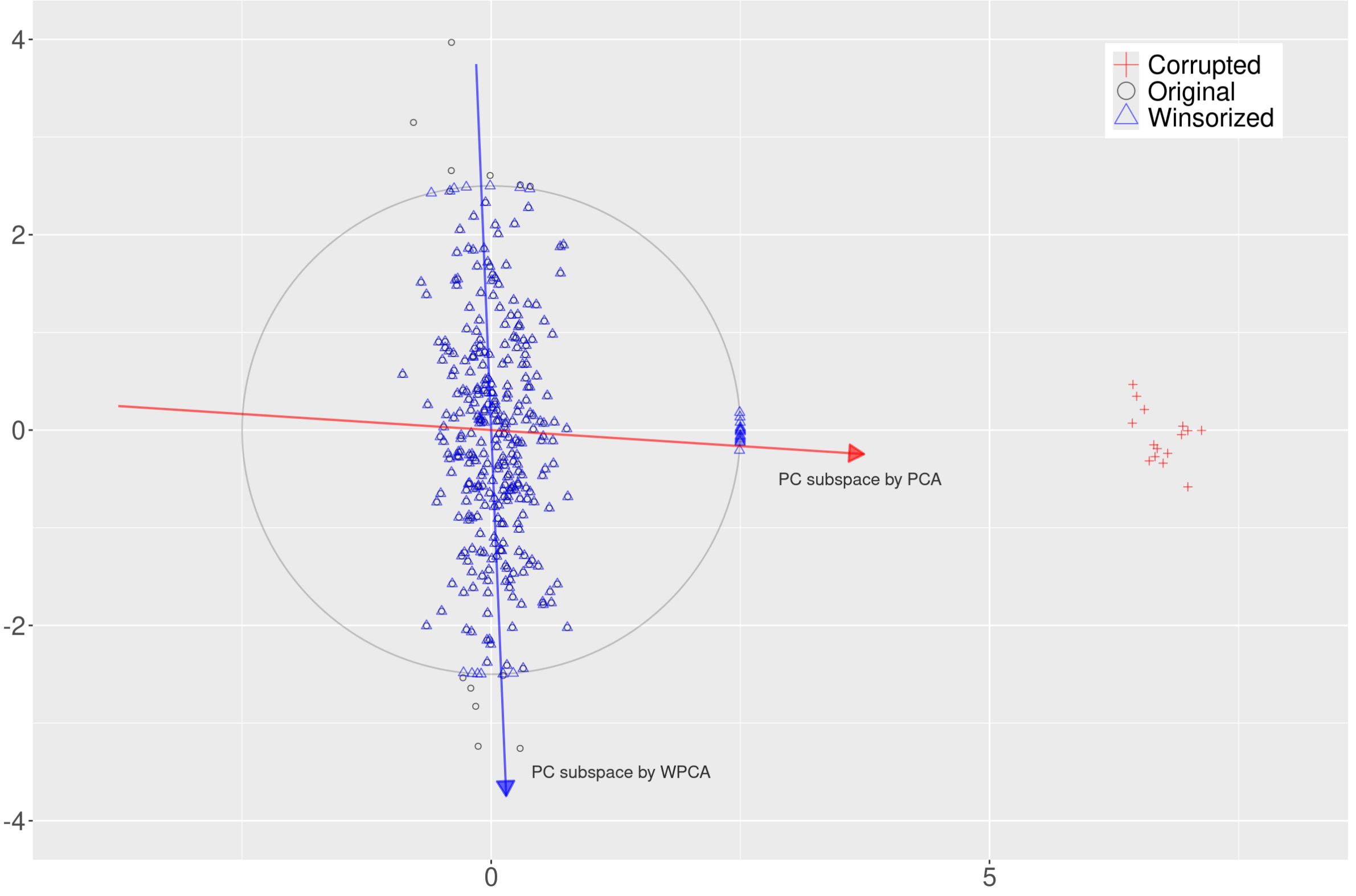


Figure 1. Recovered PC subspace by WPCA, and corrupted PC subspace by the standard PCA.

Winsorized PCA

$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]' \in \mathbb{R}^{n \times p}$: a centered, potentially contaminated data matrix consisting of n samples with p variables. $\xrightarrow{\text{PC subspace}} \mathcal{V}_d$

$\Big\Downarrow$ Winsorization with radius $r \in (0, \infty)$

$\mathbf{X}^{(r)} = [\mathbf{x}_1^{(r)}, \dots, \mathbf{x}_n^{(r)}]'$: Winsorized data where each winsorized observation is defined as: $\xrightarrow{\text{PC subspace}} \mathcal{V}_d^{(r)}$

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

Statistical Accuracy Theorem

Let $\mathcal{F}_{\mathbf{\Sigma}}$ denote a mean-zero, p-dimensional elliptical distribution (Cambanis et al., 1981), with covariance matrix $\mathbf{\Sigma} = \sum_{j=1}^{p} \lambda_j \mathbf{v}_j \mathbf{v}_j^T$. Assume that $\lambda_{j+1} > \lambda_j$ for all j and denote $\lambda_j^{(r)}$ the jth largest eigenvalue of Cov($\mathbf{x}^{(r)}$) where $\mathbf{x}^{(r)}$ is the winsorized random vector of $\mathbf{x} \sim \mathcal{F}_{\mathbf{\Sigma}}$. Let $\epsilon > 0$ be the fraction of contamination among n samples. Let $\mathcal{V}_d^{(\epsilon)}(\mathbf{X}_\epsilon)$ be the d-dimensional subspace spanned by the first d eigenvectors obtained by the proposed WPCA and $\Theta_\epsilon^{(r)} = \Theta(\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon), \mathcal{V}_d)$ be the largest principal angle between $\mathcal{V}_d^{(r)}(\mathbf{X}_\epsilon)$ and $\mathcal{V}_d$. Then for any n and p,

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

Moreover, if $\mathbf{y} := \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{x}$ is σ-subgaussian, then

$$\mathbb{E}[\sin \Theta_\epsilon^{(r)}] \le \frac{2r^2\epsilon}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}} + \frac{2^8\lambda_1(\frac{r^2}{p\lambda_p} \wedge \sigma^2)(\sqrt{\frac{p}{n}} \vee \frac{p}{n})}{\lambda_d^{(r)} - \lambda_{d+1}^{(r)}}. \tag{3}$$

Fixing the number of variables p, we define $g(n,r) = \Omega(h(n,r))$ if there exist constants $a \le b$ such that $a \le g(n,r)/h(n,r) \le b$. The convergence rates of the upper bounds in equations (2) and (3) can then be summarized as follows.

p fixed $n, r \to \infty$	heavy tail	light tail (subgaussian)
$\epsilon = 0$	$\Omega\left(r^2/\sqrt{n}\right)$	$\Omega\left(1/\sqrt{n}\right)$
$\epsilon > 0$	$\Omega\left(r^2\epsilon + r^2/\sqrt{n}\right)$	$\Omega\left(r^2\epsilon + 1/\sqrt{n}\right)$

Table 1. Convergence rate of the upper bound of $\mathbb{E}[\sin \Theta_\epsilon^{(r)}]$ when p is fixed

We now consider a high-dimensional setting where the number of variables p increases. To analyze the scenario where n, p, and r grow together, we assume $r = p^{1/2+\beta}$ with $\beta \in (-\infty, \infty)$. Here, a positive β implies fewer projected points, while negative β means more projected points. The convergence rate under this assumption can be summarized as follows.

$r = p^{1/2+\beta}$ $n, r \to \infty$	heavy tail	light tail (subgaussian)
$\epsilon = 0$	$\Omega\left(p^{2(\beta \vee 0)}\left(\sqrt{\frac{p}{n}} \vee \frac{p}{n}\right)\right)$	$\Omega\left(\sqrt{\frac{p}{n}} \vee \frac{p}{n}\right)$
$\epsilon > 0$	$\Omega\left(p^{1+2(\beta \vee 0)}\epsilon\right)$	

Table 2. Convergence rate of the upper bound of $\mathbb{E}[\sin \Theta_\epsilon^{(r)}]$ in the high-dimensional regime

Extension of Breakdown Points

- The breakdown point (Hampel, 1968; Huber and Donoho, 1983; Huber, 1984, 2011) can be defined as the minimum number of corrupted data points that cause the the furthest statistic value from the original statistic value.
- Breakdown point of real-valued statistics
  For a real-valued statistic as a function $f : \mathcal{X}^n \to \mathbb{R}$ that takes as input n data points $\mathbf{X}_0 \in \mathcal{X}^n$ and outputs a real-valued $f(\mathbf{X}_0)$, the breakdown point of f at $\mathbf{X}_0$ is

$$\mathrm{bp}(f; \mathbf{X}_0) := \min_{1 \le l \le n} \{\frac{l}{n} : \sup_{\mathbf{Z}_l} |f(\mathbf{Z}_l) - f(\mathbf{X}_0)| = \infty\}, \tag{4}$$

  where the supremum is taken over all possible corrupted collections $\mathbf{Z}_l$ that are obtained from $\mathbf{X}_0$ by replacing l data points of $\mathbf{X}_0$ with arbitrary values.

$\Big\Downarrow f \to \mathbf{v}$ extend to vector valued

- Breakdown point of vector-valued statistics
  For a vector-valued statistics, $\mathbf{v} : \mathcal{X}^n \to S^{k-1}$,

$$\mathrm{bp}(\mathbf{v}; \mathbf{X}_0) := \min_{1 \le l \le n} \{\frac{l}{n} : \sup_{\mathbf{Z}_l} \theta(\mathbf{v}(\mathbf{Z}_l), \mathbf{v}(\mathbf{X}_0)) = \frac{\pi}{2}\}, \tag{5}$$

  where $\theta(\mathbf{v}, \mathbf{w}) := \arccos(|\mathbf{v}^T\mathbf{w}|)$ is the angle between $\mathbf{v}$ and $\mathbf{w}$.

$\Big\Downarrow \mathbf{v} \to \mathcal{V}$ extend to subspace valued

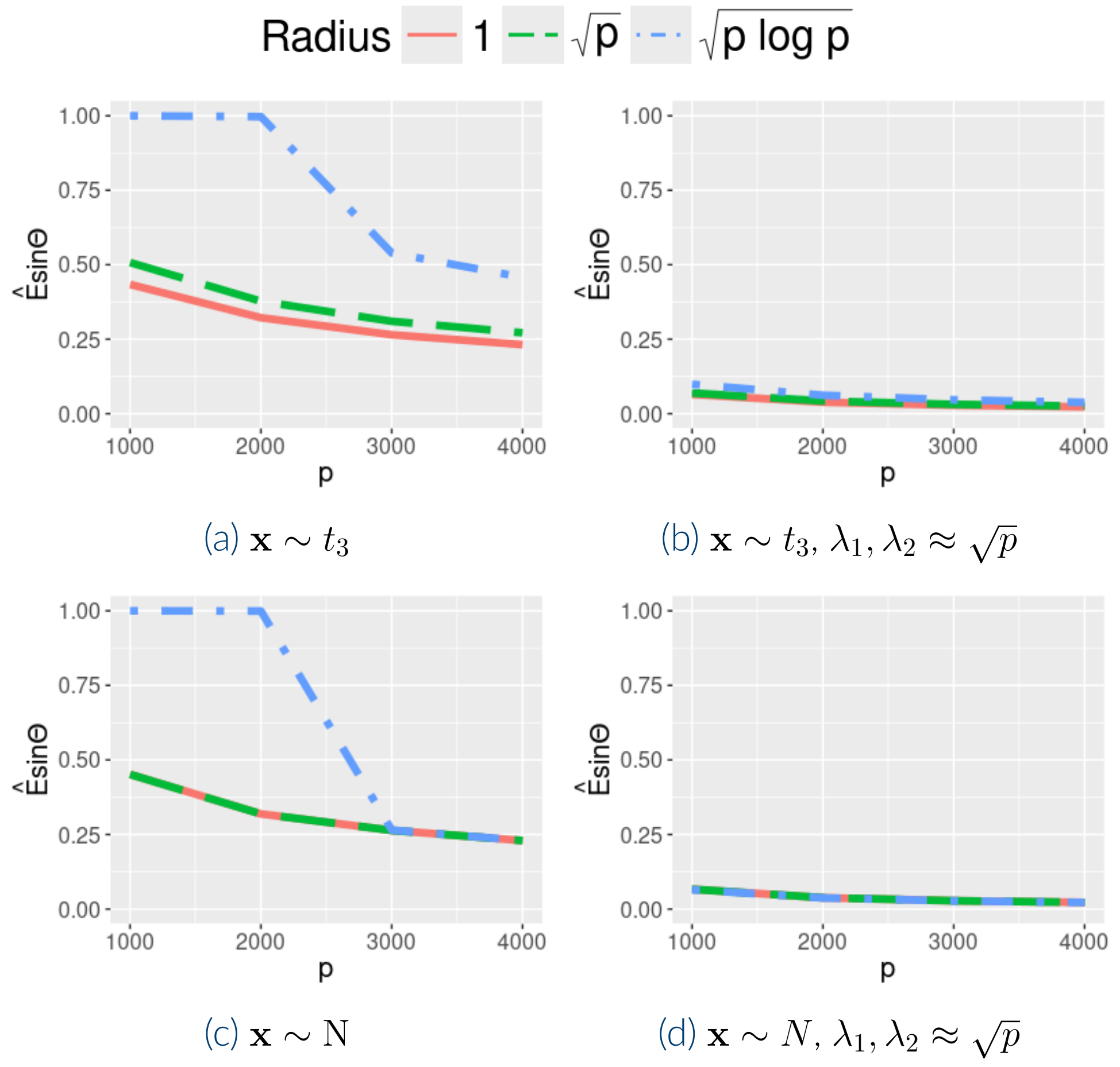
- Breakdown point of subspace-valued statistics
  For a subspace-valued statistic $\mathcal{V} : \mathcal{X}^n \to \mathrm{Gr}(d, p)$,
  (i) The breakdown point is defined as

$$\begin{aligned} \mathrm{bp}(\mathcal{V}; \mathbf{X}_0) &:= \mathrm{bp}(\mathcal{V}, \Theta; \mathbf{X}_0) \\ &= \min_{1 \le l \le n} \left\{\frac{l}{n} : \sup_{\mathbf{Z}_l} \Theta\big(\mathcal{V}(\mathbf{Z}_l), \mathcal{V}(\mathbf{X}_0)\big) = \frac{\pi}{2}\right\} \end{aligned} \tag{6}$$

  (ii) The strong breakdown point is defined as

$$\begin{aligned} \overline{\mathrm{bp}}(\mathcal{V}; \mathbf{X}_0) &:= \mathrm{bp}(\mathcal{V}, \theta; \mathbf{X}_0) \\ &= \min_{1 \le l \le n} \left\{\frac{l}{n} : \sup_{\mathbf{Z}_l} \theta\big(\mathcal{V}(\mathbf{Z}_l), \mathcal{V}(\mathbf{X}_0)\big) = \frac{\pi}{2}\right\} \end{aligned} \tag{7}$$

  where $\Theta(\cdot, \cdot)$ and $\theta(\cdot, \cdot)$ denote the largest and smallest principal angles, respectively.



(a) $\mathbf{x} \sim t_3$

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

(c) $\mathbf{x} \sim \mathrm{N}$

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

Figure 2. Empirical expectation $\widehat{\mathbb{E}}[\sin \Theta_\epsilon^{(r)}]$ 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.

The lack of Robustness of PCA

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 jth largest eigenvalue of $\mathbf{X}_0'\mathbf{X}_0/n$. Assume that $\widehat{\lambda}_d > \widehat{\lambda}_{d+1}$. Then,

$$\begin{aligned} \mathrm{bp}(\mathcal{V}_d; \mathbf{X}_0) &= \frac{1}{n}, \\ \overline{\mathrm{bp}}(\mathcal{V}_d; \mathbf{X}_0) &= \frac{d}{n}. \end{aligned} \tag{8}$$

This result highlights that traditional PCA is highly sensitive to outliers.

Robustness of WPCA

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

$$\begin{aligned} \mathrm{bp}(\mathcal{V}_d^{(r)}; \mathbf{X}_0) &\ge \frac{1}{2r^2}(\widehat{\lambda}_d^{(r)} - \widehat{\lambda}_{d+1}^{(r)}), \\ \overline{\mathrm{bp}}(\mathcal{V}_d^{(r)}; \mathbf{X}_0) &\ge \sup_{d_0 \le d} \frac{\sum_{j=1}^{d_0} \widehat{\lambda}_j^{(r)} - \sum_{j=1}^{d_0} \widehat{\lambda}_{d+j}^{(r)}}{2r^2 d_0}. \end{aligned} \tag{9}$$

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.

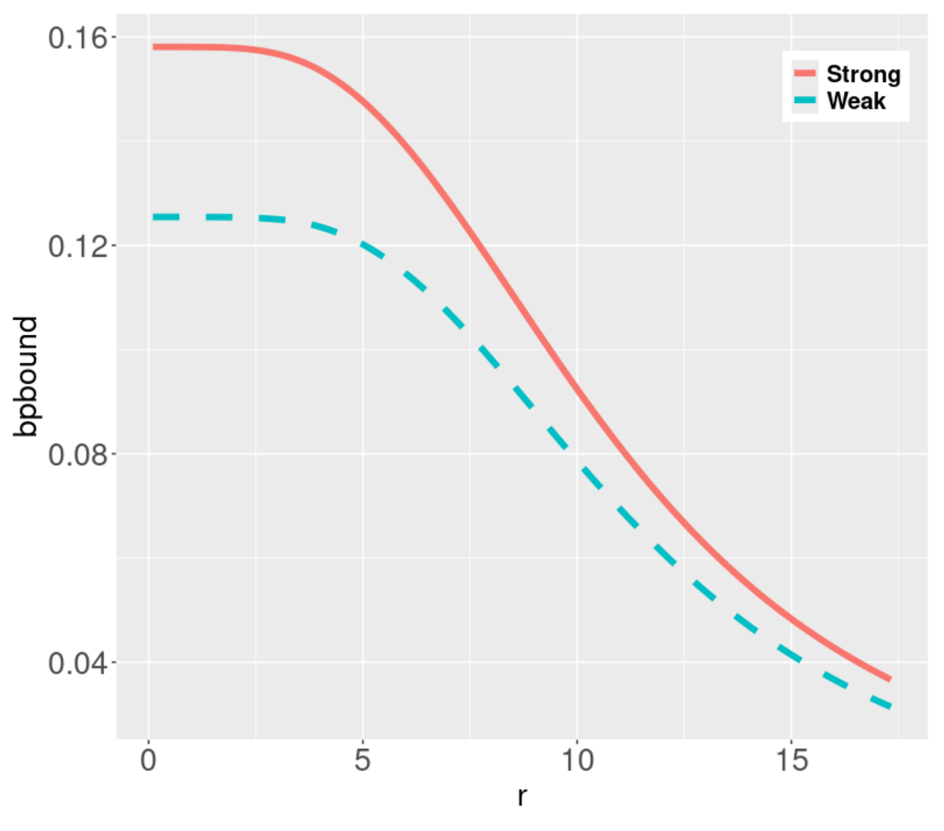


Figure 3. Estimated lower bounds for the breakdown points in (9).