Statistica Sinica Preprint No: SS-2022-0142	
Title	Outlier Detection via a Minimum Ridge Covariance Determinant Estimator
Manuscript ID	SS-2022-0142
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0142
Complete List of Authors	Chikun Li, Baisuo Jin and Yuehua Wu
Corresponding Authors	Baisuo Jin
E-mails	jbs@ustc.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

OUTLIER DETECTION VIA A MINIMUM RIDGE COVARIANCE DETERMINANT ESTIMATOR

Chikun Li, Baisuo Jin and Yuehua Wu

University of Science and Technology of China and York University

Abstract: In this paper, we propose an outlier detection procedure based on a high-breakdown minimum ridge covariance determinant estimator, which is especially for the large p/n scenario. The estimator is obtained from the subset of observations after excluding potential outliers by applying the so-called concentration steps. We explore the asymptotic distribution of the modified Mahalanobis distance related to the proposed estimator under certain moment conditions, and obtain theoretical cut-off value for outlier identification. We achieve a further improvement on the outlier detection power by adding a one-step reweighting procedure. We investigate the performance of the proposed methods by simulations and a real data analysis.

Key words and phrases: High dimension, minimum covariance determinant estimator, random matrices.

1. Introduction

Data collected in a broad range of applications frequently contain one or more atypical observations, known as outliers; that is, observations that are well-separated from the majority or bulk of the data, or in some way deviate from the general pattern of the data (Maronna et al., 2019). Outliers can not be avoided due to various reasons, and a regular data set may contain 1% to 10% of outliers, or much more than that in some specific applications (Hampel et al., 2005). Due to the ever reducing cost of collecting data, many datasets in current applications are both large and complex, sometimes with a very high number of variables. The chance of contamination or

other type of imperfections in the data increases both with the number of observations and their dimension. Thus, detecting potential outliers is either an important preprocessing step, to avoid model misspecification and biased parameter estimation, or for some specific interest in finding anomalous observations.

Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a random sample of the p-dimensional random vector $\boldsymbol{X}$ with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^\top$ and covariance matrix $\Sigma_p = (\sigma_{ij})_{p\times p}$. A common measure of outlyingness for an individual observation $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is the Mahalanobis distance

$$d_i^2(\boldsymbol{\mu}, \Sigma_p) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\top \Sigma_p^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) . \tag{1.1}$$

The famous minimum covariance determinant algorithm (Rousseeuw and Van Driessen, 1999) was designed to search for a subsequence of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ of size h with $n/2 < h < n$ whose sample covariance matrix has the the smallest determinant. Thus it is able to obtain reliable estimates of $\boldsymbol{\mu}$ and Σ_p via this subsequence for computing (1.1). To determine the cut-off value for outlying points, Hardin and Rocke (2005) presented a distributional result of (1.1) under Gaussian assumption which is superior to the commonly used chi-square cutoff. The consistency and asymptotic normality of the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985) were shown by Cator and Lopuhaä (2012). Based upon the small sample correction factors constructed by Pison et al. (2002), Cerioli (2010) proposed an iterated reweighted-MCD procedure which performs well for multiple outlier detection.

However, when p/n increases, conventional outlier detection methods which take advantage of the MCD estimator become infeasible and suffer power loss (Adrover and Yohai, 2002; Alqallaf et al., 2009). In fact, the MCD approach is often recommended when $n > 5p$ (Boudt et al., 2019). For outlier detection, Filzmoser et al. (2008) developed a computationally fast procedure by utilizing the principal component analysis to identify outliers in a transformed space when

$p/n \geq 1$. Ro et al. (2015) introduced the following alternation for (1.1):

$$d_i^2(\boldsymbol{\mu}, D_p) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\top D_p^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{1.2}$$

where $D_p = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. By replacing (1.1) with (1.2), they proposed a computationally efficient refined minimum diagonal product algorithm and made simulation studies for autoregressive correlation and moving average models. Li and Jin (2022) considered another alternation for (1.1):

$$d_i^2(\boldsymbol{\mu}, D_\Sigma) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\top D_\Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{1.3}$$

where D_Σ is the 2×2 block diagonal partition of Σ_p. They thereby developed a high-breakdown block diagonal product estimator. A few other outlier detection techniques for high-dimensional data based on (1.2) have also been proposed, including those of Yang et al. (2018) and Wang et al. (2021).

Let $\bar{\boldsymbol{x}}_n = n^{-1} \sum_{i=1}^n \boldsymbol{x}_i$, and $S_n = n^{-1} \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_n)^\top$. Denote I_p the $p \times p$ identity matrix. Motivated by the regularized Hotelling's T^2 test statistic used in the high-dimensional mean test (Chen et al., 2011; Ha et al., 2021), in this paper, we modify (1.1) as

$$d_i^2(\boldsymbol{\mu}, S_n(\lambda)) = (\boldsymbol{x}_i - \boldsymbol{\mu})^\top [S_n(\lambda)]^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}), \tag{1.4}$$

where $S_n(\lambda) = S_n + \lambda I_p$, and $\lambda > 0$ is a scalar tuning parameter. Here, the product λI_p is the perturbation that we add to the covariance estimator S_n such that the matrix $S_n(\lambda)$ is positive definite and hence invertible. Boudt et al. (2019) suggested adding a preprocessing step to standardize each $\boldsymbol{x}_i$ as

$$\boldsymbol{u}_i = D_X^{-1} (\boldsymbol{x}_i - \boldsymbol{v}_X),$$

where D_X is a diagonal matrix whose jth diagonal element is the Q_n estimator (Rousseeuw and Croux, 1993), and $\boldsymbol{v}_X$ is a location vector whose elements consist of the medians of all the

variables. Then, they defined the regularized sample covariance matrix by

$$K = \rho T + (1 - \rho)c_\alpha S_U,$$

where S_U is the sample covariance matrix of $U = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n\}$, T is a predetermined positive definite target matrix, ρ is a regularization parameter selected to bound the condition number of K, and c_α is the consistency factor defined by Croux and Haesbroeck (1999). However, there is no distributional result or reweighting step in Boudt et al. (2019), and it is not easy to obtain appropriate standardized observations in high-dimensional settings.

In this paper, we consider an outlier detection, which is especially for the large p/n scenario, where $c_1 \leq p/n \leq c_2$ with c_1 and c_2 being some positive constants. By relaxing the Gaussian assumption, we derive the exact distribution of (1.4). We then propose a high-breakdown minimum ridge covariance determinant estimator. We explore the asymptotic distribution of the modified Mahalanobis distance related to the proposed estimator under certain moment conditions, and obtain theoretical cut-off value for outlier identification, which is the basis for a proposed outlier detection procedure. We achieve a further improvement on the outlier detection power by adding a one-step reweighting procedure. We show that the proposed procedure achieves higher detection power against sparse signals in most situations than its main competitors through simulation studies and a real data analysis.

The remainder of the paper is organized as follows. In Section 2, we give our model assumptions, introduce the minimum ridge covariance determinant estimator, and present the main results. In Section 3, we examine the performance of the proposed methods by simulations and a real data analysis. We conclude the paper in Section 4. All theoretical proofs are provided in the appendix.

2. Methods and properties

2.1 Model assumptions

Let $\boldsymbol{X}_n$ be a p_n-dimensional random vector admitting the independent components model

$$\boldsymbol{X}_n = T_{p_n}\boldsymbol{Z}_n + \boldsymbol{\mu}_n, \tag{2.5}$$

where $\boldsymbol{\mu}_n = (\mu_{1,n}, \ldots, \mu_{p_n,n})^\top$ is the location vector, T_{p_n} is a $p_n \times p_n$ full rank transformation matrix, and $\boldsymbol{Z}_n$ is a p_n-dimensional random vector with independently and identically distributed (i.i.d.) components. Denote the jth component of $\boldsymbol{Z}_n$ by $z_{j,n}$. For simplicity, we suppress the subscript n in the above notations if there is no confusion in the context.

Let F^{Σ_p} denote the empirical spectral distribution (ESD) of a matrix Σ_p (Bai and Silverstein, 2010), i.e.,

$$F^{\Sigma_p}(u) = \frac{1}{p}\sum_{j=1}^{p} \mathrm{I}_{[\lambda_j,\infty)}(u),$$

where λ_j, $j = 1, \ldots, p$, are the eigenvalues of Σ_p, and $\mathrm{I}_A(\cdot)$ denotes the indicator function of the set A.

Our main assumptions are listed as follows:

Condition A1. $p, n \to \infty$ such that $c_n \triangleq p/n \to c \in (0, \infty)$.

Condition A2. $\Sigma_p \triangleq T_p T_p^\top$ is a $p \times p$ positive definite matrix.

Condition A3. F^{Σ_p} converges to a proper probability measure F as $p \to \infty$.

Condition A4. $\limsup_{p\to\infty} \|\Sigma_p\| < \infty$ and $\limsup_{p\to\infty} \left\|\Sigma_p^{-1}\right\| < \infty$, where $\|\cdot\|$ denotes the spectral norm.

Condition A5. The first four moments of z_1 match those of standard normal distribution $N(0, 1)$.

Conditions A1-A4 are commonly used in the related research on the ESD of a high-dimensional sample covariance matrix, e.g., Chen et al. (2011) and Ha et al. (2021). The four-moment matching condition in Condition A5 is required to obtain the limiting distribution of (1.4). Condition A5 is most closely related to the Four Moment theorem for random covariance matrices in the work of Tao and Vu (2012). The first and second moment conditions of z_1 are easy to meet in practice. The third moment condition is necessary for some of our lemmas, especially Lemma A.2. The fourth moment is essential for the proof of Lemma A.4 in the appendix. Extending the theoretical results with relaxed Condition A5 can be an avenue for future research.

2.2 The minimum ridge covariance determinant estimate

The classical minimum covariance determinant procedure is intended to find a subset of observations whose sample covariance matrix has the smallest determinant, through iteratively computing and sorting the Mahalanobis distances of each observation. To generalize this procedure to high-dimensional datasets, our method searches for a subset of h observations such that the determinant of the ridge sample covariance matrix is minimal. The detail is as follows.

Let $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be a collection of n observations of $\boldsymbol{X}_n$ in (2.5). Define $\mathcal{H} = \{H \subset \{1, \ldots, n\} : |H| = h,\ h > n/2\}$, the collection of all subsets of size h, where $|H|$ denotes the cardinality of H. We set $h > n/2$ since the potential outliers account for no more than fifty percent of the total observations. For any $H \in \mathcal{H}$, denote $\bar{\boldsymbol{x}}_H = |H|^{-1} \sum_{i \in H} \boldsymbol{x}_i$,

$$S_H = |H|^{-1} \sum_{i \in H} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_H)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_H)^\top,$$

and $S_H(\lambda) = S_H + \lambda I_p$ with $\lambda \in (0, \infty)$, a ridge sample covariance matrix in terms of $\{\boldsymbol{x}_i,\ i \in H\}$. It is easy to see that $S_H(\lambda)$ for a given λ is positive definite.

Definition 1. *The minimum ridge covariance determinant (abbreviated as RICD) estimate of*

$\boldsymbol{\mu}$, the multivariate location parameter, for a given $\lambda > 0$ is defined as

$$\hat{\boldsymbol{\mu}}_{\mathrm{RICD}} = \bar{\boldsymbol{x}}_{H_{\mathrm{RICD}}} \text{ with } H_{\mathrm{RICD}} = \arg\min_{H \in \mathcal{H}} \det[S_H(\lambda)]. \tag{2.6}$$

It is noted that for $p > h$, the MCD estimate (Rousseeuw, 1985) becomes ill-defined since $\det[S_H] = 0$ for such H. Denote the scatter estimate of Σ_p by $\hat{\Sigma}_{\mathrm{RICD}} = S_{H_{\mathrm{RICD}}}(\lambda)$. We remark that $\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}$ and $\hat{\Sigma}_{\mathrm{RICD}}$ can be shown to be location invariant and orthogonal equivariant but not affine equivariant. See (Lopuhaä and Rousseeuw, 1991) for the definitions of location invariance, orthogonal equivariance, and affine equivariance of a covariance estimate.

When $\mathcal{X}$ is contaminated, there exist one or more $\boldsymbol{x}_i$s that are not observations of $\boldsymbol{X}_n$ in (2.5). These $\boldsymbol{x}_i$s may be arbitrary values, or go to ∞ as $n \to \infty$. Thus, $\bar{\boldsymbol{x}}_n$ is no longer an appropriate estimate of $\boldsymbol{\mu}$ and $\|\bar{\boldsymbol{x}}_n\|_F$ may be arbitrary large so that it "breaks down", where $\|\cdot\|_F$ denotes the Frobenius norm. The finite-sample breakdown point (Maronna et al., 2019) ε_n of an estimate $\hat{\boldsymbol{\theta}}_n$ of the parameter $\boldsymbol{\theta}$ is the smallest proportion of observations from $\mathcal{X}$ that need to be replaced by arbitrary values to carry $\hat{\boldsymbol{\theta}}_n$ beyond all bounds:

$$\varepsilon_n(\hat{\boldsymbol{\theta}}_n, \mathcal{X}) = \min_{1 \le t \le n} \left\{ t/n : \sup_{\tilde{\mathcal{X}}} \left\| \hat{\boldsymbol{\theta}}_n(\mathcal{X}) - \hat{\boldsymbol{\theta}}_n\left(\tilde{\mathcal{X}}\right) \right\|_F = \infty \right\},$$

where $\tilde{\mathcal{X}} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n\}$ is a dataset having at least $(n-t)$ elements in common with $\mathcal{X}$, i.e., $|\mathcal{X} \cap \tilde{\mathcal{X}}| \ge n - t$. It is easy to see that $\varepsilon_n(\bar{\boldsymbol{x}}_n, \mathcal{X}) = 1/n$. For the finite-sample breakdown point of the proposed estimates, we have the following theorem.

Theorem 1. *Suppose that $n/2 < h < n$ and $\lambda > 0$. Then we have*

$$\varepsilon_n\left(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}, \mathcal{X}\right) = \varepsilon_n(\hat{\Sigma}_{\mathrm{RICD}}, \mathcal{X}) = \min\{(n-h+1)/n, 0.5\}. \tag{2.7}$$

Theorem 1 shows that the proposed estimates can achieve the highest breakdown value, i.e., 50%, when $h = [n/2]+1$, where $[a]$ denotes the integer part of a. To achieve the best performance

in practice while ensuring $\varepsilon_n\left(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}, \mathcal{X}\right)$ and $\varepsilon_n(\hat{\Sigma}_{\mathrm{RICD}}, \mathcal{X})$ to be as high as possible, a default choice of $h_{\mathrm{default}} = [n/2] + 1$ is recommended.

To find H_{RICD} defined in (2.6), we modify the fast minimum covariance determinant algorithm (Rousseeuw and Van Driessen, 1999) by replacing the Mahalanobis distance with its high-dimensional counterpart (1.4). However, one will encounter a problem for the case that $n < p + 1$ as the original algorithm requires a random initial subset H_{ini} containing $p + 1$ data points sampled from $\mathcal{X}$. To solve it, we can set the size of a random initial subset to be $h_{\mathrm{ini}} = h_{\mathrm{default}}$ in view that $\varepsilon_n\left(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}, \mathcal{X}\right)$ does not depend on p.

Similar to Rousseeuw and Van Driessen (1999), we refer to the construction in the following theorem as a concentration step consisting of two parts. This theorem illustrates the function of the second part of the concentration step, i.e., sorting the distances of all $\boldsymbol{x}_i$ to the center of the subset obtained in the first part. By performing this part in the concentration step, we can obtain a more concentrated h-sized subset with lower possibility to be contaminated by atypical points, which guarantees that an iteration process of repeating concentration steps leads to an optimal H, which, for convenience, is still denoted as H_{RICD}.

Theorem 2. *Let H be a subset of $\{1, \ldots, n\}$ with $|H| = h > n/2$. If $\tilde{H} \subset \{1, \ldots, n\}$ with $\left|\tilde{H}\right| = h$ is such that $\left\{d_i^2\left(\bar{\boldsymbol{x}}_H, S_H(\lambda)\right) : i \in \tilde{H}\right\} = \{d_{(1)}^2\left(\bar{\boldsymbol{x}}_H, S_H(\lambda)\right), \ldots, d_{(h)}^2\left(\bar{\boldsymbol{x}}_H, S_H(\lambda)\right)\}$, where $d_{(1)}^2\left(\bar{\boldsymbol{x}}_H, S_H(\lambda)\right) \leq \ldots \leq d_{(n)}^2\left(\bar{\boldsymbol{x}}_H, S_H(\lambda)\right)$ denote the order statistics of $\{d_i^2\left(\bar{\boldsymbol{x}}_H, S_H(\lambda)\right), i = 1, \ldots, n\}$. Then,*

$$\det\left[S_{\tilde{H}}(\lambda)\right] \leq \det\left[S_H(\lambda)\right]$$

with equality if and only if $\bar{\boldsymbol{x}}_H = \bar{\boldsymbol{x}}_{\tilde{H}}$ and $S_H(\lambda) = S_{\tilde{H}}(\lambda)$.

2.3 Asymptotic properties

The following theorem serves as a theoretical background for constructing a rule for identifying outliers. We first define two notations:

$$\begin{aligned}\Theta^{(1)}(\lambda, c, A) &= \frac{1-\lambda m_1(-\lambda)}{1-c\left[1-\lambda m_1(-\lambda)\right]},\\ \Theta^{(2)}(\lambda, c, A) &= \frac{1-\lambda m_1(-\lambda)}{\left[1-c+c\lambda m_1(-\lambda)\right]^3} - \lambda\frac{m_1(-\lambda)-\lambda m_2(-\lambda)}{\left[1-c+c\lambda m_1(-\lambda)\right]^4},\end{aligned} \tag{2.8}$$

where c is a constant, A is a $p \times p$ nonnegative definite matrix, $m_1(z)$ is defined as the Stieltjes transform of the ESD of A, $m_1(z) = \operatorname{tr}\left(A - zI_p\right)^{-1}/p$, and $m_2(z) = \operatorname{tr}\left(A - zI_p\right)^{-2}/p$.

Theorem 3. *Assume that Conditions A1–A5 hold. Let $\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}$ be i.i.d. random vectors that have the same distribution as $\boldsymbol{X}_n$ in (2.5). Then, for any k and $\lambda > 0$, we have*

$$\frac{\sqrt{p}\left((1/p)d_k^2(\bar{\boldsymbol{X}}_n, S_n(\lambda)) - \Theta^{(1)}(\lambda, c_n, S_n)\right)}{\sqrt{2\Theta^{(2)}(\lambda, c_n, S_n)}} \xrightarrow{D} \mathrm{N}(0,1), \quad \text{as } p \to \infty, \tag{2.9}$$

where $\bar{\boldsymbol{X}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{X}_{i,n}$, $S_n(\lambda) = S_n + \lambda I_p$ with $S_n = n^{-1}\sum_{i=1}^n \left(\boldsymbol{X}_{i,n} - \bar{\boldsymbol{X}}_n\right)\left(\boldsymbol{X}_{i,n} - \bar{\boldsymbol{X}}_n\right)^\top$, "$\xrightarrow{D}$" denotes the convergence in distribution.

Note that we may suppress the second subscript n in $\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}$ if there is no confusion in the context.

Since the computation of both quantities $\Theta^{(1)}(\lambda, c_n, S_n)$ and $\Theta^{(2)}(\lambda, c_n, S_n)$ does not require any knowledge on the true covariance matrix Σ_p beyond its positive definiteness, Theorem 3 provides a practical and efficient way for determining the cut-off value for identifying outliers.

2.4 The minimum ridge covariance determinant procedure

As remarked previously, we adapt the procedure in the fast minimum covariance determinant approach (Rousseeuw and Van Driessen, 1999) to solve the optimization problem (2.6), especially in a high-dimensional setting. We will present a procedure to find H_{RICD} and the raw cutoff.

We first explain what we mean by applying the concentration step described in Theorem 2 to a subset of $\{1,\ldots,n\}$ ℓ times, which is as follows: apply the concentration step to this subset, say $H_{(0)}$, and obtain a new subset of $\{1,\ldots,n\}$, say $H_{(1)}$; apply the concentration step to $H_{(1)}$, and obtain another new subset of $\{1,\ldots,n\}$, say $H_{(2)}$; repeat it $\ell-2$ times, and obtain the final subset of $\{1,\ldots,n\}$, say $H_{(\ell)}$.

Denote z_α the upper α-quantile of the standard normal distribution. Our procedure is given below.

Algorithm 1: The minimum ridge covariance determinant (RICD) procedure.

Step 1. Randomly sample c_s initial subsets $H_{j,\text{ini}}$ from $\{1,\ldots,n\}$ with $|H_{j,\text{ini}}| = [n/2]+1$, $j=1,\ldots,c_s$. Apply the concentration step to each initial subset three times, and obtain c_s concentrated subsets. Select l subsets from the above c_s concentrated subsets such that their ridge covariance determinants are the lowest.

Step 2. For each subset in the above l subsets, continue applying the concentration step until convergence and obtain l final subsets. Select the best subset with the minimum ridge covariance determinant as H_{RICD}.

Step 3. Compute $\hat{\boldsymbol{\mu}}_{\text{RICD}}$ and $\hat{\Sigma}_{\text{RICD}}$. Also compute $\Theta^{(1)}\left(\lambda, c_h, S_{H_{\text{RICD}}}\right)$ and $\Theta^{(2)}\left(\lambda, c_h, S_{H_{\text{RICD}}}\right)$ with $c_h = p/h$. For a given significance level of α, the kth observation is declared to be an outlier if

$$d_k^2\left(\hat{\boldsymbol{\mu}}_{\text{RICD}}, \hat{\Sigma}_{\text{RICD}}\right) > p\Theta^{(1)}\left(\lambda, c_h, S_{H_{\text{RICD}}}\right) + z_\alpha\left\{2p\Theta^{(2)}\left(\lambda, c_h, S_{H_{\text{RICD}}}\right)\right\}^{1/2}. \tag{2.10}$$

2.5 Refined minimum RICD procedure

Note that a one-step reweighting scheme is often an effective way to increase the efficiency of an algorithm in practice (Cerioli, 2010; Ro et al., 2015). Therefore, we improve the power of the

proposed outlier test described in Section 2.4 by adding a further reweighting step. In light of Ro et al. (2015), we first assume that the parameters $\boldsymbol{\mu}$ and Σ_p are known, and define the weights

$$W_k = \begin{cases} 0 & \text{if } d_k^2\left(\boldsymbol{\mu}, S_n(\lambda)\right) > a_\delta, \\ 1 & \text{otherwise,} \end{cases} \tag{2.11}$$

where a_δ is the upper δ-quantile of the distribution of $d_k^2\left(\boldsymbol{\mu}, S_n(\lambda)\right)$. By (A1) in Lemma A.4 given in Appendix, it follows that

$$a_\delta = \operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right) + z_\delta\sqrt{2\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2}. \tag{2.12}$$

We have the following proposition.

Proposition 1. *Assume that Conditions A1–A4 hold. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. p-dimensional random vectors from $N_p(\boldsymbol{\mu}, \Sigma_p)$. Then $\mathrm{E}\left(X_{kj} \mid W_k = 1\right) = \mu_j$, the jth element of $\boldsymbol{\mu}$, and*

$$\operatorname{Var}\left(X_{kj} \mid W_k = 1\right) = \sigma_{jj}\left[1 - \frac{2\phi\left(z_\delta\right)\left(\Sigma_p S_n(\lambda)^{-1}\Sigma_p\right)_{jj}}{\sigma_{jj}(1-\delta)\sqrt{2\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2}} + o(1)\right] \equiv \sigma_{jj}\tau_j, \tag{2.13}$$

where $\left(\Sigma_p S_n(\lambda)^{-1}\Sigma_p\right)_{jj}$ is the jth diagonal element of $\Sigma_p S_n(\lambda)^{-1}\Sigma_p$, $j = 1, \ldots, p$, and ϕ is the standard normal density function.

This proposition reveals that $\operatorname{Var}\left(X_{kj} \mid W_k = 1\right)$ is smaller than the true scatter parameter σ_{jj}. Therefore, if too many observations would be identified as outliers, it would lead to a biased Type I error. Cerioli (2010) showed by simulation that multiplying the raw MCD scatter estimate by a proportionality constant $k_{\mathrm{MCD}}(h, n, v)$ would improve the finite sample performance of its algorithm. Denote $W_{\mathrm{RICD}} = \{k_1, \ldots, k_{n_w}\}$ the set of indices of the observations $\boldsymbol{x}_k$ for which $w_k = 1$, where $w_k = 0$ if (2.10) holds, $w_k = 1$ otherwise and $n_w = \sum_{k=1}^n w_k$. In light of Cerioli (2010), we refine our estimates as follows:

$$\tilde{\boldsymbol{\mu}} = \bar{\boldsymbol{x}}_{W_{\mathrm{RICD}}}, \quad \tilde{S} = k_{\mathrm{RICD}}(h, p) S_{W_{\mathrm{RICD}}}, \tag{2.14}$$

where $k_{\mathrm{RICD}}(h,p)$ is an adjustment coefficient depending on both h and p.

It is difficult to obtain a consistent estimate of τ_j in (2.13) in a high-dimensional setting for $j=1,\ldots,p$. Nevertheless, it can be shown that

$$\mathrm{median}_{1\le j\le p}\,\tau_j^{-1} \approx \left[1+\frac{2\phi\left(z_\delta\right)\mathrm{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)}{p(1-\delta)\sqrt{2\,\mathrm{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2}}\right]\{1+o(1)\},\quad p\to\infty,$$

where $\mathrm{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)$ and $\mathrm{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2$ can be estimated more easily. By Lemma A.2 in Appendix, we can set the scaling factor $k_{\mathrm{RICD}}(h,p)$ in (2.14) as

$$k_{\mathrm{RICD}}(h,p)=1+\frac{2\phi\left(z_{\delta_w}\right)\Theta^{(1)}\left(\lambda,c_h,S_{H_{\mathrm{RICD}}}\right)}{(1-\delta_w)\sqrt{2p\Theta^{(2)}\left(\lambda,c_h,S_{H_{\mathrm{RICD}}}\right)}},\tag{2.15}$$

where $\delta_w=1-n_w/n$ is the actual proportion of observations that are effectively excluded in the reweighting step. Our refined RICD procedure for outlier detection is summarized as follows.

Algorithm 2: Refined minimum RICD procedure

Step 1. Select the significance level α. Put $h=[n/2]+1$. Choose c_s, e.g., $c_s=100$, and l, e.g., $l=10$. Apply Algorithm 1. Calculate the distance $d_k^2\left(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}},\hat{\Sigma}_{\mathrm{RICD}}\right)$, and assign a weight to each observation according to (2.10) based on an appropriately chosen δ, e.g., $\delta=\alpha/2$.

Step 2. Obtain n_w and W_{RICD}, and compute the refined location and scatter estimates $\tilde{\boldsymbol{\mu}}$ and $\tilde{S}$ by (2.14) and (2.15).

Step 3. Calculate the refined distance $d_k^2\left(\tilde{\boldsymbol{\mu}},\tilde{S}(\lambda)\right)$, update $\Theta^{(1)}\left(\lambda,c_{n_w},\tilde{S}\right)$ and $\Theta^{(2)}\left(\lambda,c_{n_w},\tilde{S}\right)$ according to (2.8) with $c_{n_w}=p/n_w$ and $\tilde{S}(\lambda)=\tilde{S}+\lambda I_p$. For a given significance level of α, the kth observation is declared to be an outlier if

$$d_k^2\left(\tilde{\boldsymbol{\mu}},\tilde{S}(\lambda)\right)>p\Theta^{(1)}\left(\lambda,c_{n_w},\tilde{S}\right)+z_\alpha\left\{2p\Theta^{(2)}\left(\lambda,c_{n_w},\tilde{S}\right)\right\}^{1/2}.\tag{2.16}$$

2.6 Choice of λ

Chen et al. (2011) suggested using the asymptotic approximation to choose the degree of regularization in their RHT test statistic (2). According to the asymptotic properties of the modified Mahalanobis distance $d^2\left(\bar{\boldsymbol{x}}_n, S_n(\lambda)\right)$, we propose a completely data driven approach to choose the degree of regularization λ. Specifically, for each λ, we first calculate $\Theta^{(1)}\left(\lambda, c_n, S_n\right)$ and $\Theta^{(2)}\left(\lambda, c_n, S_n\right)$ based on the observed data $\mathcal{X}$. Then for a target significant level α, the difference between $d^2\left(\bar{\boldsymbol{x}}_n, S_n(\lambda)\right)$ and its asymptotic approximation is measured by

$$D_\alpha(\lambda) = \text{median}_{1\le k\le n}\, d_k^2\left(\bar{\boldsymbol{x}}_n, S_n(\lambda)\right) - p\Theta^{(1)}\left(\lambda, c_n, S_n\right) - z_\alpha\left\{2p\Theta^{(2)}\left(\lambda, c_n, S_n\right)\right\}^{1/2}.$$

We select λ as

$$\hat{\lambda} = \min\left\{\lambda : \lambda \in \Xi, |D_\alpha(\lambda)| \le \varrho\right\},$$

where Ξ is a prespecified selecting range for λ, and ϱ is a small positive value. We set $\alpha = 0.05$, $\Xi = [0.05, 200]$ and $\varrho = 1$ in our simulation studies. Note that the optimal $\hat{\lambda}$ remains unchanged in the application of Algorithm 2 after it is chosen.

3. Numerical studies

3.1 Simulations

In this section, we carry out simulation studies to evaluate the performance of the proposed procedure (refined RICD). We generate the dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in the following two scenarios:

Scenario (I):

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independently distributed observations, where $\boldsymbol{x}_i$ is an observation from an ϵ-contaminated multivariate normal distribution $(1-\epsilon)N_p\left(\mathbf{0}, \Sigma_p\right) + \frac{1}{2}\epsilon N_p\left(\kappa\boldsymbol{\eta}_i, \Sigma_p\right) + \frac{1}{2}\epsilon N_p\left(-\kappa\boldsymbol{\eta}_i, \Sigma_p\right)$,

unless otherwise stated. Two cases of $\boldsymbol{\eta}_i$s are considered, which are (i) (dense mean vector case): $\boldsymbol{\eta}_i$ is the normalized p-dimensional vector $\boldsymbol{\zeta}_i$ consisting of p i.i.d. random variables from the uniform distribution $\mathrm{U}(0,1)$, i.e., $\boldsymbol{\eta}_i = \boldsymbol{\zeta}_i / \left\|\boldsymbol{\zeta}_i\right\|_F$; (ii) (sparse mean vector case) $\boldsymbol{\eta}_i$ is the normalized p-dimensional vector $\boldsymbol{\zeta}_i$ in which $[p^{0.1}]$ randomly selected elements are i.i.d. from $\mathrm{U}(0,1)$ and the others are all zeros, i.e., $\boldsymbol{\eta}_i = \boldsymbol{\zeta}_i / \left\|\boldsymbol{\zeta}_i\right\|_F$.

We fix the sample size $n = 100$, set the dimension p respectively as 100, 200, and 400, and let the contamination ratio ϵ be 0.1 or 0.2. The two settings of the covariance structure and the magnitude of abnormality κ are given below:

Case (a) (Autoregressive correlation structure setting) $\Sigma_p = \left(0.3^{|i-j|}\right)_{p\times p}$; $\kappa = 8,\ 9,\ 10$ respectively for $p = 100,\ 200,\ 400$;

Case (b) (Random structure setting) $\Sigma_p = Q^\top D_0 Q$ with D_0 being a diagonal matrix with diagonal elements $d_{jj} \overset{\text{i.i.d.}}{\sim} \mathrm{U}(1,5)$, $j = 1,\ldots,p$, and Q being an orthonormal matrix that is constructed from the spectral decomposition of $W^\top W$ ($W^\top W = Q^\top \Lambda Q$) with $W = (w_{ij})_{p\times p}$ being such that $w_{ij} \overset{\text{i.i.d.}}{\sim} \mathrm{U}(0,1)$; $\kappa = 12,\ 14,\ 16$ respectively for $p = 100,\ 200,\ 400$.

Scenario (II) (Non-Gaussian scenario):

Case (c) Let p-dimensional random vector $\boldsymbol{\xi} = 0.7827\boldsymbol{\gamma} + 0.6224\boldsymbol{\nu}$, where $\boldsymbol{\gamma}$ has i.i.d. elements with the common distribution $\mathrm{U}(-\sqrt{3},\sqrt{3})$, and $\boldsymbol{\nu}$, independent of $\boldsymbol{\gamma}$, has i.i.d. elements with the common density function

$$f(\nu) = \begin{cases} \frac{\sqrt{2}}{2} e^{-\sqrt{2}\nu}, & \text{if } \nu \geq 0, \\ \frac{\sqrt{2}}{2} e^{\sqrt{2}\nu}, & \text{if } \nu < 0. \end{cases}$$

It can be shown that the distribution of ξ_1, the first element of $\boldsymbol{\xi}$, satisfies Condition A5. Denote the distribution of $\boldsymbol{\xi}$ by F_ξ. Replace the ϵ-contaminated multivariate normal distribution in Scenatio (I) by $(1-\epsilon)F_\xi + \frac{1}{2}\epsilon N_p\left(\kappa\boldsymbol{\eta}_i, I_p\right) + \frac{1}{2}\epsilon N_p\left(-\kappa\boldsymbol{\eta}_i, I_p\right)$; $\kappa = 8, 9$ or 10 respectively for

$p = 100, 200$ or 400.

We compare the proposed procedure (RICD) with some existing methods, including the refined minimum diagonal product procedure (RMDP) of Ro et al. (2015), the block diagonal product procedure (BDP) of Li and Jin (2022) and the principal component outlier detection procedure (PCout) of Filzmoser et al. (2008) for each setting. Outlier identification performance is evaluated by the Type I error rate, i.e., the proportion of good observations that are incorrectly classified as outliers, and the detection power, i.e., the proportion of contaminated observations that are correctly flagged out. The average Type I error rate $\bar{\alpha}$ and the detection power $\bar{\beta}$ presented in this section are calculated from 500 replications.

The average Type I error rates (%) of the the proposed RICD procedure for various p and ϵ are displayed in Table 1, the nominal significance level α is set to be 0.01, 0.05 or 0.1. From this table, it can be seen that the empirical Type I error rates are close to the nominal levels in most settings.

Simulation results of the four different methods with $\alpha = 0.05$, $\epsilon = 0.1$ and 0.2 are summarized in Tables 2-3, respectively. By these two tables, it can be seen that (i) the proposed method outperforms both RMDP and BDP procedures in terms of detection power in most cases; (ii) the PCout method has comparable performance with ours in Case (i) for $\epsilon = 0.1$. However, the former has conservative Type I error rate when the contamination ratio increases to 0.2 and suffers from some power loss in Case (ii).

In Scenario (I), a radial contamination scheme (Cerioli, 2010) is also considered, which we refer to as the following:

Case (d) (Scatter outliers) $\boldsymbol{x}_i^{(\epsilon)}$ is an observation from $(1-\epsilon)N_p\left(\mathbf{0}, \Sigma_p\right) + \epsilon N_p\left(\mathbf{0}, \Sigma_{(i)}\right)$, where Σ_p is set the same as in Case (a), $[p^{0.5}]$ random diagonal components of $\Sigma_{(i)}$ are 7.5, the other

Table 1: Average Type I error (%) by the proposed procedure for various p, ϵ and α.

			$\epsilon = 0.1$			$\epsilon = 0.2$		
$\boldsymbol{\eta}_i$	Case	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
(i)	(a)	100	1.74	6.97	13.2	1.21	5.56	11.13
		200	1.41	6.59	13.15	1.00	5.28	11.16
		400	1.38	6.50	12.16	0.97	5.16	9.77
	(b)	100	1.62	6.86	12.89	1.22	5.44	10.75
		200	1.38	6.48	12.74	0.97	5.15	10.52
		400	1.26	6.05	11.02	0.83	4.50	8.03
	(c)	100	1.63	6.06	11.49	1.09	4.84	9.55
		200	1.33	6.00	11.73	0.92	4.74	9.81
		400	1.31	5.99	10.87	0.90	4.66	8.60
(ii)	(a)	100	1.75	7.00	13.17	1.24	5.60	11.16
		200	1.42	6.63	13.14	1.02	5.30	11.09
		400	1.41	6.52	12.19	1.02	5.15	9.54
	(b)	100	1.61	6.85	12.87	1.20	5.41	10.75
		200	1.39	6.49	12.71	0.98	5.17	10.50
		400	1.27	6.05	10.75	0.87	4.56	7.97
	(c)	100	1.61	6.08	11.52	1.08	4.84	9.56
		200	1.35	6.00	11.73	0.92	4.71	9.76
		400	1.29	5.99	10.76	0.92	4.65	8.40

Table 2: Average Type I error (%) and detection power (%), where $\alpha = 0.05$ and $\epsilon = 0.1$.

			RICD		RMDP		BDP		PCout	
$\boldsymbol{\eta}_i$	Case	p	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$
(i)	(a)	100	6.97	94.60	6.59	94.22	6.95	91.72	5.33	97.24
		200	6.59	92.53	6.19	92.69	7.34	90.92	5.01	97.52
		400	6.50	90.09	5.86	87.05	7.94	85.82	5.15	97.19
	(b)	100	6.86	88.47	6.10	86.72	6.61	83.31	5.65	92.91
		200	6.48	88.21	6.15	87.11	7.32	84.17	5.09	95.33
		400	6.05	83.56	5.85	83.92	8.42	83.67	4.71	95.73
	(c)	100	6.06	97.80	6.15	97.24	6.41	95.38	4.61	99.33
		200	6.00	96.61	6.12	96.29	7.57	94.98	4.15	100.00
		400	5.99	93.76	5.96	93.20	8.34	91.70	4.38	98.57
(ii)	(a)	100	7.00	97.35	6.33	92.80	6.79	95.07	6.86	30.79
		200	6.63	95.12	6.22	90.48	7.34	92.70	7.06	21.11
		400	6.52	92.42	6.17	83.68	8.47	84.60	7.62	17.64
	(b)	100	6.85	88.28	6.31	81.71	6.80	83.28	7.66	24.76
		200	6.49	88.13	6.40	81.72	7.60	83.19	7.20	16.59
		400	6.05	84.92	5.99	77.01	8.70	78.98	7.66	14.69
	(c)	100	6.08	97.66	6.16	95.18	6.59	95.66	6.98	39.96
		200	6.00	96.27	6.11	93.11	7.58	93.40	6.63	27.44
		400	5.99	94.27	5.78	86.49	8.57	85.84	7.03	19.46

Table 3: Average Type I error ($\bar{\alpha}$ %) and detection power ($\bar{\beta}$ %), where $\alpha = 0.05$ and $\epsilon = 0.2$.

			RICD		RMDP		BDP		PCout	
$\boldsymbol{\eta}_i$	Case	p	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$
(i)	(a)	100	5.56	93.38	4.89	91.80	5.25	87.92	2.22	98.12
		200	5.28	91.42	4.87	89.87	5.78	86.28	1.97	99.62
		400	5.19	88.53	4.31	84.26	6.17	81.13	1.69	99.85
	(b)	100	5.44	84.88	4.49	83.85	4.81	77.58	2.04	99.42
		200	5.15	85.12	4.55	83.43	5.73	78.54	1.71	99.96
		400	4.50	79.50	4.24	78.63	6.15	76.34	1.60	99.95
	(c)	100	4.84	96.84	4.67	96.76	5.31	93.89	1.55	99.99
		200	4.74	95.67	4.57	94.84	5.87	91.59	1.39	100.00
		400	4.66	92.22	4.46	90.84	6.62	87.62	1.26	100.00
(ii)	(a)	100	5.60	96.28	4.76	90.95	5.20	93.44	5.63	30.88
		200	5.30	94.10	4.73	86.91	5.78	89.30	6.57	19.99
		400	5.15	90.64	4.88	79.92	7.13	81.51	6.91	15.49
	(b)	100	5.41	85.48	4.90	78.20	5.31	79.20	6.24	23.06
		200	5.17	85.53	5.06	77.80	6.04	79.57	6.78	16.90
		400	4.56	81.06	4.91	72.98	7.28	74.80	7.14	14.66
	(c)	100	4.84	96.79	4.78	93.63	5.25	93.96	5.59	35.75
		200	4.71	94.92	4.82	90.21	6.23	90.76	5.53	24.93
		400	4.65	93.09	4.55	83.70	6.96	82.62	6.37	19.55

Table 4: Average Type I error ($\bar{\alpha}$ %) and detection power ($\bar{\beta}$ %) in Case (d), where $\alpha = 0.05$.

			RICD		RMDP		BDP		PCout	
Case	ϵ	p	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$	$\bar{\alpha}$	$\bar{\beta}$
(d)	0.1	100	6.89	89.00	6.27	86.74	6.70	89.25	7.23	33.86
		200	6.55	91.25	6.36	89.23	7.48	91.96	7.08	26.64
		400	6.46	94.14	5.90	92.15	8.14	94.60	6.90	25.51
	0.2	100	5.62	87.30	4.76	83.35	5.21	86.20	5.61	33.02
		200	5.28	89.45	4.87	87.13	5.97	90.26	5.82	25.67
		400	5.15	93.00	4.30	90.17	6.17	93.12	5.79	23.64

entries are the same as those of Σ_p.

We fix the significance level $\alpha = 0.05$ in this case. A comparison of results with different contamination ratios are reported in Table 4, which shows that the proposed method can maintain the desired Type I error rate and achieve high detection power simultaneously. Similar to the performance in location outlier settings, the PCout procedure appears to be insensitive to sparse signals. The Type I error rate seems not controlled well by the BDP procedure compared with the proposed method for $p \geq 200$ and $\epsilon = 0.1$.

3.2 Real data analysis

We illustrate the proposed method on the octane dataset, which consists of near-infrared absorbance spectra with $p = 226$ wavelengths collected on $n = 39$ gasoline samples. The dataset was described in Esbensen et al. (1996), and is available in the R package `rrcov`. Since this

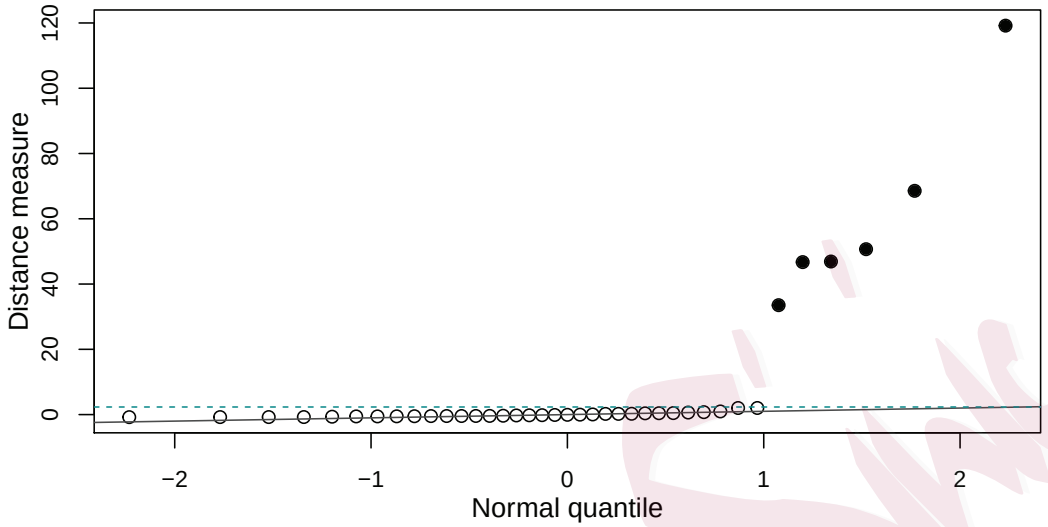

Figure 1: Q-Q plot of the distance measures based on the RICD.

dataset has a large p/n ratio, and hence the original minimum covariance determinant estimate cannot be computed. As the 25th, 26th, 36th, 37th, 38th, and 39th samples contain added ethanol, they are outliers. We apply the proposed method to this dataset at a significance level of 0.01, and record the distance measures $\left[d^2(\tilde{\boldsymbol{\mu}}, \tilde{S}(\lambda)) - p\Theta^{(1)}(\lambda, c_{n_w}, \tilde{S})\right] / \left[2p\Theta^{(2)}(\lambda, c_{n_w}, \tilde{S})\right]^{1/2}$ (see (2.16)). The Q-Q plot of the distance measures is given in Figure 1, in which the dashed horizontal line indicates the cut-off value, "good" points are around the black solid line, and the true outliers are labeled as the solid points. This figure clearly demonstrates that the proposed procedure has correctly pinpointed all the six outliers.

More numerical studies are given in the supplementary material.

4. Conclusion

In this paper, we propose a new outlier detection procedure based on the ridge sample covariance matrix. The resulting high-breakdown ridge covariance determinant estimate is well-defined for high-dimensional data and contains more information on the correlations among the variables than the MDP estimate (Ro et al., 2015). We obtain the asymptotic distribution of the modified Mahalanobis distance by relaxing the commonly used Gaussian assumption. This novel outlier detection procedure first finds a clean subset by applying the concentration step, and then identifies outliers whose modified distances are above the cut-off value. The regularization parameter is adaptively selected based on the data, thereby enhancing the robustness of the proposed method. Through simulations and a real data example, we illustrate that the proposed method is robust to masking and swamping effects of the contaminated data, and outperforms the existing methods RMDP, BDP, and PCout in certain situations.

5. Appendix

First, we give some lemmas.

Lemma A.1. *(Lemma 4 of Chen et al. (2011)) Given random variables* $\{x_n, y_n\}_{n=1}^{\infty}$. $f_n(x_n, y_n)$ *is a real function of* x_n *and* y_n. *If* $f_n \mid \mathcal{F}_n \xrightarrow{D} G$ *and distribution* G *is independent of* $\mathcal{F}_n$, *here* $\mid \mathcal{F}_n$ *denotes conditional on* $\mathcal{F}_n$ *and* $\mathcal{F}_n$ *is the* σ*-field generated by* $\{y_1, \ldots, y_n\}$, *then we have* $f_n \xrightarrow{D} G$.

Lemma A.2. *(Theorem 2.3 of Ha et al. (2021)) Assume that Conditions A1–A5 hold. Let* $\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}$ *be i.i.d. random vectors that have the same distribution as* $\boldsymbol{X}_n$ *in* (2.5). *For any*

$\lambda > 0$, we have

$$\sqrt{p}\left|\frac{1}{p}\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right) - \Theta^{(1)}\left(\lambda, c_n, S_n\right)\right| \xrightarrow{\text{p}} 0$$

and

$$\frac{1}{p}\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2 - \Theta^{(2)}\left(\lambda, c_n, S_n\right) \xrightarrow{\text{p}} 0, \quad \text{as } p \to \infty,$$

where "$\xrightarrow{\text{p}}$" denotes the convergence in probability, $\Theta^{(i)}, i = 1, 2$, are defined in (2.8).

Lemma A.3. *(Lemmas 4.2-4.4 of Ha et al. (2021)) Assume that Condition A1 holds. Let A be a $p \times p$ nonrandom symmetric matrix with bounded spectral norm, and $Z = (z_{ij})$ a $p \times n$ random matrix whose entries are i.i.d., satisfying*

$$\mathrm{E}\, z_{11} = 0, \quad \mathrm{E}\, z_{11}^2 = 1, \quad \mathrm{E}\, z_{11}^4 < \infty, \quad \text{and} \quad |z_{11}| \le \eta_n\sqrt{n},$$

where $\{\eta_n\}$ is a deterministic sequence with $\eta_n \downarrow 0$ whose convergence rate can be made arbitrarily slow. Then

$$\mathrm{E}\left|\bar{\boldsymbol{z}}_k^\top A\bar{\boldsymbol{z}}_k\right|^v \le k_v, \quad \mathrm{E}\left|\frac{1}{n}\boldsymbol{z}_k^\top A\boldsymbol{z}_k\right|^v \le k_v, \quad \mathrm{E}\left|\bar{\boldsymbol{z}}_k^\top A\boldsymbol{z}_k\right|^v \le k_v, \quad v = 1, 2, \ldots$$

where $\bar{\boldsymbol{z}}_k = \frac{1}{n}\sum_{j\neq k}^n \boldsymbol{z}_j$, $k = 1, \ldots, n$, $\boldsymbol{z}_j$ is the jth column of Z, and k_v is a constant depending on v.

When $\boldsymbol{X}_{k,n} \in \{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}\}$, it is difficult to obtain the universality of the CLT for the proposed estimator directly since $\boldsymbol{X}_{k,n}$ is not independent of the sample covariance S_n, and hence the ridge covariance $S_n(\lambda)$. Thus we divide the proof into two steps. Let $\Omega_n = \{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}, \ldots\}$ denote the complete set of random vectors generated by model (2.5). At first, Lemma A.4 is derived to characterize the asymptotic distribution of the modified distance (1.4) when the objective $\tilde{\boldsymbol{X}} \notin \{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}\}$. Define $\tilde{d}^2(\boldsymbol{\eta}, S_n(\lambda)) = (\tilde{\boldsymbol{X}} - \boldsymbol{\eta})^\top S_n(\lambda)^{-1}(\tilde{\boldsymbol{X}} - \boldsymbol{\eta})$.

Lemma A.4. *Assume that Conditions A1–A5 hold. Let* $\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}, \ldots$ *be i.i.d. random vectors that have the same distribution as* $\boldsymbol{X}_n$ *in* (2.5)*. If the random vector* $\tilde{\boldsymbol{X}}$ *is independent of* $\{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}\}$ *and* $\lambda > 0$*, we have*

$$\frac{\tilde{d}^2(\boldsymbol{\mu}, S_n(\lambda)) - \operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)}{\sqrt{2\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2}} \xrightarrow{D} \mathrm{N}(0,1), \quad p \to \infty, \tag{A1}$$

where $\tilde{d}^2(\boldsymbol{\mu}, S_n(\lambda)) = (\tilde{\boldsymbol{X}} - \boldsymbol{\mu})^\top S_n(\lambda)^{-1}(\tilde{\boldsymbol{X}} - \boldsymbol{\mu})$.

Lemma A.5. *Assume that Conditions A1–A4 hold. Let* $\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}, \ldots$ *be i.i.d. random vectors that have the same distribution as* $\boldsymbol{X}_n$ *in* (2.5) *satisfying that* $\mathrm{E}\, z_{11} = 0$ *and* $\mathrm{E}\, z_{11}^2 = 1$*. If the random vector* $\tilde{\boldsymbol{X}}$ *is independent of* $\{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}\}$ *and* $\lambda > 0$*, we have*

$$\frac{\left|\tilde{d}^2(\bar{\boldsymbol{X}}_n, S_n(\lambda)) - \tilde{d}^2(\boldsymbol{\mu}, S_n(\lambda))\right|}{\sqrt{2\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2}} = o_{\mathrm{p}}(1), \quad p \to \infty. \tag{A2}$$

The asymptotic bias between $\tilde{d}^2(\bar{\boldsymbol{X}}_n, S_n(\lambda))$ and $\tilde{d}^2(\boldsymbol{\mu}, S_n(\lambda))$ is formally given in Lemma A.5, which ensures that we can use the raw location and scatter estimators to select a cut-off value for outlier identification. Next, instead of letting $\tilde{\boldsymbol{X}}$ be independent of $\{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}\}$, we consider the modified distance (1.4) if $\tilde{\boldsymbol{X}} \in \{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}\}$.

Let $\boldsymbol{X}_{k0,n} = (X_{1,n}, \ldots, X_{k-1,n}, 0, X_{k+1,n}, \ldots, X_{n,n})^\top$, $\bar{\boldsymbol{X}}_{k0} = \frac{1}{n}\boldsymbol{X}_{k0,n}^\top \mathbf{1}_n$, $S_{n,k0} = \frac{1}{n}\boldsymbol{X}_{k0,n}\boldsymbol{X}_{k0,n}^\top - \bar{\boldsymbol{X}}_{k0}\bar{\boldsymbol{X}}_{k0}^\top$ and $S_0(\lambda) = S_{n,k0} + \lambda I_p$. Here $\mathbf{1}_n$ denotes an n-dimensional vector consisting of 1s. The asymptotic bias between $d_k^2(\bar{\boldsymbol{X}}_n, S_n(\lambda))$ and $d_k^2(\bar{\boldsymbol{X}}_{k0}, S_0(\lambda))$ is given in the following lemma.

Lemma A.6. *Assume that Conditions A1–A4 hold. Let* $\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}$ *be i.i.d. random vectors that have the same distribution as* $\boldsymbol{X}_n$ *in* (2.5) *satisfying that* $\mathrm{E}\, z_{11} = 0$*,* $\mathrm{E}\, z_{11}^2 = 1$ *and* $\mathrm{E}\, z_{11}^4 < \infty$*. For any* $\boldsymbol{X}_{k,n} \in \{\boldsymbol{X}_{1,n}, \ldots, \boldsymbol{X}_{n,n}\}$ *and* $\lambda > 0$*, the following three arguments hold:*

$$\frac{\left|d_k^2\left(\bar{\boldsymbol{X}}_n, S_n(\lambda)\right) - d_k^2\left(\bar{\boldsymbol{X}}_{k0}, S_0(\lambda)\right)\right|}{\sqrt{2\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma_p\right)^2}} = o_p(1), \tag{A3}$$

$$\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma_p\right)^2 - \operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2 = O_p(1), \tag{A4}$$

$$\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma_p\right) - \operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right) = O_p(1), \quad p \to \infty. \tag{A5}$$

Although (A1) presupposes that the estimate of $\boldsymbol{\mu}$ and $S_n(\lambda)$ are a sample without outliers, it is also expected to be roughly valid for the distance $d_k^2(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}, \hat{\Sigma}_{\mathrm{RICD}})$, where $\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}$ and $\hat{\Sigma}_{\mathrm{RICD}}$ are reliable approximations to those obtained from a clean sample. This lemma, in conjunction with (A1) and (A2), suggests that we could use normal distributions to construct a threshold rule.

Note that both $\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)$ and $\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma_p\right)^2$ in (A1) involve the unknown covariance matrix Σ_p. Thus, Σ_p needs to be estimated in order to obtain the cut-off value for outlier identification. By the Stieltjes transform of the empirical spectral measure of a random matrix, we can simply adapt the estimates $\Theta^{(1)}(\lambda, c)$ and $\Theta^{(2)}(\lambda, c)$ from Ha et al. (2021).

The proofs of Theorems 1–2, Lemmas A.4–A.6, Theorem 3 and Proposition 1 are given below. For simplicity, we suppress the subscripts of Σ_p, S_n, $S_n(\lambda)$ and $\bar{\boldsymbol{X}}_n$, and suppress the second subscript n in the subscript $\{\ell, n\}$ if there is no confusion in the context.

Proof of Theorem 1. First we prove that $\varepsilon_n(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}, \mathcal{X}) \leq (n-h+1)/n$. If we replace $(n-h+1)$ observations of the original data set $\mathcal{X}$, then the optimal subset $\tilde{H}_{\mathrm{RICD}}$ of $\tilde{\mathcal{X}}$ would contain at least one outlier, but the least square method breaks down even with one single outlier. Denote $\tilde{\boldsymbol{\mu}}_{\mathrm{RICD}} = \bar{\boldsymbol{x}}_{\tilde{H}_{\mathrm{RICD}}}$, it then follows that $\|\tilde{\boldsymbol{\mu}}_{\mathrm{RICD}}\|_F$ is not bounded.

On the other hand, to show $\varepsilon_n(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}, \mathcal{X}) \geq (n-h+1)/n$, we prove that there exists a value M, which only depends on $\mathcal{X}$ and λ, such that for every $\tilde{\mathcal{X}}$ obtained by replacing at most $(n-h)$ observations in $\mathcal{X}$, the Frobenius norm of the RICD location estimate $\tilde{\boldsymbol{\mu}}_{\mathrm{RICD}}$ based on $\tilde{\mathcal{X}}$ is still

bounded by M from above.

If we take any dataset $\tilde{\mathcal{X}}$ by replacing $(n-h)$ observations in $\mathcal{X}$, there still exists a subset $H_1 \in \mathcal{H}$ containing indices only corresponding to the data points of the original dataset $\mathcal{X}$. The determinant of $S_{H_1}(\lambda)$ is

$$\begin{aligned}\det\left[S_{H_1}(\lambda)\right] &= \prod_{k=1}^{p} \eta_k \leq \left(\frac{1}{p}\sum_{k=1}^{p}\eta_k\right)^p \\ &= \left[\frac{1}{hp}\sum_{k=1}^{p}\sum_{j\in H_1}\left\{x_{jk}-\hat{\mu}_k\left(H_1\right)\right\}^2+\lambda\right]^p \\ &\leq \left(4N^2+\lambda\right)^p,\end{aligned}$$

where $(\eta_1,\ldots,\eta_p)$ are the eigenvalues of the matrix $S_{H_1}(\lambda)$, $\hat{\mu}_k\left(H_1\right)$ denotes the kth component of $\bar{\boldsymbol{x}}_{H_1}$, and N is defined as $\max_{1\leq i\leq n, 1\leq j\leq p}|x_{ij}|$.

Let H_2 be the optimal subset corresponding to $\tilde{\mathcal{X}}$, then $\tilde{\boldsymbol{\mu}}_{\mathrm{RICD}} = \bar{\boldsymbol{x}}_{H_2}$. Since $h-(n-h)\geq 1$, the set H_2 contains one observation $\boldsymbol{x}_{i_0}$ from $\mathcal{X}$. Thus we have

$$\det\left[S_{H_2}(\lambda)\right] = \det[A+B] = \det(A)\cdot\det\left(I_p + A^{-1}B\right),$$

where

$$A = h^{-1}\left(\boldsymbol{x}_{i_0}-\bar{\boldsymbol{x}}_{H_2}\right)\left(\boldsymbol{x}_{i_0}-\bar{\boldsymbol{x}}_{H_2}\right)^\top + 2^{-1}\lambda I_p,$$

and

$$B = h^{-1}\sum_{i\in H_2, i\neq i_0}\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{H_2}\right)\left(\boldsymbol{x}_i-\bar{\boldsymbol{x}}_{H_2}\right)^\top + 2^{-1}\lambda I_p.$$

It follows that

$$\begin{aligned}\det\left[S_{H_2}(\lambda)\right] &> \det(A) \\ &= 2^{-p}\lambda^p \det\left[I_p + \frac{2}{h\lambda}\left(\boldsymbol{x}_{i_0}-\bar{\boldsymbol{x}}_{H_2}\right)\left(\boldsymbol{x}_{i_0}-\bar{\boldsymbol{x}}_{H_2}\right)^\top\right] \\ &= 2^{-p}\lambda^p + \frac{1}{h}2^{1-p}\lambda^{p-1}\left(\boldsymbol{x}_{i_0}-\bar{\boldsymbol{x}}_{H_2}\right)^\top\left(\boldsymbol{x}_{i_0}-\bar{\boldsymbol{x}}_{H_2}\right).\end{aligned}$$

Let

$$M = p^{1/2}\left[\left\{\left[\left(4N^2+\lambda\right)^p - 2^{-p}\lambda^p\right]2^{p-1}\lambda^{1-p}h\right\}^{1/2} + N\right].$$

If $\|\tilde{\boldsymbol{\mu}}_{\mathrm{RICD}}\|_F > M$, then there exists j_0 such that $|\hat{\mu}_{j_0}(H_2)| > M/p^{1/2}$. Thus,

$$\begin{aligned}
\det\left[S_{H_2}(\lambda)\right] &> 2^{-p}\lambda^p + \frac{1}{h}2^{1-p}\lambda^{p-1}\left[x_{i_0j_0} - \hat{\mu}_{j_0}(H_2)\right]^2 \\
&\geq 2^{-p}\lambda^p + \frac{1}{h}2^{1-p}\lambda^{p-1}\left[|x_{i_0j_0}| - |\hat{\mu}_{j_0}(\mathrm{H}_2)|\right]^2 \\
&\geq 2^{-p}\lambda^p + \frac{1}{h}2^{1-p}\lambda^{p-1}\left[\frac{M}{p^{1/2}} - N\right]^2 \\
&= \left(4N^2+\lambda\right)^p
\end{aligned}$$

by the definition of M. This implies $\det\left[S_{H_2}(\lambda)\right] > \det\left[S_{H_1}(\lambda)\right]$, which contradicts the definition of $\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}$. So, we conclude that $\|\tilde{\boldsymbol{\mu}}_{\mathrm{RICD}}\|_F \leq M$. Since $\hat{\Sigma}_{\mathrm{RICD}}$ is obtained from $\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}$ based on the same subset H_{RICD}, we have $\varepsilon_n\left(\hat{\Sigma}_{\mathrm{RICD}}, \mathcal{X}\right) = \varepsilon_n\left(\hat{\boldsymbol{\mu}}_{\mathrm{RICD}}, \mathcal{X}\right)$, which concludes the proof of Theorem 1.

Proof of Theorem 2. The conclusions of Theorem 2 can be derived from Theorem 1 of Boudt et al. (2019), which is briefly described below:

For a given H, Boudt et al. (2019) regularized the sample covariance matrix S_H as $K_H = \rho T + (1-\rho)S_H$, where $0 < \rho < 1$ is a scalar weight coefficient and T is a predetermined positive definite target matrix. One can thus compute the distance $d_i^2(\bar{\boldsymbol{x}}_H, K_H) = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_H)^\top K_H^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_H)$. If we take $T = I_p$ and $\rho = \lambda/(1+\lambda)$, we have $S_H(\lambda) = (\lambda+1)K_H$, $d_i^2(\bar{\boldsymbol{x}}_H, S_H(\lambda)) = (\lambda+1)^{-1}d_i^2(\bar{\boldsymbol{x}}_H, K_H)$. Thus, Theorem 1 follows from Theorem 1 of Boudt et al. (2019).

Proof of Lemma A.4. First, let $V = \Sigma^{1/2}S(\lambda)^{-1}\Sigma^{1/2}$. By Condition A2 and the definition of $S(\lambda)$, the matrix V can be decomposed as $Q^\top \Lambda Q$, where Q is an orthogonal matrix and Λ is a diagonal matrix with positive diagonal elements $\zeta_{n,1} \leq \zeta_{n,2} \leq \cdots \leq \zeta_{n,p}$. It is obvious that for any n, the largest eigenvalue of $S(\lambda)^{-1}$ is bounded above by $1/\lambda$. On the other hand, Theorem 3.6 in Bai and Silverstein (2010) implies that $F^S(x)$ tends to the M-P law under Condition A1

(see Eq.(3.1.1) in Bai & Silverstein, 2010), and hence the largest eigenvalue of S is bounded away from infinity asymptotically. Therefore, we conclude that $\{\zeta_{n,i}\}$ are bounded away from both zero and infinity asymptotically.

Next, by the definition of V, we have

$$\tilde{d}^2(\boldsymbol{\mu}, S(\lambda)) = \tilde{\boldsymbol{Y}}^\top \Lambda \tilde{\boldsymbol{Y}} = \sum_{i=1}^{p} \zeta_{n,i} \tilde{y}_i^2 = \sum_{i=1}^{p} \zeta_{n,i} w_{n,i}, \tag{A6}$$

where $\tilde{\boldsymbol{Y}} = Q\Sigma^{-1/2} T_p \tilde{\boldsymbol{Z}} = (\tilde{y}_1, \ldots, \tilde{y}_p)^\top$ with $\tilde{\boldsymbol{X}} = T_p \tilde{\boldsymbol{Z}} + \boldsymbol{\mu}$ (see the equation (2.5)), and $w_{n,i} = \tilde{y}_i^2$. Since $\tilde{\boldsymbol{X}}$ is independent of S thus independent of Q, by Conditions A2 and A5, $\mathrm{E}\, w_{n,i} = 1$, $\mathrm{E}\, w_{n,i}^2 = 3$.

Let $W_{n,i} = \zeta_{n,i}(w_{n,i} - 1)$, $\vartheta_p = \sqrt{2\sum_{i=1}^{p} \zeta_{n,i}^2}$. Denote the σ-field generated by $\{\zeta_{n,1}, \ldots, \zeta_{n,p}\}$ by $\mathcal{F}$. It is easy to see that $\sqrt{2p}\zeta_{n,1} \le \vartheta_p \le \sqrt{2p}\zeta_{n,p}$. Conditional on $\mathcal{F}$, we have $\mathrm{E}(W_{n,i} \mid \mathcal{F}) = 0$, $\mathrm{E}\left(W_{n,i}^2 \mid \mathcal{F}\right) = 2\zeta_{n,i}^2$, and $\sum_{i=1}^{p} \mathrm{E}\left(\left(\frac{W_{n,i}}{\vartheta_p}\right)^2 \mid \mathcal{F}\right) = 1$. It follows that

$$\begin{aligned}
&\sum_{i=1}^{p} \mathrm{E}\left(\left|\frac{W_{n,i}}{\vartheta_p}\right|^2 ; \left|\frac{W_{n,i}}{\vartheta_p}\right| > \epsilon \mid \mathcal{F}\right) \\
&= \frac{1}{\vartheta_p^2} \sum_{i=1}^{p} \mathrm{E}\left(\zeta_{n,i}^2 (w_{n,i} - 1)^2 ; |w_{n,i} - 1| > \epsilon\vartheta_p/\zeta_{n,i} \mid \mathcal{F}\right) \\
&\le \frac{1}{\vartheta_p^2} \sum_{i=1}^{p} \mathrm{E}\left(\zeta_{n,p}^2 (w_{n,i} - 1)^2 ; |w_{n,i} - 1| > \epsilon\vartheta_p/\zeta_{n,p} \mid \mathcal{F}\right) \\
&\le \frac{p}{\vartheta_p^2} \mathrm{E}\left(\zeta_{n,p}^2 (w_{n,i} - 1)^2 ; |w_{n,i} - 1| > \epsilon\sqrt{2p}\zeta_{n,1}/\zeta_{n,p} \mid \mathcal{F}\right) \\
&\le \frac{1}{2\zeta_{n,1}^2} \mathrm{E}\left(W_{n,p}^2 ; |W_{n,p}| > \epsilon\sqrt{2p}\zeta_{n,1} \mid \mathcal{F}\right) \\
&\xrightarrow{\mathrm{p}} 0, \quad \text{as } p \to \infty.
\end{aligned}$$

Here $\mathrm{E}(x; a \mid b)$ denotes the expected value of x restricted to a while conditioned on b. Then, according to the Lindeberg-Feller central limit theorem, we have $\frac{\sum_{i=1}^{p} W_{p,i}}{\vartheta_p} \mid \mathcal{F} \xrightarrow{D} \mathrm{N}(0,1)$. Base

on Lemma A.1 we have

$$\frac{\tilde{d}^2(\boldsymbol{\mu}, S(\lambda)) - \sum_{i=1}^{p} \zeta_{n,i}}{\sqrt{2\sum_{i=1}^{p} \zeta_{n,i}^2}} \xrightarrow{D} \mathrm{N}(0,1).$$

The proof is complete.

Proof of Lemma A.5. By (A6), we have $\tilde{d}^2(\boldsymbol{\mu}, S(\lambda)) = \tilde{\boldsymbol{Y}}^\top \Lambda \tilde{\boldsymbol{Y}}$, where $\tilde{\boldsymbol{Y}} = (\tilde{y}_1, \ldots, \tilde{y}_p)^\top$. Similarly, for each $\boldsymbol{X}_i$, $i = 1, \ldots, n$, we can also define $\boldsymbol{Y}_i = Q\Sigma^{-1/2}T_p\boldsymbol{Z}_i$ with $\boldsymbol{X}_i = T_p\boldsymbol{Z}_i + \boldsymbol{\mu}$ and $\bar{\boldsymbol{Y}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{Y}_i$, where $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{ip})^\top$. Then

$$\left|\tilde{d}^2\left(\bar{\boldsymbol{X}}, S(\lambda)\right) - \tilde{d}^2\left(\boldsymbol{\mu}, S(\lambda)\right)\right| = \left|\left(\tilde{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\right)^\top \Lambda \left(\tilde{\boldsymbol{Y}} - \bar{\boldsymbol{Y}}\right) - \tilde{\boldsymbol{Y}}^\top \Lambda \tilde{\boldsymbol{Y}}\right|$$

$$= \left|\bar{\boldsymbol{Y}}^\top \Lambda \bar{\boldsymbol{Y}} - 2\tilde{\boldsymbol{Y}}^\top \Lambda \bar{\boldsymbol{Y}}\right| \leq \left|\bar{\boldsymbol{Y}}^\top \Lambda \bar{\boldsymbol{Y}}\right| + 2\left|\tilde{\boldsymbol{Y}}^\top \Lambda \bar{\boldsymbol{Y}}\right|.$$

As discussed in the proof of Lemma A.4, by Conditions A1 and A4 and the fact that the largest eigenvalue of $S(\lambda)^{-1}$ is bounded above by $1/\lambda$, the spectral norm of Λ, $\zeta_{n,p}$, is bounded above, say by ϖ. By Conditions A2 and the definition of $\boldsymbol{Y}_i$, we have

$$\mathrm{E}\, y_{ij} = 0, \quad \mathrm{E}\, y_{ij}^2 = 1.$$

Similar arguments also hold for $\tilde{y}_j$, $j = 1, \ldots, p$. Therefore, we have, for large n and p,

$$\mathrm{E}\left(\left|\bar{\boldsymbol{Y}}^\top \Lambda \bar{\boldsymbol{Y}}\right|\right) \leq \varpi\, \mathrm{E}\left(\bar{\boldsymbol{Y}}^\top \bar{\boldsymbol{Y}}\right) \leq \varpi\, \mathrm{E}\left[\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n} y_{ij}\right)^2\right] = \varpi p/n < 2c\varpi,$$

and

$$\mathrm{E}\left(\left|\tilde{\boldsymbol{Y}}^\top \Lambda \bar{\boldsymbol{Y}}\right|\right) \leq \varpi\, \mathrm{E}\left(\tilde{\boldsymbol{Y}}^\top \bar{\boldsymbol{Y}}\right) \leq \varpi\, \mathrm{E}\left[\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n} y_{ij}\tilde{y}_j\right)\right] < 2c\varpi,$$

which concludes the lemma.

Proof of Lemma A.6. Following steps of the truncation, centralization, and rescaling similar to those in Bai and Silverstein (2004), we may assume that the random variables $\{x_{ij}\}$ satisfy that

$$\mathrm{E}\, x_{ij} = 0, \quad \mathrm{E}\, x_{ij}^2 = 1, \quad \mathrm{E}\, x_{ij}^4 < \infty, \quad \text{and} \quad |x_{ij}| \leq \eta_n\sqrt{n},$$

where $\{\eta_n\}$ is a deterministic sequence such that $\eta_n \downarrow 0$ whose convergence rate can be made arbitrarily slow. Under these assumptions, for any $\alpha > 4$, we have

$$\mathrm{E}\,|x_{ij}|^{\alpha} = O\left(\left(\eta_n\sqrt{n}\right)^{\alpha-4}\right).$$

Since

$$\bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_{k0} + \frac{1}{n}\boldsymbol{X}_k,$$

we have

$$\begin{aligned} S_n &= S_{n,k0} + a_n\boldsymbol{X}_k\boldsymbol{X}_k^{\top} - n^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^{\top} - n^{-1}\bar{\boldsymbol{X}}_{k0}\boldsymbol{X}_k^{\top} \\ &= S_{n,k+} - n^{-1}\left(\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^{\top} + \bar{\boldsymbol{X}}_{k0}\boldsymbol{X}_k^{\top}\right), \end{aligned}$$

where $S_{n,k+} = S_{n,k0} + a_n\boldsymbol{X}_k\boldsymbol{X}_k^{\top}$ with $a_n = (n-1)/n^2$. For simplicity in writing, denote $R_n = S_n(\lambda)$, $R_0 = S_0(\lambda)$, and $R_1 = S_{n,k+} + \lambda I_p$. By the inverse matrix formula,

$$R_n^{-1} = R_1^{-1} + R_1^{-1}\left(n^{-1}\boldsymbol{X}_k, \bar{\boldsymbol{X}}_{k0}\right)\Delta^{-1}\begin{pmatrix} \bar{\boldsymbol{X}}_{k0}^{\top} \\ n^{-1}\boldsymbol{X}_k^{\top} \end{pmatrix} R_1^{-1}, \tag{A7}$$

where

$$R_1^{-1} = R_0^{-1} - \frac{a_n R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^{\top}R_0^{-1}}{1 + a_n\boldsymbol{X}_k^{\top}R_0^{-1}\boldsymbol{X}_k},$$

and

$$\Delta = I_2 - \begin{pmatrix} n^{-1}\bar{\boldsymbol{X}}_{k0}^{\top}R_1^{-1}\boldsymbol{X}_k & \bar{\boldsymbol{X}}_{k0}^{\top}R_1^{-1}\bar{\boldsymbol{X}}_{k0} \\ n^{-2}\boldsymbol{X}_k^{\top}R_1^{-1}\boldsymbol{X}_k & n^{-1}\boldsymbol{X}_k^{\top}R_1^{-1}\bar{\boldsymbol{X}}_{k0} \end{pmatrix}.$$

Denote

$$\Upsilon_k = R_1^{-1}\left(n^{-1}\boldsymbol{X}_k, \bar{\boldsymbol{X}}_{k0}\right)\Delta^{-1}\begin{pmatrix} \bar{\boldsymbol{X}}_{k0}^{\top} \\ n^{-1}\boldsymbol{X}_k^{\top} \end{pmatrix} R_1^{-1}.$$

We have

$$\Upsilon_k = \frac{\Upsilon}{\left(1 - n^{-1}\boldsymbol{X}_k^{\top}R_1^{-1}\bar{\boldsymbol{X}}_{k0}\right)^2 - n^{-2}\boldsymbol{X}_k^{\top}R_1^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^{\top}R_1^{-1}\bar{\boldsymbol{X}}_{k0}}, \tag{A8}$$

where

$$\begin{aligned}\Upsilon =& n^{-1}R_1^{-1}\boldsymbol{X}_k\left(1-n^{-1}\boldsymbol{X}_k^\top R_1^{-1}\bar{\boldsymbol{X}}_{k0}\right)\bar{\boldsymbol{X}}_{k0}^\top R_1^{-1}\\ &+n^{-2}R_1^{-1}\bar{\boldsymbol{X}}_{k0}\boldsymbol{X}_k^\top R_1^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^\top R_1^{-1}\\ &+n^{-2}R_1^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^\top R_1^{-1}\bar{\boldsymbol{X}}_{k0}\boldsymbol{X}_k^\top R_1^{-1}\\ &+n^{-1}R_1^{-1}\bar{\boldsymbol{X}}_{k0}\left(1-n^{-1}\bar{\boldsymbol{X}}_{k0}^\top R_1^{-1}\boldsymbol{X}_k\right)\boldsymbol{X}_k^\top R_1^{-1}.\end{aligned} \tag{A9}$$

Let $\beta_k = 1/(1+\ a_n\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k)$. By applying the identity

$$R_1^{-1} = R_0^{-1} - a_n\beta_k R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}, \tag{A10}$$

we obtain that

$$\boldsymbol{X}_k^\top\Upsilon\boldsymbol{X}_k := \mathrm{I}+\mathrm{II}+\mathrm{III}+\mathrm{IV},$$

where

$$\begin{aligned}\mathrm{I} =& n^{-1}\beta_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\left(1-n^{-1}\beta_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\right)\\ &\times\left(\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k - a_n\beta_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right),\\ \mathrm{II} =& n^{-2}\beta_k\left(\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0} - a_n\beta_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\right)\\ &\times\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\left(\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k - a_n\beta_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right),\\ \mathrm{III} =& n^{-2}\beta_k^2\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\left(\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\right.\\ &\left.-a_n\beta_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\right)\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k,\end{aligned}$$

and

$$\begin{aligned}\mathrm{IV} =& n^{-1}\beta_k\left(\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0} - a_n\beta_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\right)\\ &\times\left(1-n^{-1}\beta_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right)\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k.\end{aligned}$$

For the first term I, we have

$$\begin{aligned}\mathrm{I} =& n^{-1}\beta_k \boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k \\ &- n^{-2}\beta_k^2 \boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k \\ &- n^{-1}a_n\beta_k^2 \boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k \\ &+ n^{-2}a_n\beta_k^3 \boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k.\end{aligned} \tag{A11}$$

Note that β_k and $\|R_i\|$ for $i = n$, 0, or 1 are all bounded by some constant. It is easy to show that the order of the difference between $1/\left(1+\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k/n\right)$ and $\beta_k = 1/\left(1+a_n\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right)$ is $O_{L_1}\left(n^{-1}\right)$, say $\iota_n = O_{L_1}\left(n^{-1}\right)$, denoting that $\mathrm{E}\left|n\iota_n\right|$ is bounded by some constant. Thus, we simplify (A11) by substituting β_k with $1/\left(1+\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k/n\right)$. Similarly, we substitute a_n with $1/n$ there. By applying Lemma A.3 and Cauchy-Schwarz inequality, we obtain that

$$\begin{aligned}&\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right| \\ \leq&\sqrt{\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^2\mathrm{E}\left|\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right|^2} \\ =&O(1), \\ &\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right| \\ \leq&\sqrt{\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^2\mathrm{E}\left|\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right|^2} \\ \leq&\sqrt{\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^2\sqrt{\mathrm{E}\left|\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\right|^4\mathrm{E}\left|\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right|^4}} \\ =&O(1), \\ &\mathrm{E}\left|n^{-2}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right| \\ \leq&\sqrt{\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^2\mathrm{E}\left|n^{-1}\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^2} \\ \leq&\sqrt{\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^2\sqrt{\mathrm{E}\left|\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right|^4\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^4}} \\ =&O(1),\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}\left|n^{-2}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right| \\
\leq&\sqrt{\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0}\right|^2\mathrm{E}\left|n^{-1}\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^2} \\
\leq&\sqrt{\mathrm{E}\left|n^{-1}\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\right|^4\mathrm{E}\left|\bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\boldsymbol{X}_k\right|^4} \\
=&O(1),
\end{aligned}$$

which imply that $\mathrm{I} = O_{L_1}(1)$. The orders of the other three terms, i.e., II, III, and IV, can be derived similarly, from which one can verify that

$$\boldsymbol{X}_k^\top\Upsilon\boldsymbol{X}_k = O_{L_1}(1).$$

Furthermore, by (4.33) of Ha et al. (2021), the denominator of Υ_k in (A8) has the order of $1 + O_{L_1}(1)$, and hence it follows that

$$\boldsymbol{X}_k^\top\Upsilon_k\boldsymbol{X}_k = O_{L_1}(1). \tag{A12}$$

Similarly, it can be shown that

$$\boldsymbol{X}_k^\top\Upsilon_k\bar{\boldsymbol{X}}_{k0} = O_{L_1}(1). \tag{A13}$$

Returning to the first argument (A3) of Lemma A.6, we have

$$\begin{aligned}
d_k^2\left(\bar{\boldsymbol{X}}, S_n(\lambda)\right) - d_k^2\left(\bar{\boldsymbol{X}}_{k0}, S_0(\lambda)\right) &= (\boldsymbol{X}_k - \bar{\boldsymbol{X}})^\top R_n^{-1}(\boldsymbol{X}_k - \bar{\boldsymbol{X}}) - (\boldsymbol{X}_k - \bar{\boldsymbol{X}}_{k0})^\top R_0^{-1}(\boldsymbol{X}_k - \bar{\boldsymbol{X}}_{k0}) \\
&= \boldsymbol{X}_k^\top R_n^{-1}\boldsymbol{X}_k + (\bar{\boldsymbol{X}}_{k0} + n^{-1}\boldsymbol{X}_k)^\top R_n^{-1}(\bar{\boldsymbol{X}}_{k0} + n^{-1}\boldsymbol{X}_k) + 2\boldsymbol{X}_k^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0} \\
&\quad - 2\boldsymbol{X}_k^\top R_n^{-1}(\bar{\boldsymbol{X}}_{k0} + n^{-1}\boldsymbol{X}_k) - \boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k - \bar{\boldsymbol{X}}_{k0}^\top R_0^{-1}\bar{\boldsymbol{X}}_{k0},
\end{aligned}$$

which, jointly with Lemma A.3, (A7), (A8), (A10), (A12) and (A13), implies that

$$d_k^2\left(\bar{\boldsymbol{X}}, S_n(\lambda)\right) - d_k^2\left(\bar{\boldsymbol{X}}_{k0}, S_0(\lambda)\right) = -a_n\beta_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k + O_{L_1}(1). \tag{A14}$$

By the end of the proof of their Lemma 4.3 on Page 14 of Ha et al. (2021), we have that for any

z_i satisfying the conditions of Lemma A.3,

$$\sum_{i=1}^{p} \mathrm{E}\left(\frac{1}{n}|z_i|^2\right)^v \leq \begin{cases} O\left(n^{-v+1}\right) & \text{if } v \leq 2, \\ O\left(\eta_n^{2v-4} n^{-1}\right) & \text{if } v > 2. \end{cases}$$

By replacing the coefficient $1/n$ of $|z_i|^2$ with $n^{-1/2}$ in the above inequality, and taking $v = 2$, it is obvious that

$$\sum_{i=1}^{p} \mathrm{E}\left(n^{-1/2}|z_i|^2\right)^2 \leq |O(1)|.$$

Thus, $\mathrm{E}(-a_n\beta_k(\boldsymbol{X}_k^\top R_0^{-1}\boldsymbol{X}_k)^2)$, the expectation of the first term of (A14), has the order of $O(1)$, which concludes the first argument of Lemma A.6.

Next, we consider the third argument of Lemma A.6, i.e., (A5). We have

$$\mathrm{tr}\left(R_0^{-1}\Sigma\right) - \mathrm{tr}\left(R_n^{-1}\Sigma\right) = \mathrm{tr}\left(R_0^{-1} - R_0^{-1} + a_n\beta_k R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1} - \Upsilon_k\right)\Sigma.$$

As it has been shown above that

$$\mathrm{tr}\left(a_n\beta_k R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\right)\Sigma = a_n\beta_k\boldsymbol{X}_k^\top R_0^{-1}\Sigma R_0^{-1}\boldsymbol{X}_k = O_{L_1}(1),$$

we only need to find the order of $\mathrm{tr}(\Upsilon_k\Sigma)$. By the first term of $\mathrm{tr}(\Upsilon\Sigma)$ in (A9), we have

$$\begin{aligned}\mathrm{tr}\left(n^{-1}R_1^{-1}\boldsymbol{X}_k\left(1 - n^{-1}\boldsymbol{X}_k^\top R_1^{-1}\bar{\boldsymbol{X}}_{k0}\right)\bar{\boldsymbol{X}}_{k0}^\top R_1^{-1}\Sigma\right) &= n^{-1}\bar{\boldsymbol{X}}_{k0}^\top R_1^{-1}\Sigma R_1^{-1}\boldsymbol{X}_k\left(1 - n^{-1}\boldsymbol{X}_k^\top R_1^{-1}\bar{\boldsymbol{X}}_{k0}\right) \\ &= O_{L_1}(n^{-1}),\end{aligned}$$

and we can also show that the rest terms are also $O_{L_1}(n^{-1})$. Thus, we obtain that $\mathrm{tr}\left(R_0^{-1}\Sigma\right) - \mathrm{tr}\left(R_n^{-1}\Sigma\right) = O_{L_1}(1)$.

We now prove the second argument of Lemma A.6, i.e., (A4). By the fact that

$$\begin{aligned}\operatorname{tr}\left(R_n^{-1}\Sigma R_n^{-1}\Sigma\right)-\operatorname{tr}\left(R_0^{-1}\Sigma R_0^{-1}\Sigma\right)=&\\ \operatorname{tr}\big[&-a_n\beta_k R_0^{-1}\Sigma R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\Sigma+R_0^{-1}\Sigma\Upsilon_k\Sigma\\ &-a_n\beta_k R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\Sigma R_0^{-1}\Sigma+\Upsilon_k\Sigma\Upsilon_k\Sigma\\ &-a_n\beta_k R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\Sigma\Upsilon_k\Sigma+\Upsilon_k\Sigma R_0^{-1}\Sigma\\ &+a_n^2\beta_k^2 R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\Sigma R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\Sigma\\ &-a_n\beta_k\Upsilon_k\Sigma R_0^{-1}\boldsymbol{X}_k\boldsymbol{X}_k^\top R_0^{-1}\Sigma\big],\end{aligned}$$

it follows that $\operatorname{tr}\left(R_n^{-1}\Sigma R_n^{-1}\Sigma\right)-\operatorname{tr}\left(R_0^{-1}\Sigma R_0^{-1}\Sigma\right)=O_{L_1}(1)$, which completes the proof of (A4).

Proof of Theorem 3. In view of Lemma A.2 and Lemmas A.2–A.6, Theorem 3 is a natural extension by applying the Slutsky's Theorem, as

$$\frac{d_k^2(\bar{\boldsymbol{X}},S_n(\lambda))-\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma\right)}{\sqrt{2\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma\right)^2}}=(C_1+C_2+C_3+C_4)\times C_5,$$

where

$$C_1=\frac{d_k^2\left(\bar{\boldsymbol{X}},S_n(\lambda)\right)-d_k^2\left(\bar{\boldsymbol{X}}_{k0},S_0(\lambda)\right)}{\sqrt{2\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma\right)^2}},\quad C_2=\frac{d_k^2\left(\bar{\boldsymbol{X}}_{k0},S_0(\lambda)\right)-d_k^2\left(\boldsymbol{\mu},S_0(\lambda)\right)}{\sqrt{2\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma\right)^2}},$$

$$C_3=\frac{d_k^2\left(\boldsymbol{\mu},S_0(\lambda)\right)-\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma\right)}{\sqrt{2\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma\right)^2}},C_4=\frac{\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma\right)-\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma\right)}{\sqrt{2\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma\right)^2}},C_5=\frac{\sqrt{2\operatorname{tr}\left(S_0(\lambda)^{-1}\Sigma\right)^2}}{\sqrt{2\operatorname{tr}\left(S_n(\lambda)^{-1}\Sigma\right)^2}}.$$

Proof of Proposition 1. We first consider the moment generating function,

$$M(\boldsymbol{T})=\mathrm{E}\left(e^{\boldsymbol{T}^\top\boldsymbol{X}_1}\mid w_1=1\right).\tag{A15}$$

Following the discussion about $\tilde{d}^2(\boldsymbol{\mu},S(\lambda))$ in the proof of Lemma A.4, we let $V=\Sigma^{1/2}S(\lambda)^{-1}\Sigma^{1/2}$.

Assume that $V = Q^\top \Lambda Q$, where $Q^\top Q = I_p$ and $\Lambda = \mathrm{diag}\,(\zeta_1, \ldots, \zeta_p)$. We have

$$\begin{aligned} M(\boldsymbol{T}) &= \frac{1}{1-\delta}\,\mathrm{E}\left\{e^{\boldsymbol{T}^\top \boldsymbol{X}_1}\mathrm{I}\,(w_1 = 1)\right\} \\ &= \frac{1}{1-\delta}\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\int_{\left\{(\boldsymbol{X}_1-\boldsymbol{\mu})^\top S(\lambda)^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu})\le a_\delta\right\}} \exp\left\{\boldsymbol{T}^\top \boldsymbol{X}_1 - (\boldsymbol{X}_1-\boldsymbol{\mu})^\top\Sigma^{-1}(\boldsymbol{X}_1-\boldsymbol{\mu})/2\right\} d\boldsymbol{X}_1 \\ &= \frac{1}{1-\delta}\frac{1}{(2\pi)^{p/2}}e^{\boldsymbol{T}^\top\boldsymbol{\mu}+\boldsymbol{T}^\top\Sigma\boldsymbol{T}/2}\int_{\left\{\boldsymbol{z}^\top\Lambda\boldsymbol{z}\le a_\delta\right\}} \exp\left\{-\left(\boldsymbol{z}-Q\Sigma^{1/2}\boldsymbol{T}\right)^\top\left(\boldsymbol{z}-Q\Sigma^{1/2}\boldsymbol{T}\right)/2\right\} d\boldsymbol{z} \\ &= \frac{1}{1-\delta}e^{\boldsymbol{T}^\top\boldsymbol{\mu}+\boldsymbol{T}^\top\Sigma\boldsymbol{T}/2}F_{\boldsymbol{T}}\,(a_\delta)\,, \end{aligned} \tag{A16}$$

where $\boldsymbol{z} = Q\Sigma^{-1/2}(\boldsymbol{X}_1-\boldsymbol{\mu})$, and $F_{\boldsymbol{T}}(a)$ is the cumulative distribution function of the non-negative definite quadratic form in non-central normal variables, that is

$$F_{\boldsymbol{T}}(a) = \mathrm{P}\left(\boldsymbol{Z}_{\boldsymbol{v}}^\top\Lambda\boldsymbol{Z}_{\boldsymbol{v}} \le a\right), \quad \boldsymbol{Z}_{\boldsymbol{v}} \sim N\left(\boldsymbol{v}, I_p\right), \quad \boldsymbol{v} = Q\Sigma^{1/2}\boldsymbol{T}.$$

Without loss of generality, we prove the proposition for $x_{11} \mid w_1 = 1$, whose moment generating function is

$$m_1\,(t_1) = \mathrm{E}\left(e^{t_1 x_{11}} \mid w_1 = 1\right).$$

In (A15), let $\boldsymbol{T} = (t_1, 0, \ldots, 0)^\top$ with $p-1$ components of 0s. Then, it follows from (A16) that

$$m_1\,(t_1) = \frac{1}{1-\delta}e^{t_1\mu_1+\sigma_{11}t_1^2/2}F_{t_1}\,(a_\delta)\,,$$

where

$$F_{t_1}(a_\delta) = \frac{1}{(2\pi)^{p/2}}\int_{\left\{\boldsymbol{z}^\top\Lambda\boldsymbol{z}\le a_\delta\right\}} \exp\left\{-\left(\boldsymbol{z}-t_1\boldsymbol{v}_1\right)^\top\left(\boldsymbol{z}-t_1\boldsymbol{v}_1\right)/2\right\} d\boldsymbol{z},$$

$\boldsymbol{v}_1$ is the first row of $Q\Sigma^{1/2}$ and $\boldsymbol{v}_1^\top\boldsymbol{v}_1 = \sigma_{11}$. Since a_δ is the upper δ-quantile of $d_k^2\,(\boldsymbol{\mu}, S(\lambda))$, by the Berry-Esseen inequality, we have

$$\frac{a_\delta - \mathrm{tr}\,(S(\lambda)^{-1}\Sigma)}{\sqrt{2\,\mathrm{tr}\,(S(\lambda)^{-1}\Sigma)^2}} = z_\delta + o(1).$$

It is straightforward to show that

$$F_{t_1}(a_\delta)|_{t_1=0} = \frac{1}{(2\pi)^{p/2}} \int_{\left\{\boldsymbol{z}^\top \Lambda \boldsymbol{z} \le a_\delta\right\}} \exp\left(-\boldsymbol{z}^\top \boldsymbol{z}/2\right) d\boldsymbol{z} = \mathrm{P}\left\{d_k^2(\boldsymbol{\mu}, S(\lambda)) \le a_\delta\right\},$$

$$\begin{aligned}
&\left.\frac{\partial F_{t_1}(a_\delta)}{\partial t_1}\right|_{t_1=0} \\
=& \frac{1}{(2\pi)^{p/2}} \int_{\left\{\boldsymbol{z}^\top \Lambda \boldsymbol{z} \le a_\delta\right\}} \left(\boldsymbol{v}_1^\top \boldsymbol{z} - \boldsymbol{v}_1^\top \boldsymbol{v}_1 t_1\right) \exp\left\{-\left(\boldsymbol{z} - t_1 \boldsymbol{v}_1\right)^\top \left(\boldsymbol{z} - t_1 \boldsymbol{v}_1\right)/2\right\} d\boldsymbol{z} \Bigg|_{t_1=0} \\
=& \frac{1}{(2\pi)^{p/2}} \int_{\left\{\boldsymbol{z}^\top \Lambda \boldsymbol{z} \le a_\delta\right\}} \left(\boldsymbol{v}_1^\top \boldsymbol{z}\right) \exp\left(-\boldsymbol{z}^\top \boldsymbol{z}/2\right) d\boldsymbol{z} = 0,
\end{aligned}$$

and

$$\begin{aligned}
&\left.\frac{\partial^2 F_{t_1}(a_\delta)}{\partial t_1^2}\right|_{t_1=0} \\
=& \frac{1}{(2\pi)^{p/2}} \int_{\left\{\boldsymbol{z}^\top \Lambda \boldsymbol{z} \le a_\delta\right\}} \left\{\left(\boldsymbol{v}_1^\top \boldsymbol{z} - \boldsymbol{v}_1^\top \boldsymbol{v}_1 t_1\right)^2 - \boldsymbol{v}_1^\top \boldsymbol{v}_1\right\} \exp\left\{-\left(\boldsymbol{z} - t_1 \boldsymbol{v}_1\right)^\top \left(\boldsymbol{z} - t_1 \boldsymbol{v}_1\right)/2\right\} d\boldsymbol{z} \Bigg|_{t_1=0} \\
=& \frac{1}{(2\pi)^{p/2}} \int_{\left\{\boldsymbol{z}^\top \Lambda \boldsymbol{z} \le a_\delta\right\}} \left(\sum_{j=1}^{p} v_{1j}^2 z_j^2\right) \exp\left(-\boldsymbol{z}^\top \boldsymbol{z}/2\right) d\boldsymbol{z} - \sigma_{11} \mathrm{P}\left(d_k^2(\boldsymbol{\mu}, S(\lambda)) \le a_\delta\right) \\
=& -\sigma_{11} \mathrm{P}\left\{d_k^2(\boldsymbol{\mu}, S(\lambda)) \le a_\delta\right\} + \sum_{j=1}^{p} v_{1j}^2 \left[\Phi\left\{\frac{a_\delta - \operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)}{\sqrt{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}}\right\} - 2\phi\left\{\frac{a_\delta - \operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)}{\sqrt{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}}\right\}\right. \\
&\left.\left\{\frac{\zeta_j}{\sqrt{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}} + \frac{a_\delta - \operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)}{\sqrt{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}} \frac{\zeta_j^2}{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}\right\} + o(1)\right].
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\mathrm{E}\left(x_{11} \mid w_1 = 1\right) &= \left.\frac{\partial m_1(t_1)}{\partial t_1}\right|_{t_1=0} \\
&= \frac{1}{1-\delta}\left\{\mu_1 F_{t_1}(a_\delta)|_{t_1=0} + \left.\frac{\partial F_{t_1}(a_\delta)}{\partial t_1}\right|_{t_1=0}\right\} \\
&= \mu_1,
\end{aligned}$$

and

$$\mathrm{Var}\left(x_{11} \mid w_1 = 1\right) = \left.\frac{\partial^2 m_1(t_1)}{\partial t_1^2}\right|_{t_1=0} - \mu_1^2 = \sigma_{11} + \frac{1}{1-\delta}\left.\frac{\partial^2 F_{t_1}(a_\delta)}{\partial t_1^2}\right|_{t_1=0}.$$

Finally, we have

$$
\begin{aligned}
&\operatorname{Var}\left(x_{11} \mid w_1 = 1\right) \\
&= \frac{1}{1-\delta} \sum_{j=1}^{p} v_{1j}^2 \left\{ (1-\delta) - 2\phi\left(z_\delta\right) \left(\frac{\zeta_j}{\sqrt{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}} + z_\delta \frac{\zeta_j^2}{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2} \right) + o(1) \right\} \\
&= \frac{1}{1-\delta} \sum_{j=1}^{p} v_{1j}^2 \left\{ (1-\delta) - 2\phi\left(z_\delta\right) \frac{\zeta_j}{\sqrt{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}} + o(1) \right\} \\
&= \sigma_{11} \left\{ 1 - \frac{2\phi\left(z_\delta\right)}{1-\delta} \frac{\left(\Sigma S(\lambda)^{-1}\Sigma\right)_{11}}{\sigma_{11}\sqrt{2\operatorname{tr}\left(S(\lambda)^{-1}\Sigma\right)^2}} + o(1) \right\} \\
&= \sigma_{11}\tau_1,
\end{aligned}
$$

which completes the proof.

Acknowledgements

The authors thank the editor, an associate editor and two anonymous referees for their insightful comments and constructive suggestions. This work is partially supported by the National Natural Science Foundation (Grants 71873128, 72111530199, 12231017, 71921001), Natural Science Foundation of Anhui Province of China (Grants 2108085J02) and Natural Sciences and Engineering Research Council of Canada (Grant RGPIN-2017-05720).

Supplementary material

Supplementary material available online includes additional simulation results and a real data example.

References

Adrover, J. and Yohai, V.J. (2002). Projection estimates of multivariate location. *The Annals of Statistics* **30**, 1760–1781.

Alqallaf, F., Van Aelst, S., Yohai, V.J. and Zamar, R.H. (2009). Propagation of outliers in multivariate data. *The Annals of Statistics* **37**, 311–331.

Bai, Z. and Silverstein, J.W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability* **32**, 553–605.

Bai, Z. and Silverstein, J.W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. 2nd Edition. Springer, New York.

Boudt, K., Rousseeuw, P.J., Vanduffel, S. and Verdonck T. (2019). The minimum regularized covariance determinant estimator. *Statistics and Computing* **30**, 113–128.

Cator, E. and Lopuhaä, H. (2012). Central limit theorem and influence function for the MCD estimator at general multivariate distributions. *Bernoulli* **18**, 520–551.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association* **105**, 147–156.

Chen, L., Paul, D., Prentice, R.L. and Wang, P. (2011). A regularized Hotelling's T^2 test for pathway analysis in proteomic studies. *Journal of the American Statistical Association* **106**, 1345–1360.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis* **71**, 161–190.

Esbensen, K., Midtgaard, T. and Schönkopf, S. (1996). *Multivariate Analysis in Practice: A Training Package*. Camo As, Oslo.

Filzmoser, P., Maronna, R. and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics and Data Analysis* **52**, 1694–1711.

Ha, G., Zhang, Q., Bai, Z. and Wang, Y. (2021). Ridgelized Hotelling's T^2 test on mean vectors of large dimension. *Random Matrices:Theory and Applications* **11**, 2250011.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (2005). *Robust Statistics: The Approach Based on*

Influence Functions. Wiley, New York.

Hardin, J. and Rocke, D.M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics* **14**, 910–927.

Li, C. and Jin, B. (2022). Outlier detection via a block diagonal product estimator. *Journal of Systems Science and Complexity* **35**, 1929–-1943.

Lopuhaä, 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.

Maronna, R.A., Martin, R.D., Yohai, V.J. and Salibián-Barrera, M. (2019). *Robust Statistics Theory and Methods (with R)*. 2nd Edition. Wiley, Oxford.

Pison, G., Van Aelst, S. and Willems, G. (2002). Small sample corrections for LTS and MCD. *Metrika* **55**, 111–123.

Ro, K., Zou, C., Wang, Z. and Yin, G. (2015). Outlier detection for high-dimensional data. *Biometrika* **102**, 589–599.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* **8**, 283–297.

Rousseeuw, P.J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association* **88**, 1273–-1283.

Rousseeuw, P.J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–223.

Tao, T. and Vu, V. (2012). Random matrices: universality of local statistics of eigenvalues. *The Annals of Probability* **40**, 1285–1315.

Wang, T., Yang, X., Guo, Y. and Li, Z. (2021). Identification of outlying observations for large-dimensional data. *Journal of Applied Statistics*. DOI: 10.1080/02664763.2021.1993799.

Yang, X., Wang, Z. and Zi, X. (2018). Thresholding-based outlier detection for high-dimensional data. *Journal of Statistical*

Computation and Simulation **88**, 2170–2184.

Chikun Li

Department of Statistics and Finance, The School of Management, University of Science and Technology of China, Hefei, China.

E-mail: ahtclzk@mail.ustc.edu.cn

Baisuo Jin

Department of Statistics and Finance, The School of Management, University of Science and Technology of China, Hefei, China.

E-mail: jbs@ustc.edu.cn

Yuehua Wu

Department of Mathematics and Statistics, York University, Toronto, Canada.

E-mail: wuyh@yorku.ca