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

Supplementary Material

This is the Supplementary Material for the main paper Li et al. (2022), hereafter referred to as the main text. We present additional simulation results and a real-data example.

S1 Additional simulation results

To further examine the properties of the proposed procedure (RICD), we conduct the following additional simulations and also make a comparison with other three methods mentioned in Section 3.1 of the main text.

Case (e) (Laplace distribution): Let p-dimensional random vector $\boldsymbol{\xi}$ be composed of i.i.d. elements with the common density function $f(\nu) = \frac{\sqrt{2}}{2}e^{-\sqrt{2}|\nu|}$. Denote the distribution of $\boldsymbol{\xi}$ by F_{ξ}. The simulated observations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independently sampled from an ϵ contaminated distribution

$(1-\epsilon)F_\xi + \frac{1}{2}\epsilon N_p(\kappa\boldsymbol{\eta}_i, I_p) + \frac{1}{2}\epsilon N_p(-\kappa\boldsymbol{\eta}_i, I_p)$; $\kappa = 8, 9$, or 10 respectively for $p = 100, 200$, or 400. Two settings of $\boldsymbol{\eta}_i$s are explored (Cases (i) and (ii)), where $\boldsymbol{\eta}_i$ are defined at the beginning of Section 3.1 in the main text. It can be shown that F_ξ has excess peakedness and sub-exponential tail in this case. Note that the first three moments of ξ_1, the first element of $\boldsymbol{\xi}$, match those of the standard normal distribution, but the fourth moment is 6.

Case (f) (Dependent data): At first, each simulated observation $\boldsymbol{x}_i$ is generated by $\boldsymbol{y}_{i+2} + 0.4\boldsymbol{y}_{i+1} + 0.25\boldsymbol{y}_i$, where $\boldsymbol{y}_i$ are independently drawn from $N_p(\mathbf{0}, I_p)$. Then $[\epsilon n]$ observations among $\boldsymbol{x}_i$ are randomly replaced by outliers from $N_p(\kappa\boldsymbol{\eta}_i, I_p)$ and $N_p(-\kappa\boldsymbol{\eta}_i, I_p)$ in each half; $\kappa = 9, 11$, or 14 respectively for $p = 100, 200$, or 400.

Case (g): Let p-dimensional random vector $\boldsymbol{\xi} = 0.6908\boldsymbol{\gamma} + 0.7230\boldsymbol{\nu}$, 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

$$f(\nu) = \begin{cases} \frac{a}{2b}(\frac{\nu}{b})^{-a-1}, & \text{if } \nu \geq b, \\ \frac{a}{2b}(\frac{-\nu}{b})^{-a-1}, & \text{if } \nu < -b, \\ 0, & \text{else.} \end{cases}$$

The two parameters a and b in the above density function are chosen as $2 + 4/\sqrt{3}$ and $\sqrt{2/(2+\sqrt{3})}$, respectively so that Condition A5 in Section

2.1 of the main text holds for ξ_1; $\kappa = 8, 9$, or 10 respectively for $p = 100, 200$, or 400.

Simulation results under Cases (e)-(g), (i)-(ii) for various values of p at 5% significance, $\epsilon = 0.1$ and 0.2 are summarized in Tables S1 and S2, respectively. By these two tables, it can be seen that the proposed method achieves the highest detection power against sparse signals (Case (ii)) among the four methods, even when there are some dependence among data. For example, detection power of the proposed method is 82.46% and 80.15% in Case (ii)(f) for $n = 100$, $p = 400$, $\epsilon = 0.1$, and 0.2, which is almost twice as much as detection power of the RMDP method in the same settings. The PCout procedure performs fairly well in Case (i)(e), but it still appears to be insensitive to sparse signals as discussed in the main text. In contrast, for dense mean vector Case (i), the proposed method can also maintain a desired efficiency.

Next, simulation results under Cases (a)-(g), (i)-(ii) for $p = 400$, $n = 40$ at 5% significance, $\epsilon = 0.1$ and 0.2 are reported in Tables S3 and S4, respectively; $\kappa = 12, 18, 12, 12, 15, 12$ respectively for Cases (a)-(c), (e)-(g). Overall, the detection power of the proposed method does not vary much when p/n increases to 10/1. The empirical size of the proposed method is adequately controlled in most cases, though it is a little conservative when

p/n becomes much larger than in the simulation settings of the main text. The RMDP and the BDP methods both do not perform well in terms of the test size in Cases (a)-(c), (i)-(ii).

For different values of p/n, the empirical sizes of all the four methods except the PCout exceed the significance level in Cases (e), (i)-(ii), and are conservative in Cases (f), (i)-(ii). These are mainly due to that the detection rule of the RICD relies on some moment and i.i.d. conditions, which are not satisfied in these cases, and that the RMDP and BDP are both detection procedures for the data i.i.d. from normal distributions.

Finally, for examining the performance of a different h, we replace the subset size $h_{\text{default}} = [n/2] + 1$ at Step 1 of Algorithm 2 in the main text by $h_{\text{default}} = [(\alpha + 0.5)n]$. We consider the setting in Case (c) and denote the simulation results under this setting but with $h = [(\alpha + 0.5)n]$ as Case (c'). Simulation results under Cases (c) and (c'), (i)-(ii) for various α, p, and ϵ are reported in Tables S5 and S6. Results from these two tables show that there is no noticeable difference in terms of empirical size and power between these two choices of default subset sizes.

Table S1: Average type-I error ($\bar{\alpha}$ %) and detection power ($\bar{\beta}$ %) under Cases (e)-(g), (i)-(ii) with $n = 100, p = 100, 200, 400$, at significance level 0.05 and contamination ratio 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)	(e)	100	21.22	98.92	22.69	99.37	23.84	98.50	5.53	99.18
		200	20.52	98.41	24.40	99.11	26.95	98.25	4.89	99.49
		400	19.47	96.04	25.73	98.11	28.13	97.62	4.67	99.10
	(f)	100	3.79	91.00	3.25	87.66	3.48	82.84	5.08	96.44
		200	2.15	86.54	1.53	78.32	2.07	76.67	4.37	94.89
		400	0.63	80.37	0.27	63.62	0.62	67.31	4.41	93.51
	(g)	100	3.21	97.15	3.20	96.68	3.66	94.71	4.61	99.31
		200	3.37	96.20	3.29	95.13	4.24	93.34	4.28	99.65
		400	3.61	93.94	3.11	91.83	5.05	90.61	4.22	99.15
(ii)	(e)	100	21.21	99.19	22.95	96.45	23.97	96.77	8.12	43.61
		200	20.57	98.36	24.35	94.93	27.06	95.11	7.78	32.01
		400	19.43	96.32	25.77	91.40	28.43	90.93	7.51	22.99
	(f)	100	3.81	90.74	3.36	78.84	3.64	83.88	7.79	29.60
		200	2.15	87.29	1.88	65.06	2.42	74.01	7.51	20.87
		400	0.63	82.46	0.59	41.84	1.04	56.05	7.15	33.24
	(g)	100	3.23	97.12	3.17	94.66	3.82	94.77	6.76	37.43
		200	3.38	96.31	3.25	92.56	4.35	93.41	6.95	28.06
		400	3.59	93.27	3.16	85.11	5.21	85.20	7.28	21.23

Table S2: Average type-I error ($\bar{\alpha}$ %) and detection power ($\bar{\beta}$ %) under Cases (e)-(g), (i)-(ii) with $n = 100, p = 100, 200, 400$, at significance level 0.05 and contamination ratio 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)	(e)	100	18.34	98.51	18.81	98.91	19.71	96.93	1.91	99.92
		200	17.72	97.65	20.42	98.68	22.59	96.70	1.60	99.99
		400	16.04	92.71	20.95	96.86	22.52	94.99	1.50	100.00
	(f)	100	2.87	87.32	2.50	83.97	2.46	75.07	1.74	100.00
		200	1.66	82.22	1.13	74.32	1.36	65.34	1.20	99.86
		400	0.54	74.72	0.20	57.09	0.41	50.76	1.18	100.00
	(g)	100	2.40	96.19	2.46	95.59	2.72	92.79	1.53	99.67
		200	2.55	94.85	2.42	93.78	3.43	90.76	1.30	99.69
		400	2.62	91.96	2.35	89.83	3.81	86.84	1.29	99.83
(ii)	(e)	100	18.32	98.63	18.99	94.25	19.60	94.55	6.52	37.71
		200	17.63	97.51	20.80	92.14	22.99	92.10	6.30	27.43
		400	15.65	93.78	21.84	87.81	24.13	86.45	6.78	21.34
	(f)	100	2.88	88.77	2.94	75.01	2.94	79.77	6.35	25.43
		200	1.69	83.87	2.00	60.47	2.31	68.13	7.10	17.23
		400	0.57	80.15	1.04	41.17	1.31	52.83	7.33	17.66
	(g)	100	2.39	96.59	2.49	93.55	2.77	93.83	5.26	35.25
		200	2.56	95.13	2.50	90.23	3.42	90.24	5.60	26.16
		400	2.58	91.62	2.45	82.65	4.15	81.63	6.30	20.16

Table S3: Average type-I error (%) and detection power (%) under Cases (a)-(g), (i)-(ii) with $n = 40, p = 400$, at significance level 0.05 and contamination ratio 0.1.

		Average type-I error				Detection power			
Case	$\boldsymbol{\eta}_i$	RICD	RMDP	BDP	PCout	RICD	RMDP	BDP	PCout
(a)	(i)	2.78	11.77	19.96	5.64	92.75	98.67	96.21	88.95
	(ii)	2.45	11.56	19.91	8.50	93.53	75.72	70.12	39.09
(b)	(i)	0.88	11.30	18.45	6.35	49.60	96.26	92.69	80.88
	(ii)	0.88	12.33	20.19	9.23	52.30	71.22	67.51	30.35
(c)	(i)	1.86	10.68	19.41	5.23	94.37	99.64	97.92	94.68
	(ii)	1.83	10.92	19.74	8.35	94.32	75.21	71.07	46.38
(d)	(i)	—	—	—	—	—	—	—	—
	(ii)	2.48	11.65	18.51	8.85	83.59	91.05	86.38	36.23
(e)	(i)	9.82	35.11	38.53	6.74	95.97	99.83	99.32	90.77
	(ii)	9.76	33.85	36.94	9.36	97.11	88.06	80.69	45.93
(f)	(i)	0.08	0.14	1.37	7.87	87.51	72.66	93.54	67.76
	(ii)	0.11	0.28	2.50	10.22	89.08	33.84	55.52	41.75
(g)	(i)	1.34	5.70	14.09	5.65	93.13	99.33	97.42	92.43
	(ii)	1.38	5.59	14.53	8.61	93.25	74.56	70.58	41.07

Table S4: Average type-I error (%) and detection power (%) under Cases (a)-(g), (i)-(ii) with $n = 40, p = 400$, at significance level 0.05 and contamination ratio 0.2.

		Average type-I error				Detection power			
Case	$\boldsymbol{\eta}_i$	RICD	RMDP	BDP	PCout	RICD	RMDP	BDP	PCout
(a)	(i)	2.35	10.58	15.52	2.45	89.29	98.40	94.72	97.12
	(ii)	2.07	11.14	18.51	7.27	90.58	70.37	66.32	37.16
(b)	(i)	0.43	8.96	14.13	2.68	38.04	93.47	86.17	95.83
	(ii)	0.46	10.81	18.82	8.24	43.19	67.31	64.73	27.12
(c)	(i)	1.91	10.28	15.66	1.95	91.35	99.28	95.65	99.49
	(ii)	1.62	9.95	18.13	6.40	91.26	72.09	68.20	39.43
(d)	(i)	—	—	—	—	—	—	—	—
	(ii)	1.95	8.72	14.21	7.56	79.80	87.58	82.70	31.95
(e)	(i)	8.40	30.80	31.54	2.76	92.58	99.90	97.49	99.02
	(ii)	7.50	29.04	32.40	8.17	92.67	82.92	75.04	39.82
(f)	(i)	0.13	0.17	1.20	3.38	84.48	74.46	86.71	93.44
	(ii)	0.13	1.18	3.89	9.74	88.79	37.38	54.63	27.66
(g)	(i)	1.23	5.41	10.71	2.31	91.04	99.06	95.12	99.01
	(ii)	1.16	5.39	13.07	7.17	90.82	69.73	66.00	37.93

Table S5: Average type-I error (%) by the proposed procedure for various p, ϵ, and α under Cases (c) and (c').

			$\epsilon = 0.1$			$\epsilon = 0.2$		
$\boldsymbol{\eta}_i$	Case	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
(i)	(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
	(c')	100	1.57	5.97	10.88	1.16	4.58	8.80
		200	1.43	5.88	11.07	1.02	4.60	8.71
		400	1.28	5.90	10.64	0.84	4.54	7.86
(ii)	(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
	(c')	100	1.57	5.94	10.97	1.15	4.63	8.99
		200	1.46	5.73	10.91	1.03	4.68	8.87
		400	1.22	5.74	10.42	0.82	4.28	7.49

Table S6: Detection power (%) by the proposed procedure for various p, ϵ, and α under Cases (c) and (c').

			$\epsilon = 0.1$			$\epsilon = 0.2$		
$\boldsymbol{\eta}_i$	Case	p	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
(i)	(c)	100	93.04	97.80	98.85	89.23	96.84	98.61
		200	87.99	96.61	98.20	85.53	95.67	97.86
		400	81.47	93.76	97.13	77.35	92.22	95.67
	(c')	100	92.14	97.69	98.92	89.85	97.02	98.49
		200	89.06	96.72	98.38	85.41	95.46	97.77
		400	81.55	94.16	96.87	76.27	91.96	95.06
(ii)	(c)	100	92.59	97.66	99.16	89.82	96.79	98.55
		200	88.70	96.27	98.20	85.41	94.92	97.64
		400	82.79	94.27	96.62	78.43	93.09	95.87
	(c')	100	92.29	97.63	98.95	90.39	97.08	98.53
		200	89.27	96.84	98.51	85.97	95.56	97.86
		400	82.54	94.30	97.11	78.43	92.43	95.69

Table S7: Stock data set: Indices of the outliers detected by the four methods for various α values.

α	RICD	RMDP	BDP	PCout
0.01	2, 8, 9, 13, 22, 23	2, 8, 13, 19, 22	2, 7, 8, 13, 22	NA
0.05	2, 7–10, 13, 19, 22, 23	2, 7, 8, 13, 19, 21, 22	2, 7, 8, 13, 21, 22	NA
0.1	2, 7–10, 13, 19, 22, 23	2, 7, 8, 13, 19, 21, 22	2, 7, 8, 13, 21, 22	NA

S2 Example: Stock data

Consider weekly returns of 300 stocks, all constituents of the CSI 300 Index, from May 22, 2017 to December 19, 2017 (25 weeks), which were actively traded in the Shanghai Stock Exchange (SSE) and the Shenzhen Stock Exchange (SZSE). The returns are calculated in log-scale by $\log(P_t) - \log(P_{t-1})$, where P_t is the closing price of the tth week, $t = 1, \ldots, 25$. Thus, for this data set, n is 25 and p is 300. Note that the data can be downloaded from the Sina Finance database, which is publicly available.

To demonstrate the performance of the proposed detection procedure, we consider the original data and the data contaminated in the following way: When $t = 2, 8, 9, 13, 22$, and 23, the returns of the first stock, Ping An Bank, are falsely recorded as $\log(P_t)$, which implies that the data on these six weeks can be considered as outliers. Figure S1 displays the time

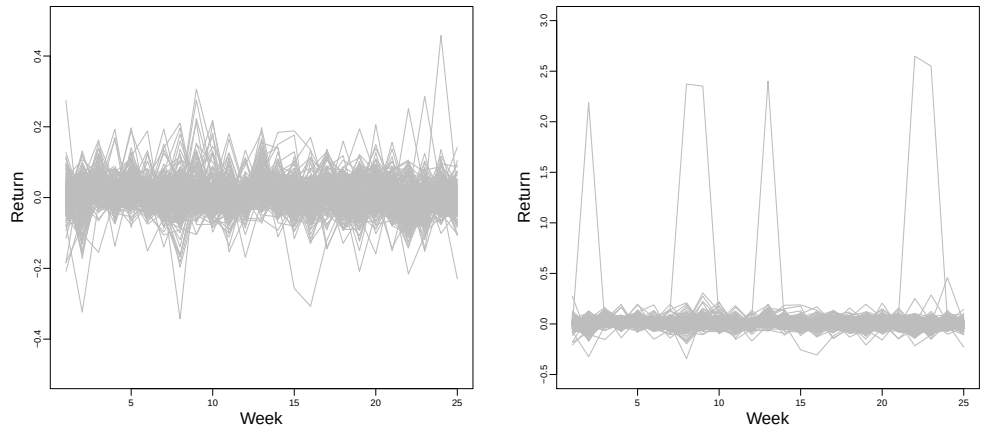

Figure S1: Time plots of the log weekly returns of all constituents of the CSI 300 Index from May 22, 2017 to December 19, 2017: The original data (left panel) and the contaminated data (right panel).

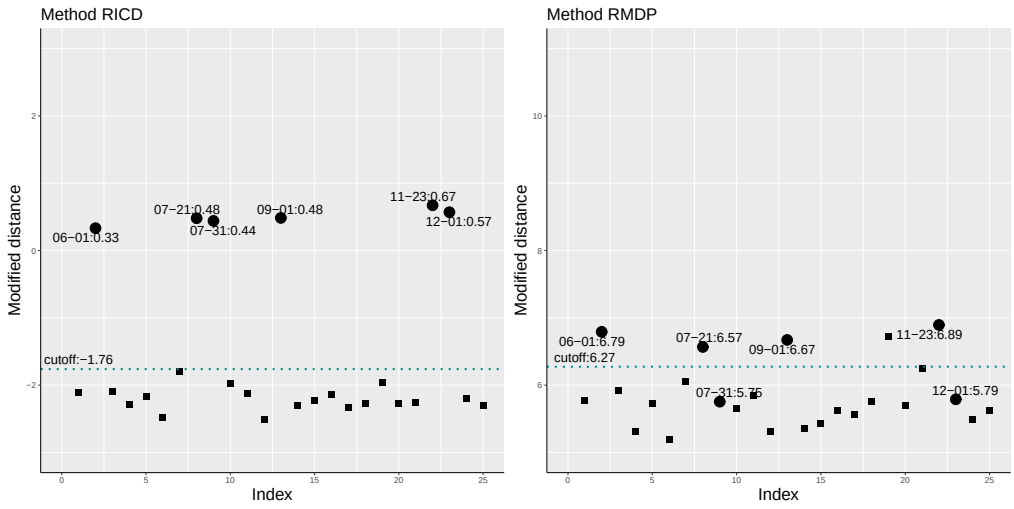

Figure S2: Plots of the modified distances based on the RICD and the RMDP.

plots of the log weekly returns of all constituents of the CSI 300 Index from May 22, 2017 to December 19, 2017 with and without contamination. We apply the aforementioned four methods to the contaminated stock data set at different significance levels and summarize the results in Table S7. It can be observed that the outliers 9 and 23 are missed by all methods except ours. The performance of the RICD at the significance level 0.01 is the best, with all six outliers detected and no false detection. We remark that the PCout procedure is not feasible for this contaminated data set.

Moreover, at a significance level of 0.01, the modified distances (in log-scale) based on the RICD and the RMDP are shown in Figure S2. These two distances with the corresponding cutoffs can be obtained by (2.16), (10), and Step 4 of Algorithm 2 in Section 2.4 of Ro et al. (2015), respectively. As depicted in Figure S2, the dashed horizontal line displays the cutoff values (in log-scale). "Good" weekly returns are marked by solid squares, while the outlying weekly returns are shown in solid circles with times and distances listed beside. This figure clearly demonstrates that the proposed procedure performs better than the RMDP in this example.

References

Li, C., Jin, B. and Wu, Y. (2022). Outlier Detection via a Minimum Ridge Covariance Determinant Estimator. *Submitted*.

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