Statistica Sinica: Supplement

Effect of heavy tails on ultra high dimensional variable ranking methods

Aurore Delaigle and Peter Hall

Department of Mathematics and Statistics, University of Melbourne, Australia.

Supplementary Material

S1 Additional simulation results

S1.1 Mean ranking

We generated p vectors of Xs and Ys from the model at (2.9), for a variety of variables ϵ_{kij} and values of μ_{kj}. In each case we produced 100 samples of size n. For each sample we ranked the p covariates according to either the values of $D_j = \bar{Y}_j - \bar{X}_j$, the values of the Student's t statistics T_j, or the values of $\bar{V}_j - \bar{U}_j$, as described in Section 2.3. We also implemented the method based on the $\bar{W}_j$s (see Remark 1), but we shall not discuss it here because it gave results very similar to, but was much slower than, the approach based on $\bar{V}_j - \bar{U}_j$. In the graphs below we show, for several representative examples, boxplots of the ranks obtained by each method for the relevant components. All boxplots were constructed from 100 samples, and in some figures we truncated the graphs to facilitate a distinction between the boxes. We considered three combinations of sample size and dimension: $(n, p) = (30, 8000)$, $(n, p) = (50, 20000)$ and $(n, p) = (200, 50000)$.

When the distributions are not heavy-tailed, transforming the data is not necessary and cannot be expected to improve results. However, in such cases, transforming the variables usually does not deteriorate the ranking much. In Section S1.3 we illustrate this fact by showing results of simulations in a simple model (see Figure 5). Below we consider more complicated models. We show only a limited summary of our results, but our conclusions were similar in the other examples we considered.

Case 1. Uniform distributions with outliers.
(a) We started with a simple model where the nonzero $\mu_{2j} - \mu_{1j}$ were all equal, and the ϵ_{1ij}s contained a small fraction $a = \lceil n/40 \rceil$ of moderate outliers. For $j = 1, \ldots, p$ we took $\mu_{1j} = 0$ and $\mu_{2j} = 2 \cdot 1_{\{j=1,\ldots,6\}}$ and defined $\mathcal{I}_j = \{j_1, \ldots, j_a\}$, where $j_1, \ldots, j_a$ are a numbers chosen at random among $1, \ldots, n$. With $\mathcal{I}_j$ defined in this way, for $i = 1, \ldots, n$, we took the ϵ_{1ij}s independent and distributed like the mixture $U[-10, 10] \cdot 1\{i \notin \mathcal{I}_j\} - U[14 + 2\mu_{1j}, 22 + 2\mu_{1j}] \cdot 1\{i \in \mathcal{I}_j\}$. The results shown in the first row of Figure 1 indicate very clearly that transforming the variables improved ranking considerably, compared to

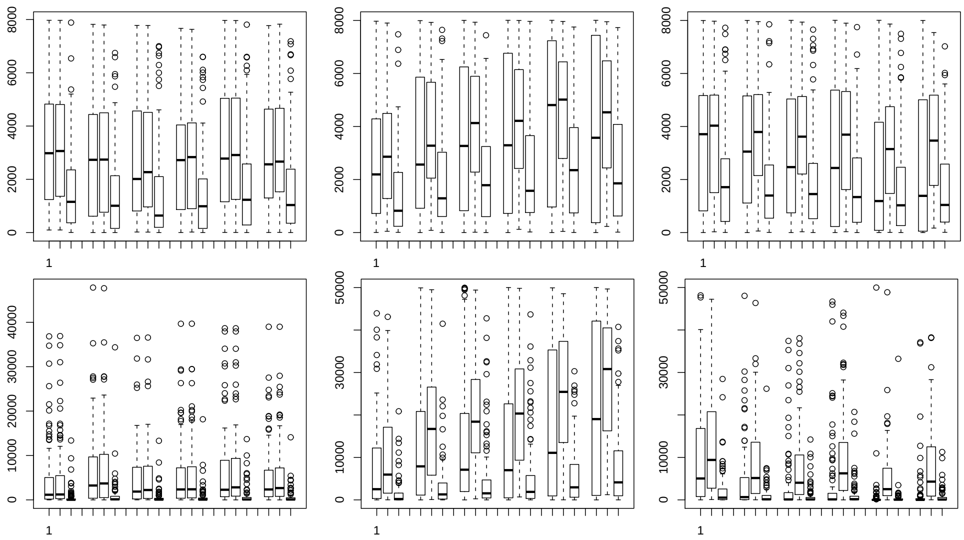

Figure 1: Boxplots of ranks for $\mu_{2j} - \mu_{1j}$ for $j = 1, \dots, 6$ in case 1(a) (first column), case 1(b) (second column), case 1(c) (third column) when $(n, p) = (30, 8000)$ (first row) and $(n, p) = (200, 50000)$ (second row). In each graph, the jth group of three boxes shows the ranks for the jth component based on the values of D_j (first boxplot), T_j (second boxplot) or $\bar{V}_j - \bar{U}_j$ (third boxplot).

both other approaches, since it assigned ranks closer to 1 for each relevant component.
(b) Next, to examine the impact of unequal variances among the relevant components, we took the same setting as in (a), except that, for $i = 1, \dots, n$, $k = 0, 1$ and $j = 1, \dots, 6$, we took the ϵ_{kij}s to be independent and distributed like the mixture $-U[14+2\mu_{1j}+5j, 22+2\mu_{1j}+5j] \cdot 1\{k = 1, i \in \mathcal{I}_j\} + U[-10-2j, 10+2j] \cdot 1\{k = 0 \text{ or } i \notin \mathcal{I}_j\}$. The graphs in the second row of Figure 1 illustrate the fact that Student's t ranking is strongly negatively influenced by the unequal variances, as noted in Section 2.3. By standardising for scale, Student's t is too focused on controlling fluctuations of rankings, and as a result it misses its target. Indeed, here the Student's t ranks are even further from 1 than the ranks based on the means $\bar{V}_{1j} - \bar{V}_{0j}$. The transformation method significantly improves the results and is much less affected by the unequal variances.
(c) Finally we took unequal μ_{kj}s and unequal variances: we used the same setting as (b), except that, for $j = 1, \dots, 6$, we took $\mu_{1j} = 2 + (j - 1)/2$. The results, shown in the third row of Figure 1, attract the same conclusions as for (b). See Figure 6 in Section S1.3 for results of simulations for cases (a) to (c) when $(n, p) = (50, 20000)$.

Case 2. Stable distributions. In our next example, for $j = 1, \dots, p$ we took $\mu_{1j} = 0$ and $\mu_{2j} = 1 \cdot 1_{\{j=1,\dots,6\}}$ and, for $i = 1, \dots, n$ and $k = 0, 1$, we took the ϵ_{kij}s to be independent with a symmetric stable distribution with parameters α and c, for several values of α (this distribution is heavy tailed and its variance does not exist). The results, shown in Figure 2, illustrate, once again, the superiority of the transformation approach.

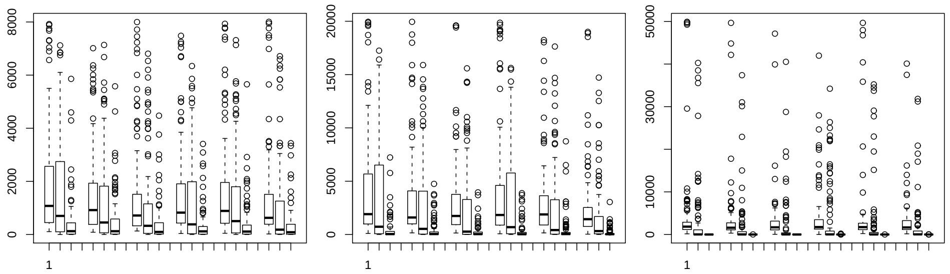

Figure 2: Boxplots of ranks for $\mu_{2j} - \mu_{1j}$ for $j = 1, \ldots, 6$ in case 2 when, from left to right, $(n, p) = (30, 8000)$, $(n, p) = (50, 20000)$ and $(n, p) = (200, 50000)$. In each graph, the jth group of three boxes shows the ranks for the jth component based on the values of D_j (first boxplot), T_j (second boxplot) or $\bar{V}_j - \bar{U}_j$ (third boxplot).

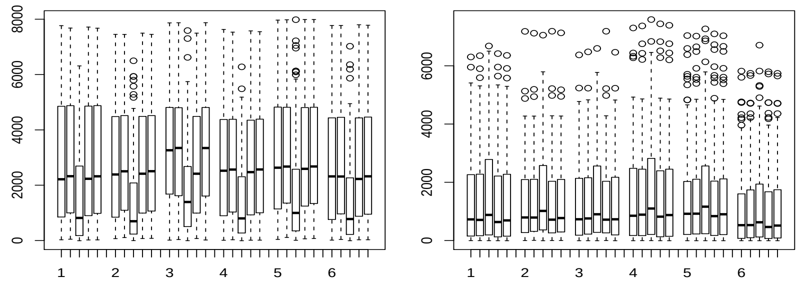

Figure 3: Boxplots of ranks for $\mu_{2j} - \mu_{1j}$ for $j = 1, \ldots, 6$ in case 1(a) (first column) and a light-tailed version of it (b) (second column), when $(n, p) = (30, 8000)$. In each graph, the jth group of five boxplots shows the ranks for the jth component based on the values of $\bar{Y}_j - \bar{X}_j$ (first boxplot), T_j (second boxplot), $\bar{V}_j - \bar{U}_j$ (third boxplot), Efron *et al.* (2001)'s criterion (fourth boxplot) and Opgen-Rhein *et al.* (2007)'s criterion (fifth boxplot).

Finally, we compared our method with variants of the t statistics proposed by Efron *et al.* (2001) and Opgen-Rhein *et al.* (2007). Like the other variants mentioned in Section 2.3, these two approaches are specifically designed to improve variance estimators by borrowing strength from other components, but not to systematically correct for outliers or heavy-tailed distributions. They are illustrated in Figure 3, where, for $(n, p) = (30, 8000)$, we compare the performance of various methods in case 1(a), where there are moderate outliers (in fact, exactly one outlier in the setting of Figure 3), and a light-tailed version of case 1(a), where we take the ϵ_{kij}s to be independent and identically distributed like $U[-10, 10]$. As we can see, the methods of Efron *et al.* (2001) and Opgen-Rhein *et al.* (2007) behave quite similarly to the student's t approach. They do not compete successfully with the variable transformation approach when the distributions are not light tailed (left column of Figure 3), whereas the transformation approach remained competitive even when the distributions were light-tailed (right column of Figure 3).

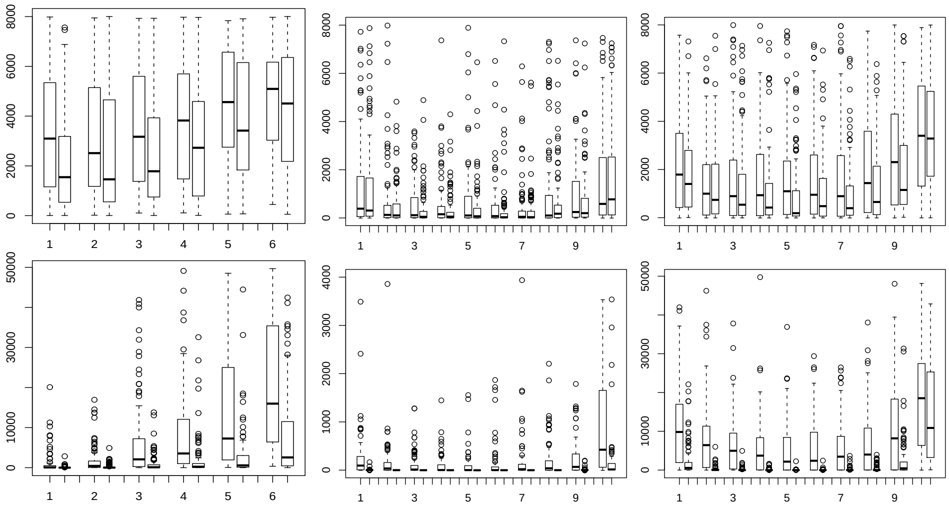

Figure 4: Boxplots of ranks for X_1 to X_6 in case 3 (first column), and for X_1 to X_{10} in cases 4(a) (second column) and 4(b) (third column) with $(\alpha, c) = (1.5, 1)$, when $(n, p) = (30, 8000)$ (first row) and $(n, p) = (200, 50000)$ (second row). In each graph, the jth group of two boxplots shows the ranks for the jth component based on the $|\hat{\rho}_j|$s (left boxplot) or on the $|\hat{\omega}_j|$s (right boxplot).

S1.2 Correlation ranking

We generated data from the model at (3.8) for a variety of distributions and values of β_j, in cases where the components were either dependent or independent. Below we present only a few cases, but we obtained similar conclusions in the other instances we considered. In each case we generated 100 samples of the form $(X_{i1}, \ldots, X_{ip}, Y_i)_{i=1,\ldots,n}$. For each sample, we ranked the p covariates according to the values of the untransformed correlations, $|\hat{\rho}_j|$, and of the transformed ones, $|\hat{\omega}_j|$. In the graphs we show boxplots of the ranks obtained, for the relevant components, using the two methods. For each component the boxplots were constructed from 100 samples.

As for the mean case, we started with a simple example involving light-tailed distributions to illustrate the fact that, in cases such as this, even though variable transformation is clearly not needed, transforming the data deteriorates the ranking a little, but the negative effect usually remains quite limited. See Figure 7 in Section S1.3. Below we consider more complicated models.

Case 3. Uniform distributions with unequal β_js and some moderate outliers. For $j = 1, \ldots, p$, let $\beta_j = \{1 - (j-1)/12\} \cdot 1_{\{j=1,\ldots,6\}}$ and let $\mathcal{I}_j = \{j_1, \ldots, j_a\}$, where $a = \lceil n/40 \rceil$, the smallest integer larger than or equal to $n/40$, and $j_1, \ldots, j_a$ are a numbers chosen at random among $1, \ldots, n$. For $i = 1, \ldots, n$, let ϵ_i be independent and identically distributed uniform $U[-10, 10]$, and let X_{ij} be independent and distributed like the mixture

$U[14,22]\cdot 1\{i \in \mathcal{I}_j\} + U[-10,10]\cdot 1\{i \notin \mathcal{I}_j\}$. Put $Y_i = \sum_{i \le 1 \le 6} \beta_j X_{ij} - L_{ij}\cdot 1\{i \in \mathcal{I}_j\} + \epsilon_i$, where the L_{ij}s are independent with distribution $U[40,50]$. Thus we introduced a few moderate outliers, corresponding to X_{ij}s where $i \in \mathcal{I}_j$. Here, the ideal ranks should be 1 to 6 for X_1 to X_6, but we can see from the second row of Figure 4 that the ranks based on the $|\hat{\rho}_j|$s are far from that, and using the $|\hat{\omega}_j|$s significantly improves the ranks. Although the distribution considered here contains only moderate outliers, the improvement obtained by data transformation is already quite impressive.

Case 4. Stable distributions with dependency between components. For $j = 1, \ldots, p$, let $\beta_j = 100 \cdot 1_{\{j=1,\ldots,10\}}$ and, for $i = 1, \ldots, n$, let the L_{ij}s and the ϵ_is be independent with a symmetric stable distribution with parameters α and c, that is, with characteristic function $\phi(t) = \exp(-c|t|^{\alpha})$. Then, let $Y_i = \sum_{1 \le i \le 10} \beta_j X_{ij} + \epsilon_i$, where, for the X_{ij}s, we consider two models: (a) dependence among relevant components: $X_{ij} = L_{ij} + 0.4\, L_{i,j-1} \cdot 1\{1 < j \le 10\}$; (b) dependence among all components: $X_{ij} = L_{ij} + 0.4\, L_{i,j-1}\cdot 1\{1 < j \le 10\} + 0.4 X_{i,\lceil 10j/p \rceil} \cdot 1\{j > 10\}$. Note that, here, the untransformed correlations do not exist, since the X_{ij}s do not have finite variance. Of course, we can calculate empirical correlations in finite samples, but we can expect them to be highly variable, since their theoretical counterparts are not well defined. In contrast, the correlations between pairs of transformed variables are well defined. We compared the two approaches for several values of α and c. In the last two rows of Figure 4 we present results for $(\alpha, c) = (1.5, 1)$. See Figure 8 for results when $(n, p) = (50, 20000)$. Unsurprisingly, in almost all cases the ranks obtained by transforming the data were much closer to 1 than the ranks obtained by untransformed correlation. This illustrates the superiority of the transformation approach in heavy-tailed contexts.

S1.3 Additional numerical results

Case 0 (Uniform distributions with equal means and variances).
For $j = 1, \ldots, p$ we took $\mu_{1j} = 0$ and $\mu_{2j} = 2 \cdot 1_{\{j=1,\ldots,6\}}$ and, for $i = 1, \ldots, n$ and $k = 0, 1$, we took the ϵ_{kij}s to be independent and identically distributed like $U[-10, 10]$. The boxplots in the first row of Figure 5 indicate that when the distributions are not heavy tailed and do not have outliers (and thus where variable transformation is not needed), transforming the variables does not deteriorate the ranking much, compared to the other two approaches to ranking.

Case 6 (Uniform distributions with unequal β_js). For $j = 1, \ldots, p$, let $\beta_j = \{1 - (j - 1)/12\} \cdot 1_{\{j=1,\ldots,6\}}$, and for $i = 1, \ldots, n$ let the X_{ij}s and ϵ_is be independent and identically distributed uniform $U[-10, 10]$. Then put $Y_i = \sum_{1 \le i \le 6} \beta_j X_{ij} + \epsilon_i$. This example illustrates situations where the distributions are very light-tailed and transforming the variables is not needed. Our goal is to investigate the negative impact of the transformation approach in such cases. Boxplots of the ranks obtained by the two methods are shown for the relevant components (i.e. $j = 1, \ldots, 6$) in Figure 7. Ideally, these six components should be ranked from 1 to 6, but because of the high dimension, a number of irrelevant components are ranked higher due to random fluctuations. This can be seen

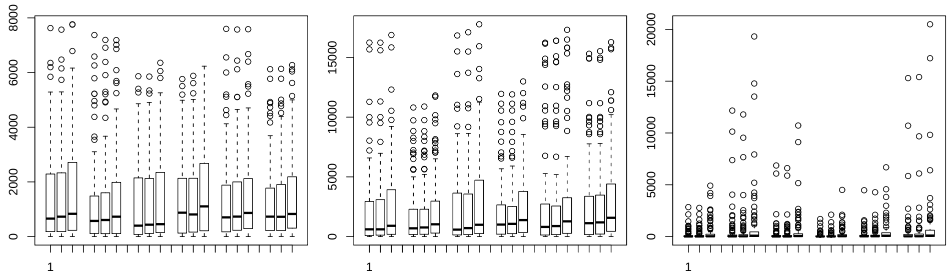

Figure 5: Boxplots of ranks for $\mu_{2j}-\mu_{1j}$ for $j=1,\ldots,6$ in case 0 when $(n,p)=(30,8000)$ (first column), $(n,p)=(50,20000)$ (second column) and $(n,p)=(200,50000)$ (third column). In each graph, the jth group of three boxes shows the ranks for the jth component based on the values of $\bar{Y}_j-\bar{X}_j$ (first boxplot), T_j (second boxplot) or $\bar{V}_j-\bar{U}_j$ (third boxplot).

from the boxplots, where the median empirical rank of each component is lower than its actual rank (1 to 6, for X_1 to X_6, respectively). In this case transforming the data deteriorated the ranking a little, but the negative effect remained quite limited. Overall, the order of importance of the six relevant components was respected by both methods.

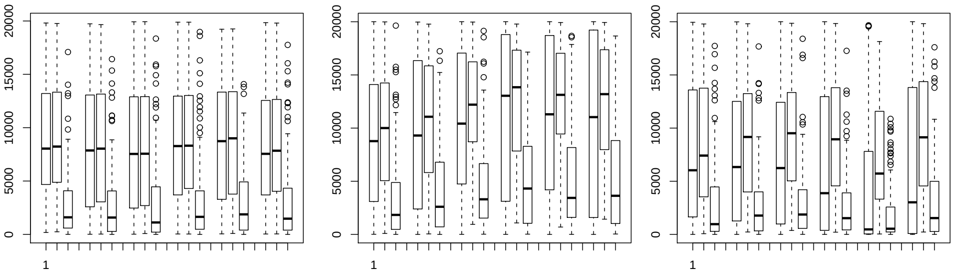

Figure 6: Boxplots of ranks for $\mu_{2j}-\mu_{1j}$ for $j=1,\ldots,6$ in case 1(a) (first column), case 1(b) (second column), case 1(c) (third column) when $(n,p)=(50,20000)$. In each graph, the jth group of three boxes shows the ranks for the jth component based on the values of D_j (first boxplot), T_j (second boxplot) or $\bar{V}_j-\bar{U}_j$ (third boxplot).

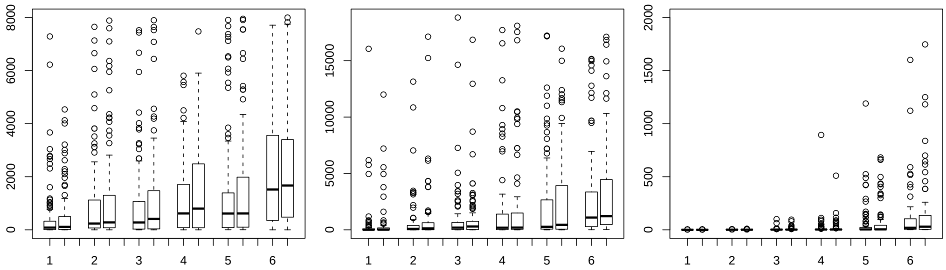

Figure 7: Boxplots of ranks for X_1 to X_6 in case 6 when $(n, p) = (30, 8000)$ (first column), $(n, p) = (50, 20000)$ (second column) and $(n, p) = (200, 50000)$ (third column). In each graph, the jth group of two boxplots shows the ranks for the jth component based on the $|\hat{\rho}_j|$s (left boxplot) or on the $|\hat{\omega}_j|$s (right boxplot).

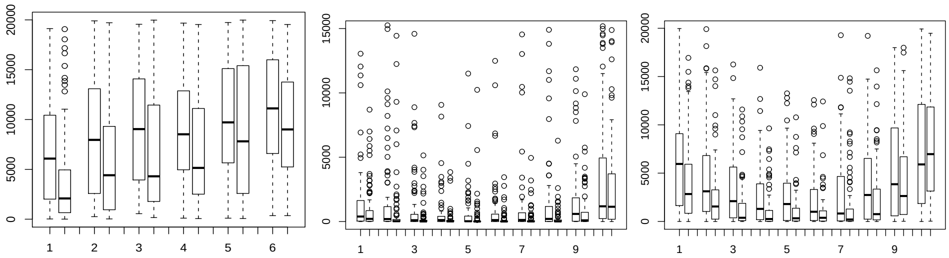

Figure 8: Boxplots of ranks for X_1 to X_6 in case 3 (first column), and for X_1 to X_{10} in cases 4(a) (second column) and 4(b) (third column) with $(\alpha, c) = (1.5, 1)$, when $(n, p) = (50, 20000)$. In each graph, the jth group of two boxplots shows the ranks for the jth component based on the $|\hat{\rho}_j|$s (left boxplot) or on the $|\hat{\omega}_j|$s (right boxplot).