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.

Abstract: Contemporary problems involving sparse, high dimensional feature selection are becoming rapidly more challenging through substantial increases in dimension. This places ever more stress on methods for analysis, since the effects of even moderately heavy tailed feature distributions become more significant as the number of features diverges. Data transformations have a significant role to play, reducing noise and enabling an increase in dimension, and for this reason they are used increasingly widely. In this paper we examine the performance of a typical transformation of this type, and study the extent to which it preserves the main attributes that lead to reliable feature selection. We show both numerically and theoretically that, in the presence of heavy tailed data, the size of the dimension for which effective variable selection is possible can be increased dramatically, from a low-degree polynomial function of sample size to one that is exponentially large.

Keywords: Correlation; feature selection; heavy tail; nonparametric statistics; Studentising; variable selection.
Short title. High dimensional data analysis.

1 Introduction

The future of genomic data analysis promises data vectors whose length, p, is not in the thousands or tens of thousands, as is common today, but in the millions or tens of millions. To appreciate the relevance of these numbers, note that there are likely between $p = 20{,}000$ and 25,000 human protein-coding genes, but, apparently, between $p = 10^6$ and 3×10^6 SNPs, or single-nucleotide polymorphisms, in the human genome. Data on SNPs, rather than on genes, are becoming common, and the availability of such high dimensional vectors means that methods have to be effective when p is much larger than the sample size, n.

In this paper we consider feature ranking in such very high dimensional problems. It has been proved in that context that if distributions are sufficiently light-tailed, standard methods work for p exponentially large as a function of n. However, genetic data are often very noisy, and can contain unusually large observations. In other words, such data are often not sufficiently light-tailed, which is a serious issue in very high dimensional settings because, as the number of features diverges, the influence of even moderately heavy tailed feature distributions becomes more significant. Therefore, techniques more appropriate to heavy-tailed distributions have to be employed.

In this setting it is increasingly common to borrow tools from the more conventional low dimensional robust literature. In particular, practitioners commonly employ simple data transformation methods. In the present paper we provide a theoretical account of the performance of such transformation methods, pointing especially to the extent of their advantages in ultra high dimensional settings. We show that they can provide substantial improvements when some or all of the components have heavy-tailed distributions. For example, we show that standard feature ranking methods, whose performance degrades when p is an exponentially large function of n, recover their capacity to deal with ultra high dimensional, noisy data when used in conjunction with an appropriate data transformation. The theoretical results in section 3 demonstrate these advantages for ranking methods based on either differences of means or on correlation. In the former setting, a standard approach consists of ranking components according to Student's t scores calculated for each component. Although Studentising helps combat heavy tailed noise, we show that transforming the variables has even more to offer. Moreover, when the distributions are light tailed and do not contain outliers, transforming the data does not have a serious negative impact on feature ranking. In section 4 we illustrate some of the main issues using two real datasets.

2 Transformation methods

2.1 Feature ranking based on correlation

Suppose we observe independent and identically distributed data pairs (X_i, Y_i), for $1 \leq i \leq n$, where $X_i = (X_{i1}, \ldots, X_{ip})$ is a p-vector, Y_i is a scalar, and we are interested in the relation between Y_i and X_i. For example, Y_i might be a score variable for a disease, and X_i a vector whose indices represent p genes or parts of a chromosome, and the interest could be in uncovering indices that are relevant to the disease, and in constructing a model for the relationship between these and Y_i. Methods for model building include the lasso (Tibshirani, 1996), non-convex penalisation (Fan and Li, 2001) and the Dantzig selector (Candes and Tao, 2007).

Fan and Lv (2008) suggest preceding model building by a massive dimension reduction step, using "correlation learning" where the p components of X_i are ranked according to the values of $|\widehat{\rho}_j|$, with, for $j = 1, \ldots, p$, $\widehat{\rho}_j = \widehat{\xi}_j/(\widehat{\sigma}_{Xj}\,\widehat{\sigma}_Y)$ denoting an empirical estimator of the theoretical correlation $\rho_j = \mathrm{corr}(X_{1j}, Y_1) = \xi_j/\{\sigma^2_{Xj}\,\sigma^2_Y\}^{1/2}$. Here we have used the notation $\xi_j = \mathrm{cov}(X_{1j}, Y_1)$, $\sigma^2_{Xj} = \mathrm{var}\,(X_{1j})$, $\sigma^2_Y = \mathrm{var}\,(Y_1)$,

$$\widehat{\xi}_j = \frac{1}{n}\sum_{i=1}^{n}(X_{ij}\,Y_i - \bar{X}_j\,\bar{Y})\,, \widehat{\sigma}^2_{Xj} = \frac{1}{n}\sum_{i=1}^{n}\left(X^2_{ij} - \bar{X}^2_j\right), \widehat{\sigma}^2_Y = \frac{1}{n}\sum_{i=1}^{n}\left(Y^2_i - \bar{Y}^2\right), \tag{2.1}$$

where $\bar{X}_j = n^{-1}\sum_i X_{ij}$ and $\bar{Y} = n^{-1}\sum_i Y_i$. Correlation learning assigns empirical rank $\widehat{\jmath}_k$ the jth component, where $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ denotes the permutation of $1, \ldots, p$ such that $|\widehat{\rho}_{\widehat{\jmath}_1}| \geq \ldots \geq |\widehat{\rho}_{\widehat{\jmath}_p}|$. For each k, $\widehat{\jmath}_k$ is an estimator of the theoretical rank j_k, where $j_1, \ldots, j_p$ is a permutation of $1, \ldots, p$ such that $|\rho_{j_1}| \geq \ldots \geq |\rho_{j_p}|$. Dimension reduction is performed by removing components with low absolute empirical correlations. See also Lv and Fan (2009) and Fan and Lv (2010).

In some analyses of genomic data, practitioners report the results of robust Wilcoxon rank tests as well as, or instead of, Student's t statistics; see Li and Fine (2010). This approach is relatively robust, and immune to heavy tail problems, although the results to which it leads are perhaps a little more difficult to interpret since they address differences between the distributions of noisy approximations to (for example) gene expression levels, rather than differences between means.

2.2 Data transformation for correlation ranking

A difficulty with the correlation learning methodology discussed in section 2.1 is that empirical correlations are sensitive to aberrations caused by heavy-tailed distributions of the explanatory vectors X_i or experimental errors ϵ_i. As a result, ranking based on correlations can be quite poor in the presence of heavy tails, since empirical correlations fluctuate so heavily that many of those corresponding to a theoretical correlation of zero take higher values than those for which the theoretical counterpart is nonzero. To reduce the fluctuations of empirical correlations, a common approach is to transform the data, replacing X_{ij} and Y_i by

$$U_{ij} = \Psi(X_{ij})\,, \quad Z_i = \Psi(Y_i)\,, \tag{2.2}$$

respectively, when computing the correlation coefficient. Here we take Ψ to be a uniformly bounded, monotone function, for example a distribution function. Then we replace ρ_j by the new correlation coefficient computed from the transformed data:

$$\omega_j = \mathrm{cov}(U_{1j}, Z_1)/\{\mathrm{var}\,(U_{1j})\,\mathrm{var}\,(Z_1)\}^{1/2}\,, \tag{2.3}$$

of which an estimator is $\widehat{\omega}_j = \widehat{\zeta}_j/(\widehat{\tau}_{Uj}\,\widehat{\tau}_Z)$ where

$$\widehat{\zeta}_j = \frac{1}{n}\sum_{i=1}^{n}\left(U_{ij}\,Z_i - \bar{U}_j\,\bar{Z}\right),\ \widehat{\tau}_{Uj}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(U_{ij}^2 - \bar{U}_j^2\right),\ \widehat{\tau}_Z^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Z_i^2 - \bar{Z}^2\right). \tag{2.4}$$

The components are ranked in the order $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ determined by $|\widehat{\omega}_{\widehat{\jmath}_1}| \geq \ldots \geq |\widehat{\omega}_{\widehat{\jmath}_p}|$.

The estimator $\widehat{\omega}_j$ is effectively the correlation estimator discussed in section 8.3 of Huber (1981). Khan *et al.* (2007) discuss a similar approach for calculating correlation in the context of the LARS method of Efron *et al.* (2004), a technique of the same type as the lasso, where one of the ingredients needed is correlation. However, they treat only numerical aspects, and only in cases where $p < n$.

Our goal here is quite different. We wish to prove, both theoretically and numerically, that in ultra high dimensional problems, where p is much larger than n, standard correlations fail to rank components correctly, whereas transformed correlations perform well. Note that it is really the conjunction of heavy-tails and high dimension that causes poor performance of correlation ranking, since in the case of heavy-tailed distributions, when p is large, with high probability many of the p empirical correlations are very far from their true values. Section 3 will shed light on the impact of heavy tails on ranking a very high number of components through empirical correlations, and prove that transforming data can greatly improve ranking.

To illustrate in practice that it is ultra high dimension that, when associated with heavy tails, causes the difficulties, we simulated data $(X_{i1}, \ldots, X_{ip}, Y_i)$, $i = 1, \ldots, n$, from the model $Y_i = \sum_{j=1}^{6} X_{ij} + \epsilon_i$, where distributions of the ϵ_is and of the X_is were, respectively, $0.98F_1 + 0.02F_2$ and $0.98F_1 + 0.02F_3$, with F_1, F_2 and F_3 denoting the distributions of, respectively, a $U[-10, 10]$, a $U[-150, -100]$, and a $U[15, 25]$. Here, the Y_is depend only on $X_{i1}, \ldots, X_{i6}$, and a small proportion of the data take unusually large values. For several values of n and p we generated 200 samples of size n from this model, and then ranked the p components $X_{i1}, \ldots, X_{ip}$ according to their absolute empirical correlations with Y_i, as described in section 2.1, and according to their transformed versions described above, where Ψ was as described in section 4. From the 200 samples

Table 1: Median and quartile ranks of X_{i1} to X_{i6} obtained by empirical correlation ranking (first block of four rows) and by transformed empirical correlation ranking (second block of four rows), when ranking $p = 20000$ components .

	X_1		X_2		X_3		X_4		X_5		X_6	
n	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3
50	712	75–3291	809	168–3600	474	45–2707	906	172–4786	672	120–3531	750	86–3638
75	236	32–1880	192	36–1394	282	30–1750	236	24–2038	190	17–1632	331	37–1951
100	91	7–806	152	15–854	110	10–836	88	9–795	145	14–776	98	9–575
200	6	3–49	7	3–56	7	2–31	10	3–59	6	3–33	6	2–56
50	174	28– 996	317	45–1394	137	14– 920	344	57–1429	262	42–1116	242	29–1570
75	48	5–202	39	5–199	40	5–244	68	7–388	25	5–175	44	5–288
100	8	3–56	15	4–76	12	3–71	8	2–37	10	3– 93	7	2–68
200	4	2–6	4	2–5	4	2–5	3	2–6	3	2–5	3	2–5

we then calculated the median (Q2) and first and third quartiles (Q1 and Q3) of ranks assigned to X_{i1} to X_{i6} (their true ranks are 1). We show the results in Table 1 for several values of n, when $p = 20000$.

As we can see, it is when n is too small that occasional outliers cause empirical correlation ranking to fail to identify $X_{i1}, \ldots, X_{i6}$ as being some of the most important components. When n is larger, the method does not rank them perfectly, but is able to rank them highly (a rank is high when close to 1). Clearly, the transformed correlation approach is able to rank highly $X_{i1}, \ldots, X_{i6}$ for smaller values of n. In this simple example the observations contain only moderate outliers, but the differences between the two methods are even more striking for higher values of p and when the distributions are more heavily tailed; see Delaigle and Hall (2011) for a more extensive simulation study. Section 3 will give a theoretical account of these empirical findings.

Note that for simplicity we focus on the simple correlation ranking method, but of course, variable transformation can be used to reduce the effect of heavy tailed distributions in variants of the correlation approach, such as for example the generalised correlation technique of Hall and Miller (2009). Conclusions similar to those drawn in this paper can be derived in such settings too.

2.3 Feature ranking based on mean differences

Transforming data prior to ranking can also be used in other contexts, such as the one based on mean differences. Suppose we observe a sample of size n of independent

p-vectors coming from two populations Π_X and Π_Y. That is, we observe $X_1, \ldots, X_{n_1}$ from Π_X, and $Y_1, \ldots, Y_{n_2}$ from Π_Y, where $n_1 + n_2 = n$, $X_i = (X_{i1}, \ldots, X_{ip})$ and $Y_i = (Y_{i1}, \ldots, Y_{ip})$. For example, X_{ij} and Y_{ij} could represent expression levels for the jth gene of the ith individual coming from population Π_X (a population of individuals suffering from a particular medical condition) and Π_Y (a population from which the condition is absent), respectively. It is often of interest in such problems to identify the components of the p-vectors for which the two populations differ the most. This is often done by detecting the components that have the largest mean differences. Let

$$E(X_{i_1 j}) = \mu_{1j} \text{ and } E(Y_{i_2 j}) = \mu_{2j}, \quad \text{for } 1 \le i_k \le n_k,\ k = 1, 2 \text{ and } 1 \le j \le p, \tag{2.5}$$

and let $j_1, \ldots, j_p$ be a permutation of $1, \ldots, p$ such that $\mu_{2j_1} - \mu_{1j_1} \ge \ldots \ge \mu_{2j_p} - \mu_{1j_p}$. When ranking is based on mean differences, the ℓth component is assigned rank j_ℓ.

In practice the means μ_{kj} are unknown and the ranks are estimated empirically from the data, for example by replacing the theoretical means by their empirical counterparts, and ranking the components according to the values of the raw differences,

$$D_j = \bar{Y}_j - \bar{X}_j, \tag{2.6}$$

where $\bar{X}_j = n_1^{-1} \sum_i X_{ij}$, $\bar{Y}_j = n_2^{-1} \sum_i Y_{ij}$.

Like the correlation approach, this ranking method is very sensitive to random fluctuations. To some extent this sensitivity can be corrected by ranking values taken by Student's t statistics T_j, corresponding to two-sample t-tests of the null hypothesis H_{0j} that $\mu_{2j} = \mu_{1j}$, against the alternative H_{1j} that $\mu_{2j} > \mu_{1j}$; that is, ranking

$$T_j = (\bar{Y}_j - \bar{X}_j) \big/ \big(n_1^{-1}\, S_{1j}^2 + n_2^{-1}\, S_{2j}^2\big)^{1/2}, \tag{2.7}$$

for $1 \le j \le p$, where S_{1j}^2 and S_{2j}^2 are conventional variance estimators computed from the data $X_{1j}, \ldots, X_{nj}$ and $Y_{1j}, \ldots, Y_{nj}$, respectively. Based on these t statistics, the components are ranked empirically by taking $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ to be the permutation of $j_1, \ldots, j_p$ such that $T_{\widehat{\jmath}_1} \ge \ldots \ge T_{\widehat{\jmath}_p}$. Adopting this technique, the most influential feature; for example, in a genomic context, the most influential gene; is the one for which H_{0j} is rejected most resoundingly in favour of H_{1j}. See also Lyons-Weiler *et al.* (2003), Xie *et al.* (2004), Papana and Ishwaran (2006) and Yang *et al.* (2009) for other motivations of the Student's t approach.

Student's t statistics T_j are less influenced than the raw differences D_j by statistical fluctuations when the data X_{ij} or Y_{ij} are relatively heavy tailed; see for example Delaigle, Hall and Jin (2011). However they can still be affected significantly, and in consequence, $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ defined above can be poor estimators of the true ranks $j_1, \ldots, j_p$. Moreover, this approach runs counter to the notion that it should emphasise location differences. To see why, note that if $j = j_{\text{small}}$ and $j = j_{\text{large}}$ denote indices for which $\sigma_j^2 = n_2^{-1} \operatorname{var}(Y_{1j}) + n_1^{-1} \operatorname{var}(X_{1j})$ is particularly small, or particularly large, respectively, then the feature with index j_{small} is likely to be ranked well ahead of that with index j_{large} even if the mean difference $\mu_{1j} - \mu_{2j}$ takes much the same value in both cases. Thus, a ranking of the T_js generally fails to adequately reflect the sizes of the differences $\mu_{1j} - \mu_{2j}$; in important cases those values are distorted by standardising for scale. This focus on differences between population means reflects the fact that, in most applications of t statistics, one generally has in mind not just the difference between two sample means, but rather the implications, conveyed by that difference, for the difference between the means of the respective populations.

Variants of t statistics have been suggested in the literature. See for example Efron *et al.* (2001), Smyth (2004), Wu (2005) and Opgen-Rhein *et al.* (2007). However, the main goal of these methods is to improve estimators of the variances σ_j^2 by borrowing strengths across similar components. In particular, they do not necessarily address the problems that concern us. For an illustration on some simulated examples, see Delaigle and Hall (2011).

The discussion above motivates methodology that focuses more sharply on the magnitudes of differences between means, yet also alleviates problems that arise when the data X_{ij}, and/or Y_{ij}, have heavy tails. Let $U_{ij} = \Psi(X_{ij})$ and $V_{ij} = \Psi(Y_{ij})$, where Ψ is a uniformly bounded, monotone increasing function. Here, the components are ranked by ranking values of empirical mean differences calculated from these transformed data. That is, with the transformation approach,

$$\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p \text{ is the permutation of } 1, \ldots, p \text{ for which } \bar{V}_{\widehat{\jmath}_1} - \bar{U}_{\widehat{\jmath}_1} \geq \ldots \geq \bar{V}_{\widehat{\jmath}_p} - \bar{U}_{\widehat{\jmath}_p}, \tag{2.8}$$

where $\bar{U}_j = n_1^{-1} \sum_i U_{ij}$, $\bar{V}_j = n_2^{-1} \sum_i V_{ij}$. To appreciate that this approach is well founded, let us elaborate on the model (2.5) by assuming that

$$X_{ij} = \mu_{1j} + \epsilon_{1ij},\ Y_{ij} = \mu_{2j} + \epsilon_{2ij} \text{ and, for any given } i \text{ and } j, \text{ the errors } \epsilon_{kij} \text{ for } k = 1, 2 \text{ are identically distributed.} \tag{2.9}$$

(Note that the quantities μ_{kj} here need not be means, and in particular, in contrast to cases where t statistics are ranked, no finite moments need be assumed.) Then,

$E\{\Psi(Y_{ij})\} \geq E\{\Psi(X_{ij})\}$, or $E\{\Psi(Y_{ij})\} \leq E\{\Psi(X_{ij})\}$, according as $\mu_{2j} \geq \mu_{1j}$, or $\mu_{2j} \leq \mu_{1j}$, respectively. These properties persist if the inequalities are taken to be strict, provided that Ψ is strictly monotone increasing on the real line. The assumption of identical distribution of errors, imposed in (2.9), is not a necessary condition for preserving the order of expectation in each case, for example when the means take only a finite number of distinct values, but it is perhaps the simplest sufficient condition.

Therefore, in the circumstances described by (2.9),

the expected value of $\bar{V}_j - \bar{U}_j$ is of the same sign as $\mu_{2j} - \mu_{1j}$, and the expected value is a monotone increasing function of μ_{2j} and a monotone decreasing function of μ_{1j}. (2.10)

Moreover, since the variables U_{ij} and V_{ij} are uniformly bounded then their associated large deviation probabilities are particularly small, even when the errors ϵ_{kij} in (2.9) are heavy-tailed. This property and (2.10) underpin the attractiveness of feature ranking based on values of $\bar{V}_j - \bar{U}_j$. Note that, since the transformation Ψ is uniformly bounded, the expected value referred to in (2.10) is always well defined even if the data are very heavy tailed.

Transforming data before calculating a mean is not new, but our goal is to investigate the advantages of such a transformation approach when used to rank a very high number of components in the presence of heavy tails. As in the correlation context, the negative impact of heavy tails for ranking means, without transforming the data, makes itself felt mostly when p is extremely large, since the probability of incorrect ranking increases with p. See section 3 for a theoretical study.

To illustrate numerically the problems due to the conjunction of high dimension and heavy tails, we generated 200 samples from $(X_{i1}, \ldots, X_{ip})$ and $(Y_{i1}, \ldots, Y_{ip})$, where, $X_{ij} - \mu_j \sim F$ and $Y_{ij} \sim F$, with F a symmetric stable distribution with characteristic function $\phi(t) = \exp(-|t|^{1.5})$. Here, only the first six components are relevant, as we took $\mu_j = 1\{j \leq 6\}$. We ranked all p components using each of the three methods described above: the method based on the raw differences at (2.6), the one based on Student's t differences at (2.7), and the transformation approach discussed above, with Ψ is as in section 4. From the 200 samples we then calculated the median, first and third quartiles of ranks assigned to X_{i1} to X_{i6} (their true ranks are 1). We show the results in Table 2 for the three methods, when $p = 20000$ and for several values of n.

As in the correlation context we see that it is when n is very small compared to p that

Table 2: Median and quartile ranks of X_{i1} to X_{i6} obtained by empirical mean ranking, when ranking $p = 20000$ components. First block of four lines: using (2.6); second block: using (2.7); third block: using $\bar{V}_j - \bar{U}_j$ at (2.8).

	X_1		X_2		X_3		X_4		X_5		X_6	
n	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3	Q2	Q1–Q3
50	1696	941–4085	1664	819–3140	1896	951–4672	1856	1042–3596	1448	793–3139	1692	908–3343
75	1442	698–3052	1378	786–2275	1382	796–3307	1344	717–2753	1318	663–2483	1492	736–3251
100	1169	694–2126	1099	676–1963	934	602–1760	1042	599–1739	1064	722–2514	1128	689–2377
200	684	446–1180	730	468–1254	694	471–1341	702	472–1386	791	506–1230	718	439–1132
50	410	50–4017	310	20–2420	627	53–4492	562	62–2595	478	39–2326	594	102–3389
75	243	7–3089	151	21–1249	380	18–2827	254	16–1890	292	14–2428	340	20–2818
100	167	9–1305	87	6–1168	81	6–892	85	5–788	108	9–1670	126	10–1670
200	11	2–515	17	3–639	11	2–280	7	2–304	25	3–615	14	3–315
50	28	5–133	36	7–157	48	6–240	39	6–165	37	4–188	35	7–240
75	6	2–40	7	3–36	6	2–27	6	2–38	6	2–41	10	2–43
100	5	2–13	4	2–9	4	2–6	4	2–6	5	2–12	4	2–8
200	3	2–5	3	2–5	4	2–5	3	2–5	4	2–5	3	2–5

the method based on (2.6) fails to identify these six components as being important, but the method does rank them increasingly highly as n increases. Ranking from Student's t differences at (2.7) already improved the results, but the transformation method gave even better results. See Delaigle and Hall (2011) for a more extensive simulation study with more complex settings, and see section 3 for a theoretical study of the empirical properties discussed here.

Of course, transforming the variables does not necessarily imply that all the components will be better ranked than by Student's t. However, the transformation approach often manages to detect components that were neglected because of heavy tailedness, and likewise can avoid false positive results caused by the same issue. Our main goal is not to rank all the components perfectly, but to identify the most important ones, that is those with highest ranks.

Remark 1. The expected value in (2.10) is not necessarily a monotone function of $\mu_{2j} - \mu_{1j}$. This need not cause difficulties, because in many settings the levels of monotonicity described in (2.10) are adequate. However, the problem can be eradicated by adjusting the methodology slightly. Specifically, define $W_{i_0 i_1 j} = \Psi(Y_{i_1 j} - X_{i_0 j})$ for $1 \leq i_0 \leq n_1$ and $1 \leq i_1 \leq n_2$, and put

$$\bar{W}_j = \frac{1}{n_1 n_2} \sum_{i_0=1}^{n_1} \sum_{i_1=1}^{n_2} W_{i_0 i_1 j} \,.$$

Then, if the errors ϵ_{kij} in (2.9) have distributions that do not depend on i, $E(\bar{W}_j)$ is a monotone increasing function of $\mu_{2j} - \mu_{1j}$ and is strictly monotone if Ψ is strictly increasing. This property motivates the following alternative method:

$$\text{define } \widehat{\jmath}_1, \ldots, \widehat{\jmath}_p \text{ to be the permutation of } 1, \ldots, p \text{ for which } \bar{W}_{\widehat{\jmath}_1} \geq \ldots \geq \bar{W}_{\widehat{\jmath}_p}. \quad (2.11)$$

We prefer the method based on (2.8), since it is considerably faster to implement in practice.

Remark 2. The transformation we use is monotone, although nonlinear, and this is a critical asset when using it for the purpose of ranking. Of course, the scale and units of the data change in a nonlinear way when transformed, and that should be borne in mind if we use the transformed data for a purpose other than ranking. This issue is not as important in the case of ranking based the correlation coefficient, which is scale and unit free and is itself a somewhat arbitrary measure of association.

3 Theoretical properties

In this section we study theoretical properties of the transformed feature ranking approaches discussed in section 2. We start with a general result in sections 3.1 and 3.2, where we prove that overall, without any restriction on the tails of the data, the empirical rankings based on variable transformation accord with their theoretical counterpart even when p is much larger than n. Next, in section 3.3 we show that this nice property is not shared by standard methods based on untransformed data. Together, these results illustrate the extent to which transforming the data effectively addresses heavy-tailedness in ultra-high dimensional feature selection. We first state, and prove, our theoretical results in the case of correlation, since this is more awkward than the context of means since correlation involves a random denominator. Subsequently, we consider the case of the mean.

3.1 Performance of the transformation method based on correlation

Suppose we observe independent and identically distributed vectors (X_i, Y_i), for $1 \leq i \leq n$, where $X_i = (X_{i1}, \ldots, X_{ip})$ is a p-vector and Y_i is a scalar. Note that we do not make any assumptions on the dependence structure of the components of the feature vectors $X_i = (X_{i1}, \ldots, X_{ip})$. Our first result, Theorem 3.1 below, shows that the variable transformation method is able to rank features in accordance with the values of the estimable but unknown correlation coefficients ω_j, defined at (2.3), even if the distributions of the X_{ij}s and/or the Y_is are heavy-tailed.

We choose the function Ψ, which we use to define our transformed variables U_{ij} and Z_i in terms of X_{ij} and Y_i at (2.2), so that following properties are satisfied:

Ψ is uniformly bounded and monotone increasing; (3.1)

var (U_{1j}) and var (Z_1) are bounded away from zero uniformly in $1 \leq j \leq p$,
as $n \to \infty$. (3.2)

For example, we can take Ψ to be the distribution function of a symmetric, unimodal, continuous distribution with unit variance and zero mean. Consider taking $\Psi = \Phi$, the standard normal distribution. For x not too large we have

$$\Phi(x) \approx \tfrac{1}{2} + (2\pi)^{-1/2}\, x \tag{3.3}$$

and for x larger than about 2 or smaller than -2 the value of $\Phi(x)$ is virtually 0 or 1, respectively. Since transformed feature ranking methods produce identical results if Ψ is replaced by $\sqrt{2\pi}\,(\Psi - \frac{1}{2})$ throughout, we see that Ψ is virtually the identity for x not too large, although with "barriers" in both tails that prevent x from taking values that are too large positive or too large negative. This explains intuitively why this approach is more resistant to problems caused by heavy tails than a method based directly on the data, or equivalently, one that takes Ψ to be linear. (Note that the approximation at (3.3) is used only as an aid to intuition, and is not employed anywhere in our analysis.)

Apart from (3.2), the only condition we assume on the distributions is the following:

The data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed; the common distribution is allowed to depend on n. (3.4)

This condition clearly does not impose any restrictive tail behaviour on the distributions. Assumption (3.2) ensures that computing the correlation at (2.3) does not involve dividing by quantities that can become arbitrarily close to zero as $n \to \infty$. Finally, we allow p to be exponentially large compared to n. Let λ_n denote a sequence diverging to infinity as $n \to \infty$, and satisfying $\lambda_n = o(n^{1/2})$. We assume that:

$p = O\{\exp(C\,\lambda_n^2)\}$ for a constant $C > 0$, where C depends on $\sup|\Psi|$ and on the lower bounds in (3.2). (3.5)

Remark 3. It is straightforward to treat cases where location and scale corrections are made empirically, using, for example, the median and standardised interquartile range, respectively. Under mild additional assumptions those quantities have the properties that, for each $\epsilon > 0$, they are within ϵ of the respective true values uniformly in $1 \le i \le p$, if p satisfies (3.5). Similar remarks apply to the results in section 3.2.

Remark 4. The normal $\mathrm{N}(0, \sigma^2)$ distribution function is attractive to use in practice, since it is symmetric and involves only one user-chooseable parameter. However, there may in some cases be advantages in using a distribution function where shape, as well as scale, can be adjusted. Examples include Student's t distribution functions, where both scale and shape, through the number of degrees of freedom, ν say, can be adjusted. The value of ν should be interpreted in the continuum.

For the statement of Theorem 3.1 below, recall that $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ are defined by $|\widehat{\omega}_{\widehat{\jmath}_p}| \ge \ldots \ge |\widehat{\omega}_{\widehat{\jmath}_p}|$. Given constants c_1 and c_2 satisfying $0 < c_1 < c_2 < \infty$, let $\mathcal{J}_1$ and $\mathcal{J}_2$ be the respective subsets of $\{1, \ldots, p\}$ for which $|\omega_j| \le c_1\, n^{-1/2}\, \lambda_n$ and $|\omega_j| \ge c_2\, n^{-1/2}\, \lambda_n$. Note particularly that (3.4) does not require the components X_{ij} of X_i to be independent or uncorrelated.

Theorem 3.1. *If* (3.1)–(3.5) *hold then, with probability converging to 1 as $n \to \infty$, all the indices from $\mathcal{J}_2$ are listed in the sequence $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ before any of the indices from $\mathcal{J}_1$.*

The implications of the theorem are perhaps most readily seen by considering a case where the indices $1, \ldots, p$ can be divided into $m + 2$ distinct classes $\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_{m+1}$, corresponding to increasingly large values of $|\omega_j|$. Here m is fixed. In this setting the theorem implies, among other properties, that for each pair (k_1, k_2) satisfying $0 \le k_1 < k_2 \le m + 1$, with probability converging to 1 as $n \to \infty$ all the indices from $\mathcal{I}_{k_2}$ are listed in the sequence $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ ahead of all those from $\mathcal{I}_{k_1}$; and moreover that, if the corresponding values of $|\omega_j|$ are at least as large as n^{-b} for some $b < \frac{1}{2}$, the sizes of these classes can be exponentially large before the claimed result breaks down. The latter issue is important, since, without imposing severe conditions on the tails of the distributions of the data, only polynomially large p is permitted if feature ranking using conventional correlation is employed. This point will be discussed in more detail in Theorems 3.3

and 3.4 of section 3.3, where we shall show, respectively, that only polynomially large p is permissible if feature ranking is based on conventional correlation, but that exponentially large p is possible if variable transformation methods are employed.

To be more specific about the classes $\mathcal{I}_0, \mathcal{I}_1, \ldots, \mathcal{I}_{m+1}$ of indices, define these sets by asking that $j \in \mathcal{I}_{m+1}$ if and only if $|\omega_j| > c$, where $c > 0$ is fixed; that $j \in \mathcal{I}_k$ (for $1 \le k \le m$) if and only if $|\omega_j| \in [c_{k1}\, n^{-b_k}, c_{k2}\, n^{-b_k}]$, where $0 < c_{k1} < c_{k2} < \infty$ and $\frac{1}{2} > b_1 > \ldots > b_m > 0$; and that $j \in \mathcal{I}_0$ if and only if $\omega_j = 0$. Theorem 3.1 implies that if $p = O\{\exp(C\, n^{1-2b_1})\}$ then, with probability converging to 1, all the indices from $\mathcal{J}_{k+1}$ are listed in the sequence $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ before any of the indices from $\mathcal{J}_k$, and that this result holds for $0 \le k \le m$. More general properties, in the case where $m = m(n)$ is permitted to diverge with n, can be derived using result (5.1) in section 5.

3.2 Performance of the transformation method based on mean differences

Analogous results hold for the feature ranking methods defined at (2.8) and (2.11). To give details of those properties we recall that in the cases represented by (2.8) and (2.11),

> the data are in the form of independent p-vectors $X_{i_1} = (X_{i_1 1}, \ldots, X_{i_1 p})$ and $Y_{i_2} = (Y_{i_2 1}, \ldots, Y_{i_2 p})$ coming from two populations, where $1 \le i_k \le n_k$ and n_k denotes the size of the sample from population k. (3.6)

Condition (3.1) on Ψ is unchanged; (3.2) is replaced by the assumption that:

> var $\{\Psi(X_{1j})\}$ and var $\{\Psi(Y_{1j})\}$ are uniformly bounded away from zero, if we are ranking features as at (2.8); var $\{\Psi(Y_{1j} - X_{1j})\}$ is bounded away from zero, if we are ranking according to (2.11); (3.7)

and (3.5) is unchanged. The following theorem describes properties of the ranking method. Its proof is almost identical to that of Theorem 3.1, hence we omit it.

Theorem 3.2. *Under* (3.1) *and* (3.5)–(3.7)*, with probability converging to 1 as* $n \to \infty$*, all the indices from* $\mathcal{J}_2$ *are listed in the sequence* $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$*, defined by* (2.11)*, before any of the indices from* $\mathcal{J}_1$*, where* $\mathcal{J}_1$ *and* $\mathcal{J}_2$ *are defined in section 3.1.*

3.3 Failure of feature ranking based on untransformed data

Next we show that feature ranking based on untransformed data is not effective when feature distributions can produce random variables taking relatively large values. For brevity, and because it is technically more difficult owing to the random denominator, we treat only the case of correlation, but similar results hold for the case of mean differences. Since our result is negative, and establishing it in generality would not confer significantly greater authority, we make simplifying assumptions that lead to a short, transparent proof. The problems that we describe here are present even more forcefully if we replace the independence and identical distribution assumption, of (3.10) below, by one where non-stationarity and correlation are present between components, since this reduces effective sample size.

In particular, we suppose we observe independent and identically distributed data vectors (X_i, Y_i), for $1 \le i \le n$, generated by a linear model,

$$Y_i = \alpha + \sum_{j=1}^{p} \beta_j \, X_{ij} + \epsilon_i \,, \tag{3.8}$$

where $\alpha, \beta_1, \ldots, \beta_p$ are scalar parameters, the experimental errors ϵ_i are identically distributed, and the ϵ_is are independent of the X_{ij}s. For simplicity we take the error ϵ_i in (3.8) to be identically zero. If feature selection methods based on conventional correlation are ineffective in this case then they will not perform well when error is present. Also for simplicity we assume that the intercept term in the model equals zero, and there are just two nonzero coefficients β_j, these being $\beta_1 = 1$ and $\beta_2 = n^{-b}$, where $0 < b < \frac{1}{2}$. Therefore the model at (3.8) reduces to:

$$Y_i = X_{i1} + n^{-b} \, X_{i2} \,. \tag{3.9}$$

We take sample size, n, to be the index of our asymptotic theory, and interpret dimension, p, as a function of n.

We assume that the design variables X_{ij} involve occasional outliers of size n^a, distributed as follows:

> First simulate independent and identically distributed random variables Q_{ij}, where the common, nondegenerate distribution is compactly supported, has zero mean and does not depend on i, j or n; and then, for each j, choose a value of i randomly (independently of the Q_{ij}s) and uniformly between 1 and n, and replace Q_{ij} by n^a. (3.10)

If we were to assume that the distribution of the X_{ij}s was regularly varying at infinity with exponent a^{-1}, for example Student's t distribution with $a-1$ degrees of freedom, then for each fixed k the k largest values of $X_{1j}, \ldots, X_{nj}$ would be of size n^a as $n \to \infty$. Since k here can be arbitrarily large, the problems experienced by feature ranking methods would be more serious than those that we shall describe in Theorem 3.3 below. On the other hand, if rather than assume that an outlier is present in each design sequence $\mathcal{X}_j$, we were to suppose that it is present with probability π_n where $\pi_n \to 0$ at rate n^{-c}, say, as $n \to \infty$, then the problems caused by outliers would be less conspicuous, but variable selection by feature ranking would still fail if p was of larger order than n^{1+c}.

Under the model described by (3.10) the variance of each X_{ij} equals $(1-n^{-1})\,\{\mathrm{var}\,(Q)+n^{2a-1}\}$, and to prevent this quantity diverging to infinity we assume, in Theorem 3.3 below, that $a < \frac{1}{2}$. On the other hand, taking a too small means that the outliers n^a have relatively little input, and so, since Theorem 3.3 seeks to identify cases where the outliers cause problems, we also place a lower bound on a in the theorem.

Recall that feature ranking by correlation orders the indices $1, \ldots, p$ as $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$, where $|\widehat{\rho}_{\widehat{\jmath}_p}| \geq \ldots \geq |\widehat{\rho}_{\widehat{\jmath}_p}|$. Now, under the linear model at (3.8), if the vector components $X_{i1}, \ldots, X_{ip}$ are independent, then we have $\rho_j = \beta_j\,(\sigma_{Xj}^2)^{1/2}/(\sigma_Y^2)^{1/2}$. Moreover, if $X_{i1}, \ldots, X_{ip}$ are also independent of the errors ϵ_i and

$$\beta_j = 0 \text{ for } j > q,\ \beta_j \neq 0 \text{ for } 1 \leq j \leq q \tag{3.11}$$

then

$$\omega_j = 0 \text{ if } j > q, \text{ and is nonzero for } 1 \leq j \leq q. \tag{3.12}$$

A proof is given in section 5.4. Hence, the actual correlations are $\rho_1 = (1+n^{-2b})^{-1/2}$, $\rho_2 = n^{-b}\,(1+n^{-2b})^{-1/2}$ and $\rho_j = 0$, for $j \geq 3$, and an ideal ranking of features in terms of the estimated correlations $\widehat{\rho}_j$ should at least preserve the order of the first two components among the ρ_js. That is, it should have $\widehat{\jmath}_1 = 1$ and $\widehat{\jmath}_2 = 2$ with probability converging to 1 as $n \to \infty$. Theorem 3.3, below, shows that if p is of sufficiently larger order than n then this property fails in respect of the second component, although it holds for the first.

Theorem 3.3. *Assume model* (3.9) *for the response variables* Y_i*, and model* (3.10) *for the design sequence* X_i*, where* $\frac{1}{4} < a < \frac{1}{2}$ *and* $1-2a < b < 2\,(1-2a)$ *in* (3.9) *and* (3.10)*; and that, as* $n \to \infty$*,* $p/n \to \infty$ *and* $p = O\{\exp(C\,n^{4a-1})\}$*, for a constant* $C > 0$

depending on the distribution of Q in (3.10)*. Then $P(\widehat{\jmath}_1 = 1) \to 1$ and $P(\widehat{\jmath}_2 = 2) \to 0$ as $n \to \infty$.*

This proves that standard correlation learning is rather ineffective in the case of heavy-tailed distributions. In contrast, we have already proved in Theorem 3.1 that correlation learning based on variable transformation performs particularly well. To simplify comparison with the negative result of Theorem 3.3, in the next theorem we give a more detailed, positive, result for transformation-based correlation learning under the simple model treated in Theorem 3.3. In particular, we show that $P(\widehat{\jmath}_1 = 1)$ and $P(\widehat{\jmath}_2 = 2)$ both converge to 1 as $n \to \infty$, when we define $\widehat{\jmath}_1, \ldots, \widehat{\jmath}_p$ by $|\widehat{\omega}_{\widehat{\jmath}_p}| \geq \ldots \geq |\widehat{\omega}_{\widehat{\jmath}_p}|$, where $\widehat{\omega}_j = \widehat{\zeta}_j / (\widehat{\tau}_{Xj}\, \widehat{\tau}_Y)$ and the quantities on the right-hand side of this equation are given by (2.4).

Theorem 3.4. *Assume the conditions of Theorem 3.3, except that the upper bound on p there is replaced by $p = O\{\exp(C\, n^{1-2b})\}$, where again $C > 0$ depends on the distribution of Q. Suppose too that the function Ψ used to transform the data is bounded and strictly monotone increasing, and has two bounded derivatives. Then both $P(\widehat{\jmath}_1 = 1)$ and $P(\widehat{\jmath}_2 = 2)$ converge to 1 as $n \to \infty$.*

It is readily seen that Theorems 3.3 and 3.4 have a non-null intersection. That is, in the context of the models described by (3.9) and (3.10), the transformation approach leads to correct rankings in cases where relatively conventional methods produce incorrect results.

4 Real-data illustrations

The simulation results reported in section 2 illustrate the theoretical properties of section 3 on simple examples, but simulations in more complex settings can be found in Delaigle and Hall (2011). In this section we show for some real-data examples that our theoretical results in section 3 translate into significant improved ranking by using the transformation approach. To apply the transformation procedure we need to choose the function Ψ. In section 3.1 we gave intuitive arguments suggesting that Ψ could be taken to be the standard normal distribution. Of course, for these intuitive arguments to be valid we need to rescale the data in some way before applying the transformation Ψ, as we need the non-aberrant data values to be located roughly between -2 and 2. Our theory is also valid when we do this; see remark 3 in section 3.1. We do not rescale the data in the same way for the methods based correlation and mean differences, so below we give details for these approaches separately.

For the correlation-based method we take $U_{ij} = \Psi\{(X_{ij} - m_j)/s_j\}$ and $Z_i = \Psi\{(Y_i - m)/s\}$, where m_j and s_j are, respectively, the sample median and standardised interquartile range of X_{ij}, and m and s are defined in the same way for Y_i. Note that, since correlation is invariant under changes of scale and location, and heavy-tailedness properties are also unaffected by such changes, the theoretical results in section 3.1 are unchanged if we replace X_{ij} and Y_i there by, respectively, $(X_{ij} - a_j)/b_j$ and $(Y_i - a)/b$, where a, b, a_j and b_j are bounded nonzero constants. See remark 3 in section 3.1 for the case of empirical location and scale.

For the mean-based method founded on (2.8) in section 2.3 we take $U_{ij} = \Psi\{(X_{ij} - m_j)/s_j\}$ and $V_{ij} = \Psi\{(Y_{ij} - m_j)/s_j\}$, where m_j is the minimum of the sample medians of X_{ij} and Y_{ij}, and s_j is the standardised interquartile range of the pooled sample of X_{ij} and Y_{ij} (with j fixed). Note that we cannot center the components X_{ij} and Y_{ij} to their respective medians, since our goal is precisely to detect differences in location. In this case, too, rescaling the data can be accommodated by our theory; see remark 3 in section 3.1.

4.1 Cardiomyopathy microarray data

Our first application concerns the ranking of genes according to their influence on overexpression of a G-protein-coupled receptor (Ro1) in mice. See Segal *et al.* (2003) and Hall and Miller (2009) for a more detailed description of the data. Here, Y_i is the measurement of Ro1 on the ith mouse, and $(X_{i1}, \ldots, X_{ip})$ are the expression levels of p genes for the ith mouse. The sample size was $n = 30$, and $p = 6319$ genes were observed. We ranked the genes using both the untransformed correlations $\widehat{\rho}_j$ and the transformed correlations $\widehat{\omega}_j$. Out of the 50 genes ranked highest by the transformed correlation method, 40 were not ranked in the top 50 by the untransformed correlation approach. In Figure 1 we show scatterplots of the (X_{ij}, Y_i)s and of the (U_{ij}, Z_i)s for the first 16 of these 40 genes. Above each scatterplot we indicate the rank assigned by the untransformed correlation method for each gene in the left panel, and the rank assigned when we used the transformation based approach in the right panel. In each scatterplot we also show

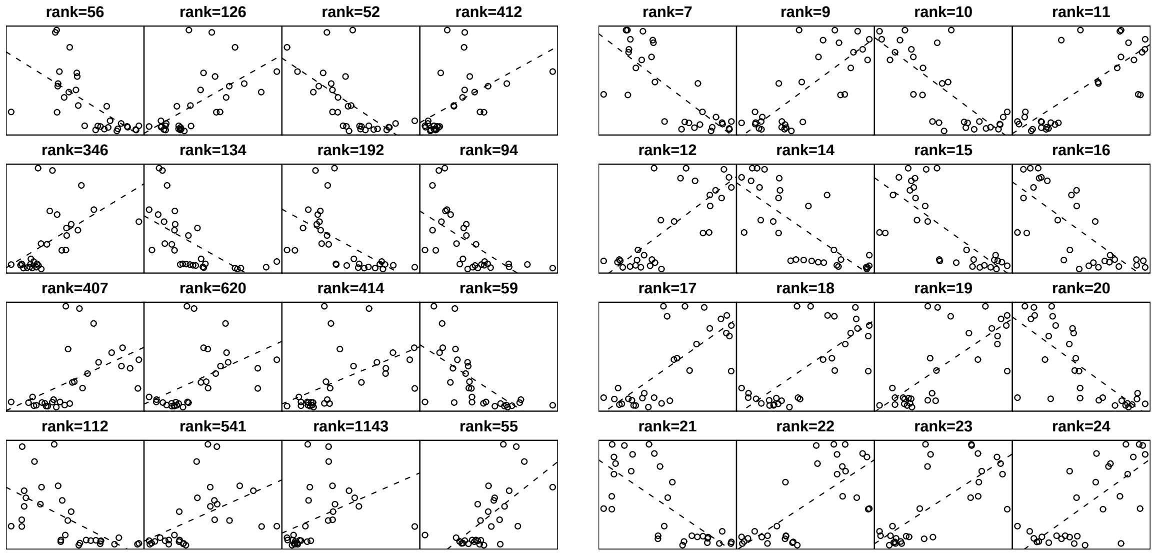

Figure 1: Ro131 study: scatterplots of the first 16 genes selected by the variable transformation approach, which were not ranked among the top 50 by the untransformed correlation method. Left group of 16: scatterplots of the (X_{ij}, Y_i)s for those 16 genes; right group of 16: scatterplots of the (U_{ij}, Z_i)s, where $U_{ij} = \Psi(X_{ij})$ and $Z_i = \Psi(Y_i)$. The oblique lines are the least squares lines.

the least squares regression line.

For a number of the 16 genes the least squares lines, and thus the empirical correlations, for the untransformed data appear to be too highly influenced by outliers. As a result, some of the genes were ranked by the untransformed correlation approach much lower than they arguably should have been. This is particularly striking for the genes ranked 346, 192, 407, 620 and 1143 by untransformed correlation, which were ranked 12, 15, 17, 18 and 23 after transforming the data. For other genes, such as the genes originally ranked 55 and 126, transforming the variables did not seem particularly useful. As we noted in section 2.3, transforming the data does not necessarily improve the rank of each gene, but it manages to identify influential genes that were wrongly disqualified because of outliers.

4.2 Affymetrix spike-in data

We used the transformed mean ranking approach to the Affymetrix spike-in data described by Cope *et al.* (2004). This dataset contains two groups with $n = 12$ observations in each group, and $p = 12626$ genes. It is available from `http://strimmerlab.`

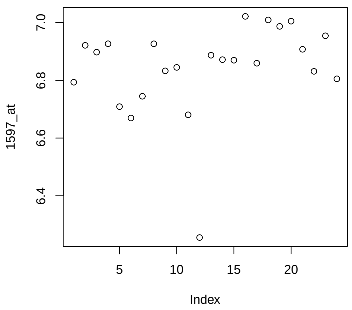

Figure 2: Scatterplot of gene 1597_at for the Affymetrix data.

`org/data.html`. In Cope *et al.* (2004), a comparison with benchmark data permitted the identification of 16 genes believed to be differentially expressed (that is, for which the mean differences between the two groups were significantly different from zero). We applied the method in section 2.3 to these data, and found that we needed to keep the 47 highest values of $|\bar{V}_j - \bar{U}_j|$ in order to select these 16 known genes. By comparison, the method based on Student's t scores needed to keep 155 components, and the method based on the untransformed means $|\bar{Y}_j - \bar{X}_j|$ had to keep 100. We also tried the shrinkage method of Opgen-Rhein *et al.* (2007), and had to keep 54 components in order to include all 16 identified genes. The method of Efron *et al.* (2001) needed 61 components, and the other methods needed even more. The better performance of the transformation based procedure comes from the fact that many components have one or more outliers. For example, the scatterplot of the observations of gene 1597_at in Figure 2 shows that the data on gene 1597_at (which, of the 16 genes, was the most difficult to pick up for most methods) contain an outlier, and this is precisely the setting in which our method can bring considerable improvements. Note that our method did not rank all the 16 known genes higher than the other methods, but managed to pick them all more efficiently, by keeping fewer components.

5 Technical arguments

5.1 Proof of Theorem 3.1

It follows from Bernstein's inequality that, if $W, W_1, \ldots, W_n$ are independent random variables for which $P(|W| \le C_1) = 1$, $E(W) = 0$ and $E(W^2) = 1$, and if we define $S = n^{-1/2} \sum_i W_i$, then for each $C_2 > 0$ there exists $C_3 > 0$, depending only on C_1 and C_2, such that $P(|S| > x) \le 2\,\exp(-C_3\,x^2)$ for all $x \in (0, C_2\,n^{1/2}]$. Applying this result to $W = \{W' - E(W')\}/(\mathrm{var}\, W')^{1/2}$, where W' equals U_{ij}, U_{ij}^2, Z_i, Z_i^2 or $U_{ij}\, Z_i$, it can be deduced that

$$P\big(n^{1/2}\, |\widehat{\omega}_j - \omega_j| > x\big) \leq 2\, \exp\big(- C_4\, x^2\big)\,,$$

for all $x \in (0, C_2\, n^{1/2}]$, where $C_4 > 0$ depends only on $\sup |\Psi|$ and C_2. Therefore, if $p = O\{\exp(C_5\, \lambda_n^2)\}$ for $C_5 > 0$ sufficiently small, then, for all $C_6 > 0$,

$$P\Big(\max_{1 \leq j \leq p} |\widehat{\omega}_j - \omega_j| > n^{-1/2}\, C_6\, \lambda_n\Big) \to 0\,. \tag{5.1}$$

Result (5.1) implies Theorem 3.1.

5.2 Proof of Theorem 3.3

Under the model at (3.9) the correlation estimator $\widehat{\rho}_j$, defined below (2.1), is given by

$$\widehat{\rho}_j = \big(\widehat{a}_{j1} + n^{-b}\, \widehat{a}_{j2}\big)\, \big(\widehat{b}_j\, \widehat{b}_Y\big)^{-1} \tag{5.2}$$

for $1 \leq j \leq p$, where $\widehat{b}_Y = \{n^{-1} \sum_i (Y_i - \bar{Y})^2\}^{1/2}$,

$$\widehat{a}_{jk} = \frac{1}{n} \sum_{i=1}^{n} X_{ij}\, \widehat{w}_{ik}\,, \quad \widehat{b}_j = \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \right\}^{1/2}, \quad \widehat{w}_{ik} = X_{ik} - \bar{X}_k\,.$$

Let $\mathcal{E}_{jk}$ denote the event that the values of n^a are in the same position in the sequences $\mathcal{X}_j$ and $\mathcal{X}_k$, and write $\widetilde{\mathcal{E}}_{jk}$ for the complement of $\mathcal{E}_{jk}$. For each $C_1 > 1$ there exists $C_2 > 0$ such that for all $j \neq k$ and all $d \in (0, \frac{1}{2})$,

$$P\big(\big|n^{1/2}\, \widehat{a}_{jk}\big| > C_1\, n^d \,\big|\, \widetilde{\mathcal{E}}_{jk}\big) = O\big\{ \exp\big(- C_2\, n^{2d}\big)\big\}\,, \tag{5.3}$$

while if $\frac{1}{4} < a < \frac{1}{2}$ then for each $C_1 > 0$ there exists $C_2 > 0$ such that, for $j \neq k$ and all $d \in (0, 2a - \frac{1}{2}]$,

$$P\big(\big|n^{1/2}\, \widehat{a}_{jk} - n^{2a-(1/2)}\big| > C_1\, n^d \,\big|\, \mathcal{E}_{jk}\big) = O\big\{ \exp\big(- C_2\, n^{2d}\big)\big\}\,. \tag{5.4}$$

(If $j \neq k$ then the left-hand sides of (5.3) and (5.4) do not depend on j and k. Each of (5.3)–(5.5) is proved using Bernstein's inequality.)

Note too that $\widehat{a}_{jj} = \widehat{b}_j^2$, and that we may assume without loss of generality that the common distribution of the variables X_{ij} that do not equal n^a has unit variance. Then, if $C_1 > 1$, there exists $C_2 > 0$ such that, for all j and $d = 2a - \frac{1}{2}$,

$$P\big(n^{1/2}\, \big|\widehat{b}_j - 1\big| > C_1\, n^d\big) = O\big\{ \exp\big(- C_2\, n^{2d}\big)\big\}\,. \tag{5.5}$$

(The left-hand side of (5.5) does not depend on j.) Furthermore,

$$P(\mathcal{E}_{jk}) = n^{-1}\,. \tag{5.6}$$

Note too that, by (5.2),

$$\widehat{b}_Y\,\widehat{\rho}_1 = \widehat{b}_1 + n^{-b}\,\widehat{a}_{12}\,\widehat{b}_1^{-1}\,,\ \widehat{b}_Y\,\widehat{\rho}_2 = \widehat{a}_{12}\,\widehat{b}_2^{-1} + n^{-b}\,\widehat{b}_2\,,\ \widehat{b}_Y\,\widehat{\rho}_j = \big(\widehat{a}_{j1} + n^{-b}\,\widehat{a}_{j2}\big)\,\widehat{b}_j^{-1}\,, \tag{5.7}$$

where the last equation will be used for $j \geq 3$.

Using (5.3), (5.4) and (5.5), each with $d = 2a - \frac{1}{2}$, it can be deduced that $\widehat{a}_{jk} = O_p(n^{-(1/2)+2a-(1/2)}) = O_p(n^{-(1-2a)})$ and $\widehat{b}_j = 1 + O_p(n^{-(1-2a)})$, uniformly in $1 \leq j \leq p$ and in $k = 1, 2$. (The choice $d = 2a - \frac{1}{2}$ is responsible for the exponent $2d = 4a - 1$ in the formula $p = O\{\exp(C\,n^{4a-1})\}$ in the statement of Theorem 4.1.) Hence, using (5.7), we obtain $\widehat{b}_Y\,\widehat{\rho}_1 = 1 + O_p(n^{-(1-2a)})$, $\widehat{b}_Y\,\widehat{\rho}_2 = O_p(n^{-(1-2a)} + n^{-b})$ and $\widehat{b}_Y\,\widehat{\rho}_j = O_p(n^{-(1-2a)})$ uniformly in $3 \leq j \leq p$. Therefore the following property holds:

$$P\big[|\widehat{\rho}_1| > \max\{|\widehat{\rho}_2|, \ldots, |\widehat{\rho}_p|\}\big] \to 1\,. \tag{5.8}$$

The same arguments show that

$$\widehat{b}_Y\,\widehat{\rho}_j = \widehat{a}_{j1} + o_p\big(n^{-b}\big) \quad \text{uniformly in } 3 \leq j \leq p\,, \tag{5.9}$$

and that

$$\begin{aligned}\widehat{b}_Y\,\widehat{\rho}_2 &= \widehat{a}_{12}\,\widehat{b}_2^{-1} + n^{-b}\,\widehat{b}_2 = n^{-(1-2a)}\,I(\mathcal{E}_{12}) + O_p\big(n^{-2(1-2a)}\big) + n^{-b} + O_p\big(n^{-b-(1-2a)}\big)\\ &= n^{-(1-2a)}\,I(\mathcal{E}_{12}) + n^{-b}\,\{1 + o_p(1)\} = n^{-b}\,\{1 + o_p(1)\}\,,\end{aligned}$$

where the second last identity follows from the fact that, by assumption, $b < 2(1 - 2a)$, and the last identity since $P(\mathcal{E}_{12}) = n^{-1} \to 0$. Therefore the following property holds:

$$\widehat{b}_Y\,\widehat{\rho}_2 = n^{-b}\,\{1 + o_p(1)\}\,. \tag{5.10}$$

Results (5.3) and (5.4) imply that, for a constant $C_3 > 0$,

$$P\big(\big|n^{1/2}\,\widehat{a}_{jk} - n^{2a-(1/2)}\,I(\mathcal{E}_{jk})\big| > C_1\big) = O\big\{\exp\big(-C_3\,n^{2d}\big)\big\}\,.$$

This result, and the fact that $n^{-1/2} = o(n^{2a-1})$ (by assumption, $a > \frac{1}{4}$), imply that if $d = 2a - \frac{1}{2}$ and

$$p = o\big\{\exp\big(C\,n^{2d}\big)\big\}\,, \tag{5.11}$$

where the constant C here and in the statement of the theorem satisfies $C < C_3$, then

$$\widehat{a}_{j1} = n^{2a-1}\, I(\mathcal{E}_{j1}) + o_p\big(n^{2a-1}\big) \quad \text{uniformly in } 3 \le j \le p\,. \tag{5.12}$$

Conditional on the dataset $\mathcal{X}_1 = \{X_{11}, \ldots, X_{n1}\}$, the events $\mathcal{E}_{j1}$, for $3 \le j \le p$, are independent and satisfy $P(\mathcal{E}_{j1} \,|\, \mathcal{X}_1) = n^{-1}$. Therefore if $p/n \to \infty$ then $P(\mathcal{E}_{31} \cup \ldots \cup \mathcal{E}_{p1}) \to 1$. Combining this property with (5.12) we deduce that if (5.11) holds then for each $\epsilon > 0$,

$$P\Big\{\widehat{a}_{j1} > (1-\epsilon)\, n^{2a-1} \quad \text{for at least one } j \text{ in the range } 3 \le j \le p\Big\} \to 1\,. \tag{5.13}$$

Property (5.8) implies that $P(\widehat{\jmath}_1 = 1) \to 1$. Properties (5.9) and (5.13), and the inequality $1 - 2a < b$ assumed in Theorem 3.3, imply that for each $\epsilon > 0$,

$$P\Big\{\widehat{b}_Y\, \widehat{\rho}_j > (1-\epsilon)\, n^{2a-1} \quad \text{for at least one } j \text{ in the range } 3 \le j \le p\Big\} \to 1\,,$$

which, together with (5.10) and the property $1 - 2a < b$, implies that $P(\widehat{\jmath}_2 = 2) \to 0$.

5.3 Proof of Theorem 3.4

Put $U_{ij} = \Psi(X_{ij})$ and $Z_i = \Psi(Y_i)$ where Y_i is as at (3.9), and define

$$\tilde{a}_j = \frac{1}{n} \sum_{i=1}^{n} Z_i\, \widetilde{w}_{ij}\,, \quad \tilde{b}_j = \bigg\{\frac{1}{n} \sum_{i=1}^{n} (U_{ij} - \bar{U}_j)^2\bigg\}^{1/2}, \quad \widetilde{w}_{ij} = U_{ij} - \bar{U}_j\,, \tag{5.14}$$

$\tilde{b}_Z = \{n^{-1} \sum_i (Z_i - \bar{Z})^2\}^{1/2}$ and $\widehat{\omega}_j = \tilde{a}_j/(\tilde{b}_j\, \tilde{b}_Z)$. Without loss of generality, the common distribution of the variables $\Psi(X_{ij})$ when X_{ij} does not equal n^a has unit variance. (This property is achievable by simple rescaling of Ψ, since Ψ is strictly monotone and, by (3.10), X_{ij} given that $X_{ij} \ne n^a$ has nonzero variance.) Result (5.15) below replaces both (5.3) and (5.4) in the present setting, and (5.16) replaces (5.5). Both are derived using Bernstein's inequality, treating separately the instance where $X_{ij} = n^a$ (for each j, this is true for exactly one i) and the contrary case. For each $C_1 > 1$ there exists $C_2 > 0$ such that, for all $d \in (0, \frac{1}{2})$,

$$\sup_{1 \le j \le p} P\big\{\big|n^{1/2}\, (1-E)\, \tilde{a}_j\big| > C_1\, n^d\big\} = O\big\{\exp\big(-C_2\, n^{2d}\big)\big\}\,, \tag{5.15}$$

$$\sup_{1 \le j \le p} P\big(n^{1/2}\, \big|\tilde{b}_j - 1\big| > C_1\, n^d\big) = O\big\{\exp\big(-C_2\, n^{2d}\big)\big\}\,. \tag{5.16}$$

For a given value of i let $\mathcal{E}_j$ denote the event that j is an index for which $X_{ij} \neq n^a$. Then if $\mathcal{E}_1 \cap \mathcal{E}_2$ holds,

$$\Big|\Psi(Y_i) - \big\{\Psi(X_{i1}) + n^{-b}\, X_{i2}\, \Psi'(X_{i1})\big\}\Big| \leq C\, n^{-2b}\,, \tag{5.17}$$

where C denotes a constant. (Here we have used the assumption that Ψ is twice differentiable.) Therefore,

$$\begin{aligned} &\operatorname{cov}\{\Psi(Y_i), \Psi(X_{ij}) \,|\, \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_j\} \\ &= n^{-b} \operatorname{cov}\Big\{X_{i2}, \Psi(X_{ij}) \;\Big|\; \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_j\Big\}\, E\Big\{\Psi'(X_{i1}) \;\Big|\; \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_j\Big\} + O\big(n^{-2b}\big)\,, \end{aligned} \tag{5.18}$$

uniformly in i and $j \geq 2$.

If $j = 2$ then, since Ψ is strictly monotone increasing, $\operatorname{cov}\{X_{i2}, \Psi(X_{ij}) \,|\, \mathcal{E}_1 \cap \mathcal{E}_2\} > 0$ and $E\{\Psi'(X_{i1}) \,|\, \mathcal{E}_1 \cap \mathcal{E}_j\} > 0$, and so

$$\gamma \equiv \operatorname{cov}\{X_{i2}, \Psi(X_{i2}) \,|\, \mathcal{E}_1 \cap \mathcal{E}_2\}\, E\{\Psi'(X_{i1}) \,|\, \mathcal{E}_1\} > 0\,.$$

It follows from (5.17) that

$$\operatorname{cov}\{\Psi(Y_i), U_{i1} \,|\, \mathcal{E}_1 \cap \mathcal{E}_2\} = \operatorname{var}(U_{i1} \,|\, \mathcal{E}_1) + O\big(n^{-b}\big)\,. \tag{5.19}$$

In view of (5.18),

$$\operatorname{cov}\{\Psi(Y_i), U_{i2} \,|\, \mathcal{E}_1 \cap \mathcal{E}_2\} = n^{-b}\,\gamma + O\big(n^{-2b}\big)\,, \tag{5.20}$$

and, more simply,

$$\operatorname{cov}\{\Psi(Y_i), U_{ij}\} = 0 \quad \text{for} \quad 3 \leq j \leq p\,. \tag{5.21}$$

Results (5.15) and (5.19), and the fact that (by assumption) $\operatorname{var}\{\Psi(X_{ij})\} = 1$ when $X_{ij} \neq n^a$ (i.e. $X_{ij} = Q_{ij}$, as in (3.10)), imply that for each $C_1 > 0$ there exists $C_2 > 0$ such that

$$P\Big\{|\tilde{a}_1 - 1| > C_1\left(n^{d-(1/2)} + n^{-b}\right)\Big\} = O\big\{\exp\big(-C_2\, n^{2d}\big)\big\} \tag{5.22}$$

whenever $d \in (0, \frac{1}{2})$.

By (5.15), (5.20) and (5.21), for each $d \in (0, \frac{1}{2})$ and each $C_1 > 0$ we can choose $C_2 > 0$ such that

$$P\Big\{\big|\tilde{a}_2 - n^{-b}\,\gamma\big| > C_1\left(n^{d-(1/2)} + n^{-b}\right)\Big\} = O\big\{\exp\big(-C_2\, n^{2d}\big)\big\}\,, \tag{5.23}$$

$$\sup_{3 \leq j \leq p} P\Big\{|\tilde{a}_j| > C_1\left(n^{d-(1/2)} + n^{-b}\right)\Big\} = O\big\{\exp\big(-C_2\, n^{2d}\big)\big\}\,. \tag{5.24}$$

In progressing from (5.19)–(5.21) to (5.22)–(5.24) we used the fact that, for each j, there is just one i in the sequence $i = 1, \ldots, n$ for which $X_{ij} = n^a$, and that, by the definition of $\tilde{a}_j$ at (5.14), and since $P(|Z_i \, \widetilde{w}_{ij}| \leq 2 \sup \Psi^2) = 1$, this i (which is randomly chosen without regard for the values of X_{ij} that do not equal n^a) contributes no more than $2 \, n^{-1} \sup \Psi^2$ to the value of $\tilde{a}_j$. That contribution is of strictly smaller order than n^{-b} in (5.22) and (5.24).

Combining (5.16) and (5.22)–(5.24) we deduce that, taking $d = \frac{1}{2} - b$ and assuming that $p = O\{\exp(C \, n^{1-2b})\}$ for $C > 0$ sufficiently small, we have, as $n \to \infty$, $\tilde{b}_Z \, \widehat{\omega}_1 = 1 + o_p(1)$, $\tilde{b}_Z \, \widehat{\omega}_2 = n^{-b} \gamma + o_p(n^{-b})$ where $\gamma > 0$, and $\tilde{b}_Z \, \widehat{\omega}_j = o_p(n^{-b})$ uniformly in $3 \leq j \leq p$. These three properties imply that $P(\widehat{\jmath}_1 = 1)$ and $P(\widehat{\jmath}_2 = 2)$ converge to 1 as $n \to \infty$, as claimed in the theorem.

5.4 Proof that (3.8) and (3.11) imply (3.12)

Note that under (3.8) and (3.11) and the independence assumption above (3.11), we have, trivially, $\rho_j = 0$ if $j > q$, and is nonzero for $1 \leq j \leq q$. This result will be useful in the proofs. To prove (3.12), it suffices to show that if Ψ is uniformly bounded then the covariance between $\Psi(Y_i)$ and $\Psi(X_{ij})$ vanishes when $\beta_j = 0$, and if in addition Ψ is strictly monotone then the covariance is nonzero when $\beta_j \neq 0$. The first of these results is trivial since, under (3.8) and the independence assumption, $\rho_l = 0$, $\Psi(Y_i)$ and $\Psi(X_{ij})$ are independent if $\beta_j = 0$. To derive the second result it suffices to show that if random variables V_1 and V_2 are independent, if V_1 is essentially bounded and nondegenerate, and if functions Ψ_1 and Ψ_2 are strictly monotone increasing and bounded, then $\gamma \equiv \text{cov}[\Psi_1\{\Psi_2(V_1) + V_2\}, V_1] > 0$. The function $\Psi_3 = \Psi_2(\,\cdot + EV_1)$ is monotone if Ψ_2 is, and in this notation $\gamma = \text{cov}[\Psi_1\{\Psi_3(V_1 - EV_1) + V_2\}, V_1 - EV_1]$, so without loss of generality $E(V_1) = 0$, in which case, since V_1 and V_2 are independent, $\gamma = E[\Psi_1\{\Psi_2(V_1) + V_2\} \, V_1] = E(E[\Psi_1\{\Psi_2(V_1) + V_2\} \, V_1 \,|\, V_2]) = E(\text{cov}[\Psi_1\{\Psi_2(V_1) + V_2\}, V_1 \,|\, V_2]) = E\{\Psi_4(V_2)\}$, where $\Psi_4(v) = E(\text{cov}[\Psi_1\{\Psi_2(V_1) + V_2\}, V_1 \,|\, V_2 = v])$. That is, $\gamma = E\{\Psi_4(V_2)\}$. Therefore it suffices to consider the case where V_2 is identically constant. In this case we can absorb V_2 into the definition of Ψ_1, and so it is sufficient to take $V_2 = 0$. Therefore we must show that when Ψ_1 and Ψ_2 are strictly monotone increasing and Ψ_1 is bounded, $\text{cov}[\Psi_1\{\Psi_2(V_1)\}, V_1] > 0$, or equivalently, if Ψ_3 is strictly monotone increasing and bounded (and V_1 is essentially bounded and not degenerate) then $\gamma_1 \equiv \text{cov}\{\Psi_3(V_1), V_1\} > 0$. To appreciate that this inequality holds, let V be a random variable distributed as V_1 and independent of V_1. Note that $\{\Psi_3(V_1) - \Psi_3(V)\}\,(V_1 - V) \geq 0$ and is strictly positive whenever $V_1 \neq V$. Therefore, $E[\{\Psi_3(V_1) - \Psi_3(V)\}\,(V_1 - V)] > 0$. Since $E(V_1) = E(V) = 0$ then the left-hand side here equals $2\,\text{cov}\{\Psi_3(V_1), V_1\}$.

Acknowledgements

Research supported by grants and fellowships from the Australian Research Council.

References

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313–2351.

Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z. and Speed, T. P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **20**, 323–331.

Delaigle, A. and Hall, P. (2011). Effect of heavy-tails on ultra high dimensional variable ranking methods: simulation results. *Available at* `http://www.ms.unimelb.edu.au/~aurored/Simulations/AddSimul.pdf`.

Delaigle, A., Hall, P. and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *J. Roy. Statist. Soc.* Ser. B **73**, 283-301.

Efron, B., Tibshirani, R., J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* **96**, 1151–1160.

Efron, B. E., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *Ann. Statist.* **32**, 407–451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. (With discussion.) *J. Roy. Statist. Soc.* Ser. B **70**, 849–911.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.* **18**, 533–550.

Huber, P .J. (1981). *Robust statistics.* Wiley series in probability and statistics, Wiley, New-York.

Khan, J.A., Van Aelst, S., and Zamar, R.H. (2007). Robust Linear Model Selection Based on Least Angle Regression. *J. Amer. Statist. Assoc.* **102**, 1289–1299.

Li, J. and Fine, J. P. (2010). Weighted area under the receiver operating characteristic curve and its application to gene selection. *J. Roy. Statist. Soc.* Ser. C **59**, 673–692.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, **37**, 3498–3528.

Lyons-Weiler, J., Patel, S. and Bhattacharya, S. (2003). A classification-based machine learning approach for the analysis of genome-wide expression data. *Genome Res.* **13**, 503–512.

Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage Approach. *Statist. Appl. Genet. Mol. Biol.* **6**, 9.

Papana, A. and Ishwaran, H. (2006). CART variance stabilization and regularization for high-throughput genomic data. *Bioinformatics* **22**, 2254–2261.

Segal, M. R., Dahlquist, K. D., Conklin, B. R. (2003). Regression approach for microarray data analysis. *J. Computational Biol.* **10**, 961–980.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.* **3**, 3.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc.* Ser. B **58**, 267–288.

Wu, B. (2005). Differential gene expression detection using penalized linear regression models: the improved SAM statistic. *Bioinformatics* **21**, 1565–1571.

Xie, Y., Jeong, K.S., Pan, W., Khodursky, A. and Carlini, B.P. (2004). A case study on choosing normalization methods and test statistics for two-channel microarray data. *Comp. Funct. Genomics* **5**, 432–444.

Yang, X., Zhou, Y., Jin, R. and Chan, C. (2009). Reconstruct modular phenotype-specific gene networks by knowledge-driven matrix factorization. *Bioinformatics* **25**, 2236–2243.