LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Anne-Laure Boulesteix, Vincent Guillemot, Willi Sauerbrei

Use of pre-transformation to cope with outlying values in important candidate genes

Technical Report Number 083, 2010
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Use of pre-transformation to cope with outlying values in important candidate genes

Anne-Laure Boulesteix[1*] Vincent Guillemot[1,2] Willi Sauerbrei[3]

[1] Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninistr. 15, 81377 Munich, Germany

[2] Département SSE, Ecole Supélec, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

[3] Institute for Medical Biometry and Medical Informatics, Universitätsklinikum Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany

Abstract

Outlying values in predictors often strongly affect the results of statistical analyses in high-dimensional settings. Although they frequently occur with most high-throughput techniques, the problem is often ignored in the literature.

We suggest to use a very simple transformation, proposed before in a different context by Royston and Sauerbrei, as an intermediary step between array normalization and high-level statistical analysis. This straightforward univariate transformation identifies extreme values and reduces the influence of outlying values considerably in all further steps of statistical analysis without eliminating the incriminated observation or feature. The use of the transformation and its effects are demonstrated for diverse univariate and multivariate statistical analyses using nine publicly available microarray data sets.

R-codes for reproducing the whole analysis are available from the companion website: http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/outliers/, such that the study is completely reproducible.

*Corresponding author. Email: boulesteix@ibe.med.uni-muenchen.de.

1 Introduction

Preprocessing and normalization of microarray experiments, gene ranking and prediction based on high-dimensional data have been the subject of thousands of articles in the last decade. Outliers have comparatively not focused much attention, although they may considerably affect the results of high-level analyses. In the microarray literature, the term "outlier" most often refers to outlying arrays. In the context of class prediction, mislabeled arrays are a special type of outliers. Such mislabeled arrays can be seen as outliers with respect to their class and can easily be detected in the context of prediction, since they are usually consistently misclassified by standard classification algorithms. Independently of the investigated class prediction problem, some arrays may yield atypical gene expression measurements, hence the term "abnormal sample" used in this context by (Shieh and Hung, 2009). Methods for detecting outlying arrays are suggested by Baty et al. (2008) and Shieh and Hung (2009).

The problem of outlying values in the predictors is a different one, and probably more difficult to handle than outlying samples in high-dimensional settings. In the context of differential gene expression, a few methods have been proposed that allow to identify genes with outlying values. For instance, Tibshirani and Hastie (2007) suggest the "outlier sum" as a criterion to identify genes with outlying values in the two-groups setting, e.g. normal versus cancer. The focus of this procedure is on the identification of genes that have outlying values in one of the two groups (the cancer group in their example). The criterion is essentially asymmetric, but the examination of outliers on both sides is possible by interchanging the two groups. However, the approach critically depends on a suitable threshold. The current proposal to use permutations and estimate false discovery rates for several thresholds is computer intensive and needs further development and investigations. In addition, it does not aim at transforming the data for subsequent analyses. In a related approach by Tomlins et al. (2005a) termed "cancer profile outlier analysis" (COPA), the *median* of each gene is centered to zero and the median absolute deviation is scaled to one. For each gene, Tomlins et al. (2005b) suggest to consider the rth percentile of the disease group's transformed values (with $r = 75\%, 90\%$ or 95%) for identifying genes that are over-expressed in a subgroup of the disease group. Genes with outlying values can be seen as extreme cases of such genes, with subgroups of size, say, one, two or three. As a competing approach, Wu (2007) proposes an "outlier robust t-statistic" based on medians with the aim to efficiently identify outliers or subgroups with outlying values. In Wu's paper, genes with outlying values or genes with differentially expressed subgroups are addressed simultaneously. However, both problems are different in general. While subgroups are often assumed to be "biologically interesting", single outlying values may rather be the consequence of undesired events such as measurement error, technical errors in the lab, or the particular characteristics of a single patient that are not relevant to the disease. In contrast to the above methods based on test statistics, Haury et al. (2010) suggest to consider as outliers the values outside a given interval defined in terms of standard deviation.

The standard traditional way to cope with such outliers in multivariate analyses is to detect them

and eliminate them from the further analyses, as already recommended by Edgeworth (1887) in the context of least squares regression: *"The method of least squares is seen to be our best course when we have thrown overboard a certain portion of our data – a sort of sacrifice which has often to be made by those who sail upon the stormy seas of Probability"*. The problem is that, in the context of high-dimensional data analysis, it is most often impossible to "throw overboard" all the observations that have an outlying value for at least one feature. Doing that, we would eliminate too many if not all observations.

While the methods reviewed above allow to detect outlying values in individual features, it is often unclear how we should cope with them in further analyses like feature ranking, multivariate model building or estimation of correlation graphs. This is an important problem, since such analyses are well-known to be extremely unstable against small changes in the available data set, see e.g. Boulesteix and Slawski (2009) for the case of ranked gene lists from univariate analyses.

Robust statistical procedures form the second main family of methods handling outlying values, as summarized by Rousseeuw and Leroy (2003) in low-dimensional settings. Many statistical methods can be adapted to achieve robustness against outlying values. However general agreement and clear guidance are still missing. The field is still in its infancy as far as high-dimensional data are concerned. A further problem is that robust methods have then to be used at all stages of the analysis, i.e. we would have to use, say, a robust t-statistic for differential gene expression analysis, a robust prediction method for multivariate model building, a robust method for the estimation of correlation graphs, etc. While it may be easy to compute a robust t-statistic, methods addressing the other issues are far less developed.

In this paper we suggest to use a simple transformation, proposed before in a different context (Royston and Sauerbrei, 2007), as an intermediary step between array normalization and high-level statistical analysis. This transformation considerably reduces the influence of outlying values in all steps of statistical analysis without eliminating the incriminated observation or feature. It can be seen both as a procedure to detect such outlying values and to transform the data for further analyses.

This manuscript is organized as follows. Section 2 presents the transformation and the method for outlier detection and gives a brief introduction into the standard methods used to rank genes and build predictive models. Based on nine real-life microarray data sets, we show in Section 3 that the rank of a few features may be substantially affected by the transformation and find that these features have strong outliers. We further examine the impact of outliers and the role of the proposed transformation on multivariate model selection using two well-known strategies to derive sparse prediction models. In a nutshell, we show that outlying values may make us overlook interesting features or conversely lead to the selection of uninteresting features ("false positives").

2 Methods

Let Y denote the random variable corresponding to the outcome of interest, for instance a binary class membership ($Y = 0, 1$) or a time-to-event with corresponding censoring indicator. The covariates are denoted as $X_1, \ldots, X_p$, where p is typically large. We consider a sample of n independent realizations of $(Y, X_1, \ldots, X_p)$, which are denoted as $(y_i, \mathbf{x}_i')$ with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ and $i = 1, \ldots, n$. For example, $X_1, \ldots, X_p$ may be gene expression levels, but our methodology can be directly applied to other continuous covariates like, e.g., protein expression levels. Furthermore, for binary Y, n_0 and n_1 denote the number of observations with $Y = 0$ and $Y = 1$, respectively.

2.1 Proposed preliminary data transformation

In a completely different context, namely improving robustness of fractional polynomial modelling in low dimension settings, Royston and Sauerbrei (2007) suggested a simple transformation of the covariates. The transformed version of x_{ij} (for $j = 1, \ldots, p$) is given as

$$x_{ij}^* \;=\; \left[\ln\left(\frac{\phi(z_{ij}) + \epsilon}{1 - \phi(z_{ij}) + \epsilon}\right) + \varepsilon^*\right] / (2\varepsilon^*), \tag{1}$$

where

- ϕ stands for the standard normal cumulative distribution function,
- z_{ij} is simply the standardized form of x_{ij}, i.e.

  $$z_{ij} = (x_{ij} - \overline{x}_j)/s_j,$$

  with $\overline{x}_j$ and s_j denoting the sample average and the unbiased standard deviation of covariate X_j, respectively,
- ε is a parameter and $\varepsilon^* = \ln[(1 + \varepsilon)/\varepsilon]$. Royston and Sauerbrei (2007) recommend the choice $\epsilon = 0.01$. This value is used here.

As can be seen from Eq. (1), the transformation is univariate in the sense that it is performed for each covariate separately. It is also monotonic (hence rank-invariant) and unsupervised (i.e. it does not refer to the outcome variable Y). The transformation x_{ij}^* is constructed to be nearly linear over the bulk of the observations (more precisely, within about $\overline{x}_j \pm 2.8 s_j$). It smoothly tapers to zero as $x \to -\infty$ and to one as $x \to \infty$. As an example, the transformation is depicted in Figure 2.1 for simulated data randomly generated from the normal distribution with mean 7 and standard deviation 1.

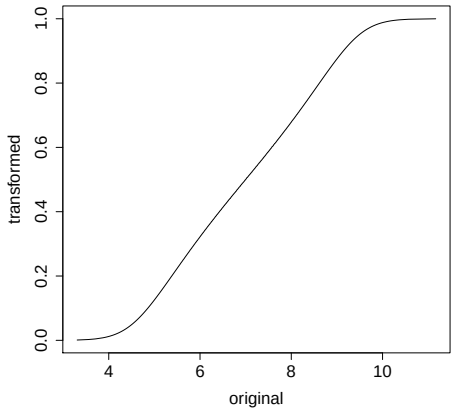

Figure 1: Transformed vs. original data for for a simulated covariate randomly generated from the normal distribution with $\mu = 7$ and $\sigma = 1$.

2.2 Gene ranking

Many microarray studies begin with a univariate gene ranking. For example, the features (genes) are typically ordered according to the absolute value of the two-sample t-statistic or one of its variants in the case of a binary Y. Readers are referred to Opgen-Rhein and Strimmer (2007); Boulesteix and Slawski (2009) for recent overviews on appropriate statistics for measuring differential expression. When the outcome of interest is a time-to-event with censoring, it is usual to rank the features based on the p-value of the likelihood-ratio (LR) test in a univariate Cox regression with the considered feature as a predictor. In the present article, we stick to the p-value of the standard two-sample t-test for binary Y and to the LR p-value in univariate Cox regression for time-to-event outcomes.

Moreover, in the case of a binary Y we also consider a moderated two-sample t-statistic as suggested in Opgen-Rhein and Strimmer (2007). The modified t-statistic is obtained by replacing the unbiased variance estimator in the denominator through a biased but more stable shrinkage estimator defined as the weighted sum of the sample variance and the median of the sample over the p covariates. The weight parameter is chosen to minimize the mean squared error of the resulting estimator. The obtained "shrinkage t-statistic" is expected to be more robust against outlying values, similar to related approaches such as Limma (Smyth, 2004).

2.3 Our rank discrepancy measure

We suggest to rank the covariates $X_1, \ldots, X_p$ based on the untransformed covariates $(x_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$ and based on the transformed data $(x^*_{ij})_{\substack{i=1,\ldots,n \\ j=1,\ldots,p}}$. Correspondingly, we obtain two rankings $(r_j)_{j=1,\ldots,p}$ and $(r^*_j)_{j=1,\ldots,p}$, where r_j and r^*_j are the ranks of covariate X_j based on the untransformed and transformed data, respectively. For instance, if the p-values from the two-sample test are used as a

ranking criterion, r_j is defined as

$$r_j \ = \ \sum_{l=1}^{p} I(p_l < p_j), \tag{2}$$

where p_j denotes the p-value of covariate j and I stands for the indicator function. We propose to compare the ranks $(r_j)_{j=1,\ldots,p}$ and $(r_j^*)_{j=1,\ldots,p}$, for instance using a scatter-plot. Genes with severe differences in ranks due to outlying values can be identified by computing

$$\Delta r_j = \frac{(r_j^* - r_j)}{\min(r_j, r_j^*)} \tag{3}$$

for each feature. With the above definition, a large absolute value of Δr is obtained both for features that rank much better with the transformed data (i.e. potentially interesting features that would have been overlooked because of the outlying value) and features that rank much better with the untransformed data (i.e. features that "seem" interesting due to an outlying value but are probably not so interesting). This criterion is inspired from the Bland-Altman plot proposed to assess agreement between two continuous measurements that would represent $r_j^* - r_j$ against $(r_j^* + r_j)/2$ (Bland and Altman, 1986). Here we adopt it and replace the denominator by $\min(r_j, r_j^*)$, as we are mainly interested in features from the left side of the distributions of r_j and r_j^* with large difference in ranks.

2.4 Multivariate models

In this paper, we also investigate the impact of the transformation on model selection in multivariate analyses. The considered multivariate methods are briefly introduced here.

2.4.1 L_1-penalized regression

L_1-penalized regression, also denoted as lasso, has grown to one of the major prediction methods in high-dimensional settings including but not limited to survival analysis (Benner et al., 2010). The lasso is attractive as a regularization method because it simultaneously performs variable selection and shrinkage: the resulting models are usually (very) sparse, depending on the applied penalty parameter. In this paper, we use the implementation provided in the package **penalized** by Goeman (2010). Since complexity selection is not our primary focus, we successively consider different values of the L_1-penalty λ_1 corresponding to signatures of usual size ($\lambda_1 = 1, 5, 10, 15$, after standardization).

2.4.2 Boosting regression

Boosting regression is an alternative model selection approach which performs well in high-dimensional data analysis, see Bühlmann and Hothorn (2007) for a recent overview and the R package **mboost** (Hothorn and Bühlmann, 2006) for a user-friendly implementation. In this paper, we use the standard boosting variant with componentwise linear least squares as a base learner and the default value $\nu = 0.1$ for the shrinkage factor. See the extensive review in Bühlmann and Hothorn (2007) and

GSE	outcome	n	$n_0 : n_1$	Events	p	Disease	Reference
9960	control/sepsis	70	16:54	-	54675	sepsis	Tang et al. (2009)
6613	control/Parkinson	105	55:50	-	22283	parkinson	Scherzer et al. (2008)
	control/tumor	47	22:25	-	22283	colon cancer	Ancona et al. (2006)
	control/tumor	102	50:52	-	12625	prostate cancer	Singh et al. (2002)
2434	relapse/no relapse	286	179:107	-	22283	breast cancer	Wang et al. (2005)
1456	overall survival	159	-	29	22283	breast cancer	Pawitan et al. (2005)
	relapse-free survival	159		40			
2603	lung metastasis-free survival	82	-	14	22283	breast cancer	Minn et al. (2005)
3494	overall survival	236	-	55	22283	breast cancer	Miller et al. (2005)
2990	metastasis-free survival	179	-	40	22283	breast cancer	Sotiriou et al. (2006)

Table 1: Summary of the data sets' characteristics.

the synthetic presentation of this boosting variant in Boulesteix and Hothorn (2010) for more details. The negative binomial log-likelihood is used as a loss function for binary Y variables, while the loss function is weighted by the inverse probability of being censored (Hothorn and Bühlmann, 2006) for time-to-event regression. Since complexity selection is not our primary focus, we successively consider several arbitrary usual values of the number m_{stop} of boosting steps ($m_{\text{stop}} = 20, 50, 100, 200$).

3 Results

3.1 Data and preprocessing

We consider a collection of nine Affymetrix gene expression data sets: five breast cancer data sets and four other diseases. In four of the data sets, the focus is on a time-to-event outcome, while the variable of interest is binary in the five other data sets (either diseased/healthy or relapse/no relapse). The raw data are publicly available as CEL files for all data sets, thus allowing to perform normalization using different methods – an important criterion for our study. We present the results obtained with all investigated data sets, not just the most favorable to our method. An overview of the considered data sets is given in Table 1.

All data sets are normalized using the RMA method as implemented in the standard function 'justRMA' from the Bioconductor package **affy** (Irizarry et al., 2009). The aim of our paper is not to compare different normalization procedures. However, we also perform additional analyses using other normalization approaches to make sure that the problem of outlying values identified here is not specific to RMA. The alternative normalization methods considered here are MAS5 (function 'mas5' from **affy** package) and GCRMA (function 'justGCRMA' from **affy** package). Some results with these alternative normalization approaches are available in additional files.

The R-codes as well as the pre-processed data sets used in this study are available for the purpose of reproducibility from the companion website (see address in the abstract) as recommended by Baggerly and Coombes (2009).

3.2 Comparison of univariate rankings before and after transformation

We compute Δr as defined in Eq. (3) for each feature in each considered data set and with each considered criterion. Moreover, having a potential influence on multivariate models in mind, it makes sense to concentrate on "relevant" features, i.e. features ranked in the top-list based on either transformed or untransformed data. Here, we consider top-lists of length 100, which is a usual choice in practice – i.e. relevant features are those with $r_j \leq 100$ or/and $r_j^* \leq 100$. The numbers of relevant features with $\Delta r > 1, 2, 5$ and $\Delta r < -1, -2, -5$ are given in Table 2 for all RMA normalized data sets. The same analyses performed on MAS5 and GCRMA normalized data sets result in different numbers, indicating that normalization affects outliers (data not shown). But the figures are in average similar to those obtained with RMA normalized data sets. In this paper we display only the results obtained with RMA because it is the most widely used normalization technique in practice (not because the results look more "favorable" with RMA).

For studies with a binary outcome results are given in the top of Table 2 for the standard t-statistic. The numbers of features with large absolute values of Δr are almost the same with the t-statistic and with the shrinkage t-statistic (data not shown), indicating that the problem of outliers cannot be simply solved by using a regularized statistic. Note that the transformed covariates have to be rescaled to their original variance before computing the shrinkage t-statistic, otherwise no shrinkage occurs. It can be seen from the bottom part of Table 2 that survival data sets have much larger numbers of outliers.

For illustrative purpose, the scatterplots of r_j and r_j^* are depicted for a data set with many outlying values (Sotiriou) and a data set without outlying values (Ancona) in Figure 2. The points are all near the diagonal for the Ancona data set, indicating good agreement between the ranks before and after transformation. In contrast, a substantial part of the relevant genes from the Sotiriou data set have severe differences in ranks before and after transformation.

Δr	$< (-5)$	$< (-2)$	$< (-1)$	> 1	> 2	> 5
Tang (/106)	0	0	0	0	0	0
Scherzer (/104)	1	1	2	0	0	0
Ancona (/103)	0	0	0	0	0	0
Singh (/106)	0	0	1	0	0	0
Wang (/107)	0	0	3	0	0	0
Pawitan (/109)	1	3	5	1	0	0
" " (/108)	1	3	12	4	2	0
Minn (/105)	0	0	0	0	0	0
Miller (/118)	2	8	13	10	5	2
Sotiriou (/114)	6	8	14	9	5	3

Table 2: **Top:** Number of relevant features (i.e. features with $r_j \leq 100$ or/and $r_j^* \leq 100$) with $\Delta r > 1, 2, 5$ and $\Delta r < -1, -2, -5$ based on the two-sample t-statistic for binary Y. **Bottom:** Number of relevant features (i.e. features with $r_j \leq 100$ and/or $r_j^* \leq 100$) with $\Delta r > 1, 2, 5$ and $\Delta r < -1, -2, -5$ based on the LR p-value in univariate Cox regression for censored Y. The total number of relevant features (with $r_j < 100$ or/and $r_j^* < 100$) is given in parentheses.

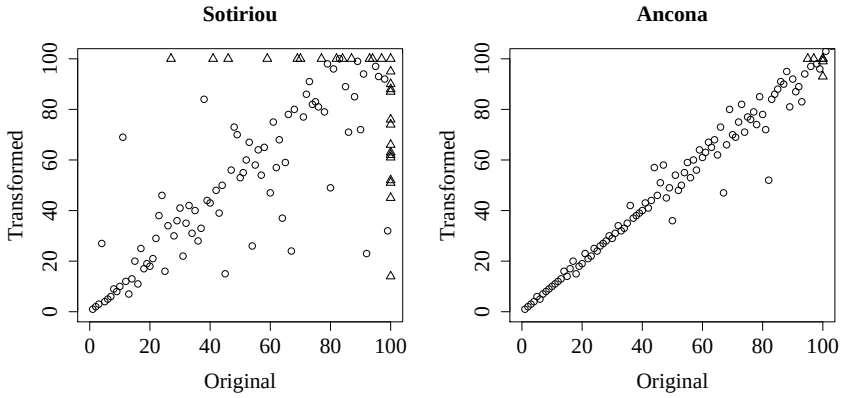

Figure 2: Scatterplots of rankings for the relevant features obtained from original data sets and transformed data sets with the Sotiriou data (left) and Ancona data (right). Features with $r_j > 100$ or $r_j^* > 100$ are marked as triangles.

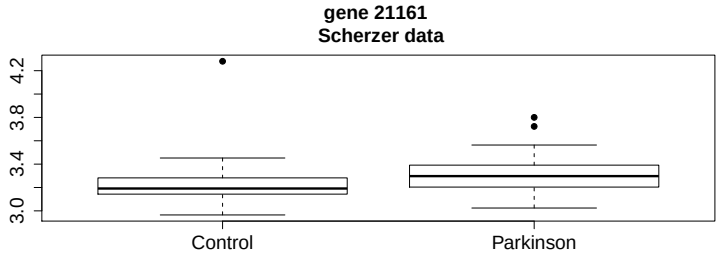

Figure 3: Boxplot of feature 21161 in control and Parkinson patients (Scherzer data).

Obviously, there are both features with negative and positive Δr. In other words, there are both features ranking noticeably better before and after transformation. Hence, after transformation, potentially important features appear at the top of list, although without transformation they may have been overlooked. Conversely, some features look important only due to outliers and the transformation allows to identify them as false positives.

To better understand the effect of the transformation and to show its influence, let us consider graphical representations of the features with large $|\Delta r|$. For example, Figure 3 displays the boxplots of feature 21161 with $\Delta r = -10.13$ found in the Scherzer data set. This feature has rank $r^*_{21161} = 84$ based on transformed data but only $r_{21161} = 935$ based on the original data. As can be seen from Figure 3, the distributions of cases and controls are different but an extreme outlying value in the control group is responsible for the bad rank obtained with untransformed data. Similarly, Figure 4 presents the boxplots of the relevant features with $\Delta r < (-5)$ for the concerned survival data sets Pawitan, Sotiriou and Miller. Obviously, these features contain strong outliers. The ranks of the corresponding features before and after transformation are displayed in Table 3.

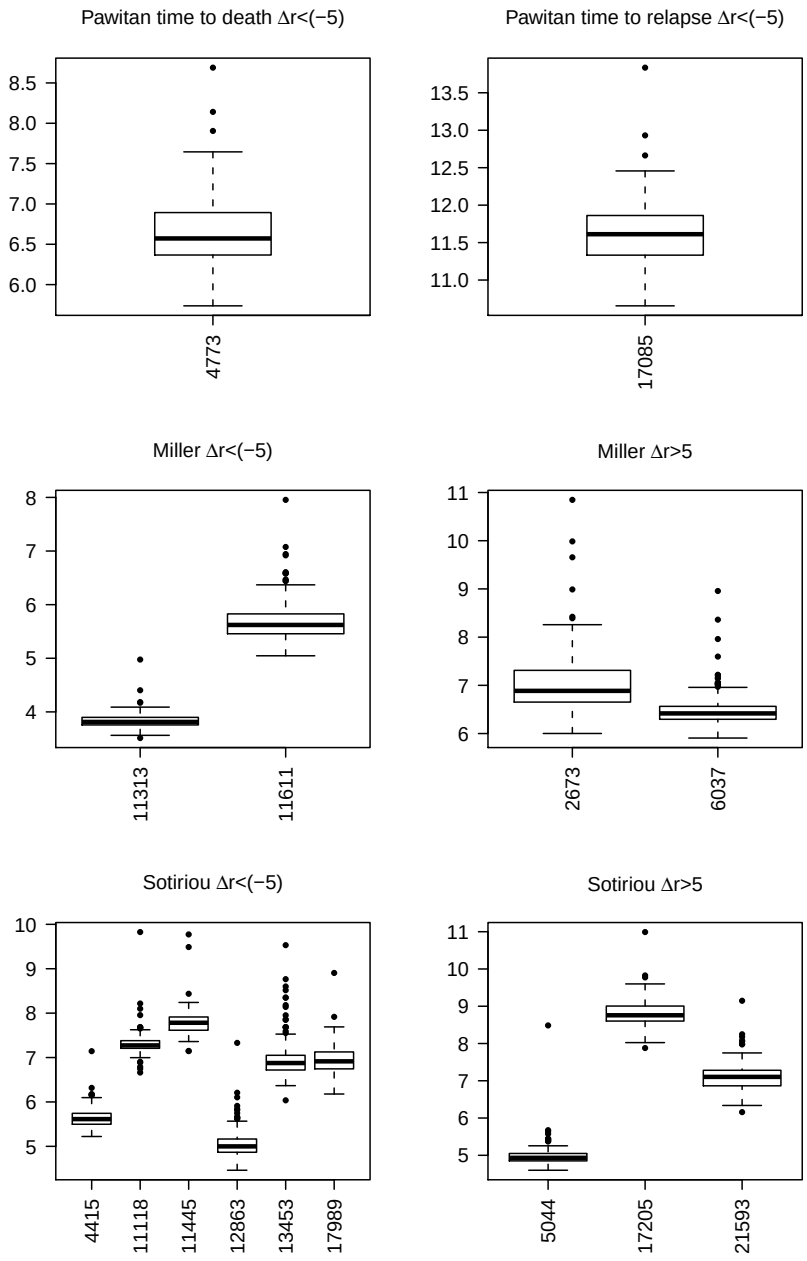

Figure 4: Boxplots of the relevant feature with the $|\Delta r| > 5$ in the survival data sets. Note that the genes do not have the same scale. However, outliers can clearly be identified.

	Δr	feature	r_j	r_j^*
Scherzer	$< (-5)$	21161	935	84
Pawitan time to death	$< (-5)$	4773	290	44
Pawitan time to relapse	$< (-5)$	17085	490	79
Miller	$< (-5)$	11313	1463	75
		11611	151	10
Miller	> 5	2673	5	193
		6037	26	316
Sotiriou	$< (-5)$	4415	120	14
		11118	281	45
		11445	507	62
		12863	403	51
		13453	373	61
		17989	385	63
Sotiriou	> 5	5044	27	344
		17205	11	69
		21593	4	27

Table 3: Ranks before transformation (r_j) and after transformation (r_j^*) for the features with $\min(r_j, r_j^*) \leq 100$ and $|\Delta r| > 5$, see Table 2.

3.3 Multivariate prediction models

For boosting regression (top) and lasso (bottom), Table 4 gives the number $|\mathcal{X} \setminus \mathcal{Z}|$ of regression coefficients that are non-zero with untransformed data but zero with transformed data, the number $|\mathcal{Z} \setminus \mathcal{X}|$ defined in the same way, and the number $|\mathcal{X} \cap \mathcal{Z}|$ of regression coefficients that are non-zero with both transformed and non-transformed data. The "signatures" selected with untransformed and transformed data show quite large overlap, especially for small numbers of boosting steps. However, a few features are selected based on transformed data but not based on untransformed data, or vice-versa. As can be seen from Table 4 the proportion of overlap decreases noticeably with the size of the signatures, yielding only moderate overlaps in the most complex models for some data sets. On the whole, multivariate analyses seem more sensitive to outliers than univariate analyses.

4 Discussion

While robust methods are specific to a particular problem and have to be developed for each research question anew, our preliminary transformation can be used as a single preliminary step to all kinds of statistical analyses. It identifies outliers and proposes a simple way to handle them. If no extreme value is identified - as in the univariate approaches in some of the data sets considered - it confirms that outliers do not seriously influence the univariate ranking of the relevant features. Such a finding summarizes the assessment for several thousands of features with different types of distributions. Obviously, this is important to know and increases confidence in the further analyses. In some of

	$\|\mathcal{X}\backslash\mathcal{Z}\|$	$\|\mathcal{Z}\backslash\mathcal{X}\|$	$\|\mathcal{X}\cap\mathcal{Z}\|$	$\|\mathcal{X}\backslash\mathcal{Z}\|$	$\|\mathcal{Z}\backslash\mathcal{X}\|$	$\|\mathcal{X}\cap\mathcal{Z}\|$	$\|\mathcal{X}\backslash\mathcal{Z}\|$	$\|\mathcal{Z}\backslash\mathcal{X}\|$	$\|\mathcal{X}\cap\mathcal{Z}\|$	$\|\mathcal{X}\backslash\mathcal{Z}\|$	$\|\mathcal{Z}\backslash\mathcal{X}\|$	$\|\mathcal{X}\cap\mathcal{Z}\|$
	mstop=20			mstop=50			mstop=100			mstop=200		
Tang	1	1	13	8	3	20	9	7	29	11	13	41
Scherzer	2	3	15	5	5	26	10	6	41	12	10	53
Ancona	0	0	13	5	4	19	12	9	24	13	16	27
Singh	2	2	6	5	4	16	3	6	22	5	5	27
Wang	3	4	12	3	4	34	11	11	51	21	21	74
Pawitan	2	2	2	3	0	8	4	2	14	5	7	23
	0	1	6	1	3	10	3	2	18	9	3	29
Sotiriou	3	0	6	3	2	12	3	1	20	8	8	31
Minn	0	1	3	1	1	8	1	0	12	4	5	15
Miller	0	1	5	3	2	10	8	5	18	19	11	28
	$\lambda_1 = 15$			$\lambda_1 = 10$			$\lambda_1 = 5$			$\lambda_1 = 1$		
Tang	0	0	3	0	1	10	3	2	21	6	11	25
Scherzer	1	1	10	2	3	27	9	7	46	16	16	59
Ancona	0	0	2	0	0	6	0	2	14	9	11	22
Singh	0	1	5	1	1	12	0	3	21	8	3	32
Wang	12	11	61	17	16	83	37	35	103	52	57	113
Pawitan	3	5	14	8	14	26	20	21	38	32	27	48
	5	4	24	15	12	42	27	33	53	34	42	61
Sotiriou	3	6	23	15	20	41	27	35	59	41	45	58
Minn	0	0	6	2	1	11	5	7	17	13	17	21
Miller	20	17	44	28	25	65	42	51	83	49	58	91

Table 4: The table gives the number $|\mathcal{X} \setminus \mathcal{Z}|$ of regression coefficients that are non-zero with untransformed data but zero with transformed data, the number $|\mathcal{Z} \setminus \mathcal{X}|$ defined the other way around, and the size $|\mathcal{X} \cup \mathcal{Z}|$ of the union. **Top:** Multivariate model selection with mboost and different numbers of boosting steps ($m_{\text{mstop}} = 20$, $m_{\text{stop}} = 50$, $m_{\text{mstop}} = 100$, $m_{\text{stop}} = 200$). **Bottom:** Multivariate model selection with lasso and different penalty parameters ($\lambda_1 = 1, 5, 10, 15$).

the data sets of our study (mainly with a survival outcome) outliers often induce the selection of features that would not have been selected otherwise or conversely lead us to overlook important features. Ignoring outlying values in statistical analyses could be seen as a simple and common error (Baggerly and Coombes, 2009). In the context of multivariate prediction differences between results with and without the transformation are larger. As for the handling of outlying values in general, the specific implications of differences in results have to be discussed with subject-matter knowledge in mind. As the transformation is easy to apply, we propose to use it at least as a sensitivity analysis. This is in line with the proposal in the original paper where the transformation was developed to improve robustness of fractional polynomial functions (Royston and Sauerbrei 2007).

Another usual way to cope with outlying values in low dimensional settings consists to simply eliminate from the data set the observations with outlying values. In high-dimensional settings, this approach may be acceptable for univariate ranking analyses. However, it can often not be applied in multivariate analyses, because many observations may have outlying values in at least one of the many features. Excluding these observations would lead to a further reduction of small sample sizes and may introduce biases. Eliminating the covariates with outlying values would also make poor sense, since they may be relevant to the investigated problem. Although we restrict to univariate rankings and prediction models in the Results section of this paper, the transformation is potentially useful

for all types of high-level analyses. For example, outliers may also affect the results of correlation networks approaches or clustering.

Of course, one could argue that outlying values may have biological relevance and should not be switched off. However, they first have to be identified anyway. The transformation with the corresponding rank discrepancy measure can be used for this purpose. While it could be interesting to investigate why the considered patient has an outlying value for the considered feature, we claim that statistical analyses should definitely not be strongly affected by a single (or a few) outlying values. In this spirit, the proposed transformation makes results more robust without completely eliminating the incriminated observation or feature. Note that other transformations may show similar effects. We consider the transformation of Eq. (1) because it is very simple, appropriate for all types of statistical analyses, and already investigated in a previous article from the related research field of medical statistics.

Furthermore, there is a potential connection between our transformation and resampling approaches. Among other advantages, resampling-based analyses may be used to detect outliers (Sauerbrei and Schumacher, 1992), although this aspect of resampling procedures has not often been addressed in the literature on high-dimensional data. In a resampling analysis, random subsamples or bootstrap samples are drawn from the original data sets and the whole analysis (e.g., univariate ranking or model building) is repeated for each subsample/bootstrap sample successively. We carried out preliminary analyses that suggest that features with very large Δr or very small Δr also have more variable ranks over the resampling iterations in resampling analyses. Roughly, this can be explained as follows. For example, if bootstrap resampling is performed, a particular outlying value is excluded from the data set in about 36.8% of the bootstrap iterations. If Δr is very large for the considered variable, i.e. if the outlying value has a strong impact on the rank, the ranks obtained when the outlying value is excluded strongly differs from the ranks obtained in the other iterations. Moreover, the effect of an outlier is amplified if it is included in the bootstrap sample more than once, thus yielding a high variability over the bootstrap iterations. If the problem of outliers is the main motivation to apply a resampling procedure, our simple transformation may be a valuable and computationally cheap alternative. However, using the bootstrap in high-dimensional data raises additional issues such as the choice between resampling with or without replacement (Binder and Schumacher, 2008). More research is needed to examine the connections between both approaches before a general recommendation can be formulated.

5 Conclusion

We suggest to apply the very simple and computationally cheap transformation previously proposed by Royston and Sauerbrei (2007) in the context of high-dimensional gene expression data. Together with the rank discrepancy measure it allows to detect features with outlying values and considerably reduces their influence in subsequent statistical analyses of all types such as univariate rankings or

multivariate model selection. By applying this transformation to nine publicly available Affymetrix microarray data sets, we show that the results of both univariate and multivariate analyses are noticeably affected by the presence of single atypical values. Based on transformed data, features that would have been overlooked otherwise show up in the top-list. Conversely, some features "look relevant" in the original data set due to a single outlying value, but disappear from the top-list if the strong effect of the outlying value is switched off via the transformation.

The particularity of our transformation is its simplicity – with several positive consequences. Firstly, it is able to handle new test data collected later. One simply has to apply Eq. (1) to the new test observations with the mean and standard deviation estimates from the training set used to do the transformation. In contrast, most pre-processing methods for microarray data such as RMA cannot simply be applied to new test observations. In this sense, the simplicity of our transformation is an important advantage. Secondly, our transformation is not specific to a particular type of data. While an improved normalization variant could perhaps correctly address outliers in a special case for a special data type, our transformation is general enough to be directly applied to all types of (metric) predictors. Thirdly, it is also a secure choice in the sense that it never produces unexpected results: it does not involve, say, complicated fitting procedures with potential convergence or instable estimation steps. Last but not least, it can be easily implemented in any statistical software tool in a few minutes.

Funding:
This project was partially supported by the French-Bavarian cooperation center for universities, by the German Science Foundation (grant BO-3139/2-1), and the LMU-innovativ Project BioMed-S: Analysis and Modelling of Complex Systems in Biology and Medicine.

References

Ancona, N., Maglietta, R., Piepoli, A., D'Addabbo, A., Cotugno, R., Savino, M., Liuni, S., Carella, M., Pesole, G., Perri, F., 2006. On the statistical assessment of classifiers using DNA microarray data. BMC Bioinformatics 7, 387.
URL http://dx.doi.org/10.1186/1471-2105-7-387

Baggerly, K. A., Coombes, K. R., 2009. Deriving chemosensitivity from cell lines: Forensic bioinformatics and reproducible research in high-throughput biology. Annals of Applied Statistics 3, 1309–1334.

Baty, F., Jaeger, D., Preiswerk, F., Schumacher, M. M., Brutsche, M. H., 2008. Stability of gene contributions and identification of outliers in multivariate analysis of microarray data. BMC Bioinformatics 9, 289.

Benner, A., Zucknick, M., Ittrich, T. H. C., Mansmann, U., 2010. High-dimensional cox models: the choice of penalty as part of the model building process. Biometrical Journal 52, 50–69.

Binder, H., Schumacher, M., 2008. Adapting prediction error estimates for biased complexity selection in high-dimensional bootstrap samples. Statistical Applications in Genetics and Molecular Biology 7, 12.

Bland, J. M., Altman, D. G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.

Boulesteix, A. L., Hothorn, T., 2010. Testing the additional predictive value of high-dimensional molecular data. BMC Bioinformatics 11, 78.

Boulesteix, A. L., Slawski, M., 2009. Stability and aggregation of ranked gene lists. Briefings in Bioinformatics 10, 556–568.

Bühlmann, P., Hothorn, T., 2007. Boosting algorithms: regularization, prediction and model fitting (with discussion). Statistical Science 22, 477–505.

Edgeworth, F. Y., 1887. On observations relating to several quantities. Hermathena 6, 279–285.

Goeman, J. J., 2010. L1 penalized estimation in the cox proportional hazards model. Biometrical Journal 52, 70–84.

Haury, A. C., Jacob, L., Vert, J. P., 2010. Increasing stability and interpretability of gene expression signatures. arXiv:1001.3109v1.

Hothorn, T., Bühlmann, P., 2006. Model-based boosting in high dimensions. Bioinformatics 22, 2828–2829.

Irizarry, R. A., Gautier, L., Bolstad, B. M., Miller, C., 2009. affy. BioconductorR package version 1.24.2: `http://www.bioconductor.org/packages/2.5/bioc/html/affy.html`.

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., Bergh, J., 2005. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Science 102, 13550–13555.

Minn, A. J., Gupta, G. P., Siegel, P. M., 2005. Genes that mediate breast cancer metastasis to lung. Nature 436, 518–524.

Opgen-Rhein, R., Strimmer, K., 2007. Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. Statistical Applications in Genetics and Molecular Biology 6, 9.

Pawitan, Y., Bjöhle, J., Amler, L., 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. Breast Cancer Research 7, R953–964.

Rousseeuw, P. J., Leroy, A. M., 2003. Robust regression and outlier detection. Wiley, Wiley, NJ.

Royston, P., Sauerbrei, W., 2007. Improving the robustness of fractional polynomial models by preliminary covariate transformation: A pragmatic approach. Computational Statistics & Data Analysis 51, 4240–4253.

Sauerbrei, W., Schumacher, M., 1992. A bootstrap resampling procedure for model building: Application to the cox regression model. Statistics in Medicine 11, 2093–2109.

Scherzer, C. R., Eklund, A. C., Morse, L. J., 2008. Molecular markers of early parkinsons disease based on gene expression in blood. Proceedings of the National Academy of Science 104, 3.

Shieh, A. D., Hung, Y. S., 2009. Detecting outlier samples in microarray data. Statistical Applications in Genetics and Molecular Biology 8, 13.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., Sellers, W. R., 2002. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.

Smyth, G., 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3, 3.

Sotiriou, C., Wirapati, P., Loi, S., 2006. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of the National Cancer Institute 98, 262–272.

Tang, B. M., McLean, A. S., Dawes, I. W., Huang, S. J., Lin, R. C., 2009. Gene-expression profiling of peripheral blood mononuclear cells in sepsis. Critical Care Medicine 37, 882–888.

Tibshirani, R., Hastie, T., 2007. Outlier sums for differential gene expression analysis. Biostatistics 8, 2–8.

Tomlins, S. A., Rhodes, D. R., Perner, S., 2005a. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. Science 310, 644–648.

Tomlins, S. A., Rhodes, D. R., Perner, S., Dhanasekaran, S. M., Mehra, R., Sun, X. W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J. E., Shah, R. B., Pienta, K. J., Rubin, M. A., Chinnaiyan, A. M., 2005b. Recurrent fusion of tmprss2 and ets transcription factor genes in prostate cancer. Science 310, 644–648.

Wang, Y., Klijn, J. G., Zhang, Y., 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 365, 671–679.

Wu, B., 2007. Cancer outlier differential gene expression detection. Biostatistics 8, 566–575.