Biostatistics (2018), **0**, 0, *pp.* 1–27
doi:10.1093/biostatistics/biosts3

Sequential rank agreement methods for comparison of ranked lists

CLAUS THORN EKSTRØM*, THOMAS ALEXANDER GERDS, ANDREAS KRYGER JENSEN

Biostatistics, Department of Public Health, University of Copenhagen, Øster Farimagsgade 5 B, P.O.B. 2099, DK-1014 Copenhagen K, Denmark

ekstrom@sund.ku.dk

Summary

The comparison of alternative rankings of a set of items is a general and common task in applied statistics. Predictor variables are ranked according to magnitude of association with an outcome, prediction models rank subjects according to the personalized risk of an event, and genetic studies rank genes according to their difference in gene expression levels. We propose a sequential rank agreement measure to quantify the rank agreement among two or more ordered lists. This measure has an intuitive interpretation, it can be applied to any number of lists even if some are partially incomplete, and it provides information about the agreement along the lists. The sequential rank agreement can be evaluated analytically or be compared graphically to a permutation based reference set in order to identify changes in the list agreements. The usefulness of this measure is illustrated using gene rankings, and using data from two Danish ovarian cancer studies where we assess the within and between agreement of different statistical classification methods.

*To whom correspondence should be addressed.

Key words: rank agreement; incomplete ranking; order statistic; methods comparison; permutation

1. Introduction

Ranking of items or results is common in scientific research and ranked lists occur naturally as the result of many statistical applications. Regression methods rank predictor variables according to magnitude of their association with an outcome, prediction models rank subjects according to their risk of an event, and genetic studies rank genes according to their difference in gene expression levels across samples. Two common research questions are of interest when several rankings of the same items are available: 1) To what extent do the lists agree on the rankings and how will that change as we go through the lists, and 2) Is it possible to identify an optimal rank until which the lists agree on the items?

A typical situation where these questions arise is in high-dimensional genomics studies such as genome-wide association studies where several analysis methods (e.g., regression methods, lasso, random forest) can be used identify and rank millions of gene variants according to their association with the outcome. The importance (i.e., ranking) of each gene variant may vary from method to method and a consensus summary of agreement of the findings is needed to determine which gene variants to investigate more closely in subsequent validation studies. To minimiize expenses it is only of interest to consider gene variants that have high ranking across the different methods. Another situation where multiple ranked lists appear is in machine learning where the stability of the ranks produced by a "black-box"-technique can be evaluted by bootstrapping the data repeatedly and comparing the ranks obtained from training the models on the bootstrapped data. Assessing which items are stable (i.e., consistant across bootstrap samples) will help to weed out possible fluke findings.

In this article we introduce sequential rank agreement for measuring agreement among ranked lists. The general idea is to define agreement based on the sequence of ranks from a subset of

the first d items in each list. As agreement metric we adapt the limits of agreement known from agreement between quantitative variables (Altman and Bland, 1983; Carstensen, 2010) but any measure for agreement could essentially be used. Our proposed approach allows us to compare multiple lists simultaneously, it provides a dynamic measure of agreement as a function of the depth in the lists, it places higher weight on items at the top of the list, it accommodates partially observed lists of varying lengths, and has a natural interpretation that directly relates to the ranks. Graphical illustration of sequential rank agreement allows us to infer any changepoints, i.e., a list depth where a substantial change in the agreement of the lists occur but we also provide asymptotical and randomization-based graphical tools to compare the observed rank agreement to the expected agreement found in non-informative data. In this sense our approach is a combination and generalization of some of the ideas of Carterette (2009) and Boulesteix and Slawski (2009). The former compares two rankings based on the distance between them as measured by a multivariate Gaussian distribution and the latter presents an overview of approaches for aggregation of ranked lists including bootstrap and leave-one-out jackknife approaches. We show the asymptotic distribution of the endpoint of agreement and discuss how to infer the distribution in small-sample situations using computational methods. This enables us to make inferences about the endpoint of agreement even for situations where there is no actual changepoint present but just a gradual decline in agreement.

Other recent approaches consider the intersection of lists as the basis for a similarity measure. However, simple intersection also places equal weights on all depths of the list and therefore Fagin *and others* (2003) and Webber *and others* (2010) proposed weighted intersections which put more emphasis on the top of the lists. Specifically, Webber *and others* (2010) define their rank-biased overlap (RBO) by weighting with a converging series to ensure that the top is weighted higher than the potentially non-informative bottom of the lists. It is possible to use the existing methods to calculate agreement of lists until a given depth, i.e., limited to the d items of each list. However,

the interpretation may not be straightforward, especially in the case of more than two lists, and they may not accommodate partial rankings.

Very recently, Hall and Schimek (2012) proposed a method for comparing pairwise rankings and derived the asymptotic distribution of the endpoint where the two ranked lists no longer are in agreement. Their approach was based on anchoring one of the two lists and subsequently generating a sequence of 0s and 1s depending on whether the ranks in the second list was close to the rank from the anchored list. Sampath and Verducci (2013) followed up on this idea for pairwise comparison of lists but used penalties based on a truncated geometric probability instead of a 0-1 process and they evaluated the distribution of the endpoint of agreement by computational approaches. The asymptotic distribution in the Hall and Schimek (2012) paper is based on letting the number of *lists* increase to infinity which is a situation that is only relevant in special cases, whereas the simulation-based null distribution approach of Sampath and Verducci (2013) does not rely on asymptotic results to evaluate their pairwise findings.

The manuscript is organized as follows: The next section defines sequential rank agreement for multiple ranked lists and discuss how to handle incomplete lists. In section 3 we discuss approaches to evaluate the results obtained from sequential rank agreement. Finally we apply the sequential rank agreement to two Danish ovarian cancer studies and compare our method to the method of Hall and Schimek (2012) in a small sample simulation study before we discuss the findings along with possible extensions. The approaches presented in this manuscript are available in the R package `SuperRanker` which can be found on CRAN.

2. Methods

Consider a set of P different items $X = \{X_1, \ldots, X_P\}$ and a ranking function $R : \{X_1, \ldots, X_P\} \to \{1, \ldots, P\}$, such that $R(X_p)$ is the rank of item X_p. The inverse mapping R^{-1} gives the item $R^{-1}(r)$ that was assigned to rank $r \in \{1, \ldots, P\}$. An ordered list is the realization of a ranking

function R applied to the set of items X. Panels (a) and (b) of Table 1 show a schematic example of these mappings. Thus if $R_l^{-1}(1) = X_{34}$ then item X_{34} is ranked first in list l and similarly $R_l(X_{34}) = 1$.

In all what follows we consider a fixed set of items and consider the ranking function to be a random variable. Thus, let $R_1(X), \ldots, R_L(X)$, $L \geqslant 2$, be a sample of L independent identically distributed draws from an unknown probability distribution function Q. One aim is then to test how much Q resembles the uniform distribution which assigns probability $1/P!$ to each of the $P!$ different possible rankings.

The agreement among the lists regarding the rank given to an item X_p can be measured by the variance across the lists

$$\begin{aligned} A(X_p) &= \mathbb{E}_Q \left[\left(R(X_p) - \mathbb{E}_Q \, R(X_p) \right)^2 \right] \\ &= \sum_{r \in \Pi} \left(r(X_p) - \mathbb{E}_Q \, R(X_p) \right)^2 Q(r), \end{aligned} \tag{2.1}$$

where Π is the set of all permutations of X, Q is a probability mass function on Π, and $\mathbb{E}_Q \, R(X_p) = \sum_{r \in \Pi} r(X_p) Q(r)$. The empirical counterpart is

$$\widehat{A}_L(X_p) = \frac{1}{L-1} \sum_{i=1}^{L} (R_i(X_p) - \overline{R}_L(X_p))^2, \quad \overline{R}_L(X_p) = \frac{1}{L} \sum_{i=1}^{L} R_i(X_p). \tag{2.2}$$

For each item, the function $\widehat{A}_L$ has an interpretation as the expected Euclidean distance of the individual rankings from the expected ranking over the L lists, and it corresponds to the same measure that is used in method comparison studies to compute the limits of agreement (Altman and Bland, 1983).

For an integer $1 \leqslant d \leqslant P$ we define the expected set of unique items found by merging the first d elements across the possible lists:

$$S(d) = \left\{ X_p; \left(\sum_{r \in \Pi} 1 \left(r(X_p) \leqslant d \right) Q(r) \right) > \varepsilon \right\} \tag{2.3}$$

where $1(\cdot)$ denotes the indicator function, and where $\varepsilon \in [0, 1)$ is a pre-specified constant that sets the minimum proportion of lists that an item must be present in before it is added to $S(d)$.

When $\varepsilon = 0$ then an item is included as soon as it is present in just one list. The empirical counterpart is the set of unique items ranked less than or equal to d in any of the L lists:

$$\widehat{S}_L(d) = \left\{ X_p; \left(\frac{1}{L} \sum_{l=1}^{L} 1\left(R_l(X_p) \leqslant d\right) \right) > \varepsilon \right\}, \tag{2.4}$$

which is exemplified in Panel (c) of Table 1.

We define the *sequential rank agreement* as the weighted expected agreement of the items found in the set $S(d)$:

$$\text{sra}(d) = \begin{cases} \frac{1}{|S(d)|} \sum_{p \in S(d)} A(X_p) & \text{when } |S(d)| > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

where $|S(d)|$ is the cardinality of the set $S(d)$. As stated, we are only interested in $\text{sra}(d)$ when $|S(d)| > 0$ so the value 0 above is somewhat arbitrary. The empirical counterpart when $|S(d)| > 0$ is equivalently given by

$$\widehat{\text{sra}}_L(d) = \frac{\sum_{\{p \in \widehat{S}_L(d)\}} (L-1) \widehat{A}_L(X_p)}{(L-1)|\widehat{S}_L(d)|}. \tag{2.6}$$

Values of sra close to zero when $|S(d)| > 0$ suggest that the lists agree on the rankings while larger values suggest disagreement. If $|S(d)| = 0$ then no items were sufficiently frequent among the observed lists and we can conclude that the lists do not agree above the threshold ε. If the ranked lists are identical then the value of sequential rank agreement will be zero for all values of d.

2.1 *Interpreting and applying sequential rank agreement*

The sequential rank agreement is equivalent to the pooled variance of the items found in $S(d)$. Thus, the square root of the sequential rank agreement measures the average of the average difference in rankings among the lists for the items we have included until depth d.

In method comparison studies the observed agreement is compared to a pre-specified acceptable limit and we can do similarly. For easy visualization of the rank agreement we suggest to plot $\sqrt{\widehat{\text{sra}}_L(d)}$ corresponding to the pooled SD against d.

As an example consider the data by Golub (1999) (also found in Dudoit *and others* (2002)) where 3051 gene expression values measured on 38 tumor mRNA samples were used to classify between acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Several possible analysis methods can be used for this for example marginal unequal variances two-sample t tests, marginal logistic regression analyses, elastic net logistic regression (Friedman *and others*, 2010), and marginal maximum information content correlations (MIC) (Reshef *and others*, 2011), and we would like to identify a set of genes that consistently are most likely to be associated to leukemia.

For the first two methods, the genes were ranked according to minimum p value, for logistic regression the genes were ordered by size of the corresponding coefficients (after standardization), and MIC was ordered by absolute correlation which resulted in the top rankings seen in Table 2. The ranked lists appear to agree that genes 2124 and 829 are among the most interesting while the best ranked gene from MIC, gene 378, is not found in the top 10 for two of the other methods.

The sequential rank agreement curve (using $\varepsilon = 0$ to decide that an item should be present in just a single list before it is included) shown in the left plot of Figure 1 show the average distance in ranks for the genes considered among the first d positions. Not surprisingly, the sequential rank agreement is better towards the top of the lists (smaller values on the y axis corresponds to *better* agreement) than towards the bottom of the lists. It is clear from the curve in Figure 1 that there is a substantial deterioration in agreement (higher sra) after depth 5. Thus, if we were to restrict attention to a small set of predictors then our prime focus would be on the items found among the top-5 lists from Table 2). The choice of which depth until which we think the lists agree can be chosen either from a pre-specified threshold for an acceptable difference in rankings or from a pre-specified item set size. A changepoint analysis on the sequential rank agreement would be able to identify depths where a substantial increase/change in rank agreement occurs and would be another way to identify sets of items that share agreement among the lists if a pre-specified

acceptable rank agreement threshold is not given.

Generally, changes in the level of rank agreement suggest that there are sets of items that are ranked similarly in all lists while other items constitute set(s) that have been assigned vastly different ranks. When the lists are likely to agree on a few top ranked items then the sequential rank agreement curve will start at a low level and then increase until it levels off exactly as is seen in Figure 1.

2.2 *Analysis of incomplete lists*

Incomplete or partial lists are a common occurrence that arise, for example, in case of missing data (items), when comparing top d list results from publications, or when some methods only rank a subset of the items. For example, penalized regression such as the Lasso provides a sparse set of predictors that have non-zero coefficients. There is no obvious ordering of the set of predictors whose coefficient has been shrunk to zero and thus we end up with a partial ordering. Incomplete lists also occur if for example the analyst restricts attention to the ranks of items that have been found to be statistically significant.

Sequential rank agreement can be generalized to incomplete lists in the following way. Let $\Lambda_l \subset X$ be the subset of d_l items that have been ranked highest in list l. The case where all lists are incomplete at the same depth d corresponds to $d_1 = \cdots = d_L = d$. For incomplete lists the rank function becomes

$$\widetilde{R}_l(X_p) = \begin{cases} \{R_l(X_p)\} & \text{for } X_p \in \Lambda_l \\ \{d_l + 1, \ldots, P\} & \text{for} X_p \notin \Lambda_l \end{cases} \tag{2.7}$$

where we only know that the rank for the unobserved items in list l must be larger than the largest rank observed in that list.

The agreement, $A(X_p)$, cannot be computed directly for all predictors in the presence of incomplete lists because the exact rank for some items will be unknown. Also, recall that the rankings within a single list are not independent since each rank must appear exactly once in each

list. Thus, we cannot simply assign the same number (*e.g.*, the mean of the unassigned ranks) to the missing items since that would result in *less* variation of the ranks and hence less variation of the agreement, and it would artificially introduce a (downward) bias of agreement for items that are missing in multiple lists.

Instead we randomize the ranks $\{d_l + 1, \ldots, P\}$ to the items that do not occur in list Λ_l. One realization of the L rankings of the set X is obtained by randomizing the missing items of each list. By randomizing a large number of times we can compute (2.5) for each realization, and then compute the sequential rank agreement as the pointwise (for each depth) average of the rank agreements. The algorithm is described in detail in Algorithm 1.

The proposed approach is based on two assumptions: 1) that the most interesting items are found in the top of the lists, and 2) that the ranks that are missing from the lists provide so little information that it is reasonable to assume that they can be represented by a random order. The first assumption is justifiable because we have already accepted that it is reasonable to rank the items in the first place. The second assumption is fair in the light of the first assumption provided that we have a "sufficiently large" part of the top of the lists available.

When the two assumptions are satisfied then it is clear that the interesting part of the sequential rank agreement curves is restricted to depths where the number of items without ranks available is low. Similar to fully observed lists we generally expect the sequential rank agreement to start low and then increase unless the lists are completely unrelated (in which case the sequential rank agreement will be constant at a high level) or if the lists mostly agree on the ranking (in which case the sequential rank agreement will also be constant but at a low level). For incomplete ranked lists we also expect a changepoint around the depth where the lists are become incomplete. This is an artefact stemming from the fact that we assume that the remainder of the lists can be replaced by a simple permutation of the missing items. Note that if an item is ranked highly in a few lists but unranked in the remaining lists then it gets poor rank agreement since

we only compare whether the lists agree on their rankings and in this case they clearly do not.

The right-most plot of Figure 1 shows the impact of restricting the Golub data such that only top-20 lists are available instead of full lists of length 3051 (20 was chosen to resemble the list lengths that might be harvested from published manuscripts). The sequential rank agreement increases much quicker because the incomplete lists introduce more noise in the estimation of the agreement, but it is still possible to see that the top of the list has a sequantial rank agreement that is not substantially different from the full lists.

3. Evaluating sequential rank agreement results

To evaluate the sequential rank agreement values we propose two different benchmark values corresponding to two different hypotheses. We wish to determine if we observe better agreement than would be expected if there were no relevant information available in the data.

The first reference hypothesis is

$$H_0 : \text{The list rankings correspond to completely randomly permuted lists} \tag{3.8}$$

which not only assumes that there is no information in the data on which the rankings are based but also that the methods used to provide the rankings are completely independent.

Alternatively, we can remove the restriction of the independence among the methods used to generate the L ranked lists and only require that there is no information contained in the rankings but that the rankings are all based on applying the method/approaches to the same data

$$\widetilde{H}_0 : \text{The list rankings are based on data that contain no association to the outcome.}$$

This alternative null hypothesis addresses the fact that some ranking methods are more likely

to provide similar rankings of the same data because the ranking methods focus on the same features of the data rather than because of any information contained in the data.

3.1 *Permutation-based inference*

H_0 is a quite unrealistic null hypothesis but we can easily obtain realizations from that null hypothesis simply by permuting the items within each list and then computing the sequential rank agreement for the permuted lists. In the fully observed case each experiment contains L lists of random permutations of the items in X. For the incomplete case we first permute the items $X_1, \ldots, X_P$ and then assign missing ranks for list l from d_l to P (*i.e.*, each list has the same number of observed rankings as was observed for list l in the original dataset). The sequential rank agreement curve from the original lists can then be compared to, say, the pointwise 95% quantiles of the observed rank agreements obtained under H_0.

To obtain the distribution under $\widetilde{H}_0$ the idea is to repeat the ranking procedures for unassociated data many times. For each resample, we

1. first permute the outcome variable in the dataset. This removes any association between the predictor variables and the outcome while keeping the structure in the predictors, and

2. we apply the same methods that was used for the original data to the permuted dataset to generate L new rankings and compute the sra for the unassociated data.

Note that we only permute the outcomes and thus preserve the internal structure of the predictors. This randomization approach requires that the original data is available and as such it may not be possible to evaluate $\widetilde{H}_0$ in all situations.

If the sequential rank agreement for the original data lies substantially below the distribution of the sequential rank agreements obtained under either H_0 or $\widetilde{H}_0$ then this suggests that the original ranked lists agree *more* than expected in data with no information, and therefore that

the information in the lists is significantly more in agreement than what would be expected.

Figure 1 shows the empirical distributions of sequential rank agreement under H_0 and $\widetilde{H}_0$ each based on 400 permutations of the Golub data from Section 2.1. Figure 1 indicates that the observed sequential rank agreement for the Golub data is significantly better than what would be expected by chance for data that contain no information since it lies below the reference areas if the lists were just random (H_0 corresponding to the blue area). However, if we consider $\widetilde{H}_0$ then the sequential rank agreement is just inside the red area and we conclude that the agreement seen in the Golub data is *not* significantly better than what we would expect when we remove the association between the predictors and the outcome.

The incomplete data also suggests that there may be at most 1 or 2 ranked items towards the top of the lists that yield a result better than what would be expected (the bottom-right plot). Not surprisingly, the sequential rank agreement under $\widetilde{H}_0$ is lower than the sequential rank agreement under H_0 because the four methods used to rank the data (t test, logistic regression, elastic net, and MIC) generally tend to identify (and rank) similar predictors even if there are only spurious associations.

It is important to stress that neither H_0 nor $\widetilde{H}_0$ are related to questions regarding the association between the outcome and the predictors in the dataset. Both hypotheses are purely considering how the rankings agree in a situation where there is no relevant information available in the data used for creating the rankings. It is also worth pointing out, that if the lists are short (P low) and there are few lists (L low) then the number of possible different permutations under the null is small and the p value obtained may be fluctuating if the number of permutations is small. We have found that a number of permutations over 500 works well for smaller samples. The number of permutations can be lowered for larger data sizes.

3.2 *Asymptotic inference of change in agreement*

In many applications it is of interest to estimate a list depth which satisfies a changepoint criterion since that corresponds to a change in agreement among the list ranks. In particular, a changepoint will provide a data-driven indicator as to the depth until the lists exhibit a change in rank agreement, and would consequently be an obvious choice for identifying the set of items that the lists agree the upon the most. In this section we investigate the theoretical properties of our proposed method for this specific task. As in Hall and Schimek (2012) we consider an infinite set of lists and study the asymptotic behaviour for $L \to \infty$. The list lengths are not allowed to change with L since the lengths are fixed in most applications.

We start by showing that $\widehat{\text{sra}}_L$ is a consistent estimator of sra for $L \to \infty$.

THEOREM 3.1 Assume that $\{R_l(X)\}_{l=1}^{L}$ are independent draws from a probability distribution Q on the set of lists Π. Then, $\|\widehat{\text{sra}}_L - \text{sra}\|_\infty = o_P(1)$. *Proof.* See Appendix A in the supplementary material. □

We now define the changepoint as the first crossing point of the sequential rank agreement and a threshold function $q\colon \{1, \ldots, P\} \mapsto \mathbb{R}_{\geqslant 0}$. The values of q could be a deterministic constant or, for example, the limits-of-agreement obtained in randomly permuted lists corresponding to the null-hypothesis in equation (3.8). We define the superlevel set of the sequential rank agreement with respect to q as

$$\mathcal{L}(q) = \{d : \text{sra}(d) \geqslant q(d)\}\,. \tag{3.9}$$

A changepoint $d^*(q)$ in the list agreement is then defined by the position

$$d^*(q) = \begin{cases} \inf(\mathcal{L}(q)) & |\mathcal{L}(q)| > 0 \\ P & |\mathcal{L}(q)| = 0 \end{cases} \tag{3.10}$$

corresponding to the first list depth where the sequential rank agreement exceeds the threshold if such a position exists. Otherwise, the full list is in agreement according to q and the changepoint

is set to the full length of the lists. The empirical superlevel set is similarly defined as

$$\widehat{\mathcal{L}}_L(\widehat{q}_L) = \{d : \widehat{\mathrm{sra}}_L(d) \geqslant \widehat{q}_L(d)\} \tag{3.11}$$

where we allow the threshold function to depend on the sample size as well. The estimated changepoint is therefore

$$\widehat{d^*_L}(\widehat{q}_L) = 1(|\widehat{\mathcal{L}}_L(\widehat{q}_L)| > 0) \inf \widehat{\mathcal{L}}_L(\widehat{q}_L) + 1(|\widehat{\mathcal{L}}_L(\widehat{q}_L)| = 0)P. \tag{3.12}$$

The consistency of the estimated changepoint, $\widehat{d^*_L}(\widehat{q}_L)$, follows from Theorem 3.1 by the following corollary.

Corollary 3.1 Let $\widehat{q}_L$ be a positive threshold function such that $\|\widehat{q}_L - q\|_\infty = o_P(1)$ for some limiting function q. Then, $\widehat{d^*_L}(\widehat{q}_L) \xrightarrow{P} d^*(q)$ for $L \to \infty$. *Proof.* See Appendix B in supplementary material. □

Corollary 3.1 indicates that we can use the threshold function $\widehat{q}_L$ estimated under the null hypothesis as discussed in the previous section as a limiting threshold function for inferring the depth d, where the observed sequential rank agreement first crosses the threshold of the null threshold, *i.e.*, the depth until which the observed ranked lists are in better agreement than expected under the null hypothesis. In that sense the threshold function serves the same role as the limits of agreement in method comparison studies, except that the threshold function is not constant but can accommodate the changing nature of the number of items used for the computation of the sequential rank agreement for a given depth. In practice we can compute an estimate of the threshold function under the null using the permutation approach sketched in the previous section which makes it relevant even for small sample settings.

4. Application to ovarian cancer data

We now consider an application of the sequential rank agreement to two datasets consisting of MALDI-TOF (Matrix-Assisted Laser Desorption/Ionization Time Of Flight) mass spectra obtained from blood samples from patients with either benign or malignant ovarian tumors. The datasets are sub-samples of the Danish MALOVA and DACOVA study populations.

The MALOVA study is a multidisciplinary Danish study on ovarian cancer (Hogdall *and others*, 2004) where all Danish women diagnosed with an ovarian tumor and referred for surgery from the participating departments of gynecology were enrolled continuously from December 1994 to May 1999. For the purpose of illustration we use a random sub-sample of 119 patients with a total of 58 patients with malignant ovarian cancers as cases and 61 patients with benign ovarian tumors as controls. The DACOVA study is another multidisciplinary Danish study on ovarian cancer which included about 66% of the female population of Denmark (Bertelsen, 1991). The study aimed to continuously enroll all patients that were referred to surgery of an ovarian tumor clinically suspected to be cancer during the period from 1984 to 1990. Similarly, we use a random sub-sample from the DACOVA study of 113 patients with a total of 54 malignant ovarian cancers and 59 benign ovarian tumors/gynecologic disorders.

Each spectrum consists of 49642 samples over a range of mass-to-charge ratios between 800 to 20000 Dalton which we downsample on an equidistant grid of 5000 points by linear interpolation. We then preprocess the downsampled spectra individually by first removing the slow-varying baseline intensity with the SNIP algorithm (Ryan *and others*, 1988) followed by a normalization with respect to the total ion count. Finally, we standardize the 5000 predictors to have column-wise zero mean and unit variance in each dataset.

We use the two datasets to illustrate how the sequential rank agreement can be applied in two different scenarios. In the first scenario we assess the agreement of four different statistical classification methods in how they rank the predictors according to their importance for distin-

guishing benign and malignant tumors. In the second scenario we assess the agreement among rankings of individual predicted risks of having a malignant tumor. The first scenario is relevant in the context of biomarker discovery and the latter is important e.g., when ranking patients according to immediacy of treatment.

Four classification methods are considered: Random Forest (Breiman, 2001) implemented in the R package `randomForest` (Liaw and Wiener, 2002), logistic Lasso (Tibshirani, 1996) and Ridge regression (Segerstedt, 1992) both implemented in the R package `glmnet` (Friedman *and others*, 2010), and Partial Least Squares Discriminant Analysis (PLS-DA) (Boulesteix, 2004) implemented in the R package `caret` (Kuhn, 2014). All four methods depend on a tuning parameter. The tuning parameter for Lasso and Ridge regression is the degree of penalization, and for PLS-DA it is the number of components (the dimensionality of the subspace). We estimate these separately for each sub-sample by a 20 times repeated 5-fold cross-validation procedure. For the Random Forest we grow a fixed number of 5000 trees and let the tuning parameter be the number of predictors randomly sampled at each split. We estimate this by a binary search with respect to minimizing the Out-of-Bag classification error estimate.

In both scenarios we use the MALOVA data to train the statistical models, and in both situations the agreements are assessed with respect to perturbations of the training data in the following manner. We repeatedly draw a random sub-sample (without replication) consisting of 90% of the MALOVA observations and train the four models on each sub-sample. We use 1000 iterations for the sub-sampling procedure.

The implementation of Lasso and Ridge regression in the `glmnet` package offers three different cross-validated optimization criteria for the penalty parameter: total deviance, classification accuracy and area under ROC. We apply all three criteria to our data to investigate their effect on the agreements. Note also that the Lasso models produce incomplete lists depending on the value of the penalty parameter.

4.1 *Agreement of predictor rankings*

For each of the four methods, each of the 1000 models trained on the 1000 sub-samples of the MALOVA data produces a ranking of the 5000 predictors according to their importance for discriminating between the tumor types. For the Random Forest classifier the predictors are ranked according to the Gini index, while for the logistic Lasso and Ridge regression models we order by absolute magnitude of the estimated regression coefficients. For the PLS-DA model the importance of the predictors is based on a weighted sum of the absolute coefficients where the weights are proportional to the reduction in the sums of squares across the components.

The right panel of Figure 2 shows the sequential rank agreement of the estimated importance of the 5000 predictors. For clarity of presentation we zoom in on the agreement up to list depth 600. At deeper list depths all agreement curves are approximately constant. As expected, most of the sequential rank agreement curves start low, indicating good agreement, followed by an increase until they approximately become constant. This has the interpretation that the agreement across the different sub-samples is higher in the top as compared to the tail of the lists for all these classification methods. The changepoints where the curves become approximately constant are the list depths where the ranks of the remaining items become close to uniformly random.

A not expected shape of the agreement curves is seen for the Ridge models for all three tuning criteria. They all show higher disagreement in the top of the lists followed by a decrease. The reason behind this behavior is rather subtle. Looking at the distribution of the absolute value of the regression coefficients we see that a large proportion of them are numerically very close to zero and have almost equal absolute value. This is a general feature of the Ridge models in this dataset and seen for all the 1000 trained models. This implies that when predictors are ranked according to the magnitude of their coefficients, their actual order becomes more uncertain and more close to a random permutation. This problem can be alleviated by truncating all predictors with absolute coefficient values below a given threshold thereby introducing an artificial incompletion

of the lists. For the Ridge models tuned with the deviance criterion, Figure 3 (left) shows the the sequential rank agreement where for each of the 1000 trained models the predictors were artificially censored when their absolute coefficient value was lower than the 0.1% quantile of the 5000 absolute coefficient values. The curve was calculated using Algorithm 1 with $B = 1000$ and $P = 5000$. The corresponding curve from Figure 2 (right panel) is shown for comparison. Even though the number of predictors with missing ranks is very small compared to the total number of predictors, the effect on the sequential rank agreement is substantial and with the artificial censoring the shape of the curves is as expected, starting low and then increasing.

Looking at the agreement curves for the Lasso models in Figure 2 (right) we clearly see the effect of the sparsity inducing penalization giving rise to incomplete lists. These curves were similarly calculated using 1 and 1000 random permutations. Under the deviance optimization criterion the median number of non-zero coefficients was 33 (range 16 to 50) and for the class accuracy criterion 14 (range 4 to 56). These values correspond to the list depths where the agreement curves become constant as a result of the subsequent censoring.

4.2 *Agreement of individual risk predictions*

To assess the stability of the individual risk predictions we apply the predictors from the DACOVA dataset to each of the models. The predicted probabilities are then ranked in decreasing order such that the patients with the highest risk of a malignant tumor appears in the top of the list. Figure 2 (left) shows the sra separately for each method, based on the 1000 risk predictions obtained from the models trained in the same 1000 random sub-samples of the MALOVA data.

Most curves start low and then increase indicating higher agreement among high risk patients. This is expected if we rank the individuals according to highest risk of disease. However, it is also expected that individuals with very low risk also show high agreement. In this case we order the patients according to (high) risk prediction but we could essentially also have reversed the order

to identify the patients that have low risk prediction.

An exception is the risk prediction agreement for the Lasso tuned with the AUC criterion which shows very low agreement among the high values of the predicted risks. The reason is that optimizing the penalty parameter with respect to the AUC criterion tends to favor a very high penalty value causing only a single predictor to be selected in each of the 1000 iterations. This results in a lack of generalizability to the DACOVA data which gives rise to the higher disagreement in the predicted risks. In the extreme case where the penalty becomes so high that none of the predictors are selected by the Lasso, the sequential rank agreement for the predicted probabilities becomes undefined since all the ranks will be ties.

Comparing the left and right panels of Figure 2 it can further be seen that some of the methods show better agreement with respect to the predicted probabilities than for ranking the importance of the predictors and vice versa. Ridge regression shows higher agreement across training sets for the risk predictions than PLS-DA, and PLS-DA shows higher agreement for predictor importance than Ridge regression.

Lasso shows similar agreement for ranking the risk predictions as PLS-DA (except for the AUC criterion), and poorer agreement for ranking predictors. This reason for the latter is the high auto-correlation between the intensities in the mass spectra which leads to collinearity issues in the regression models. It is well-known that variable selection with the Lasso does not perform very well when the predictors are highly correlated. The collinearity does, however, not affect the agreement of the risk predictions (Figure 2, left panel), since the specific variable selected is not that important for a group of highly correlated predictors when the purpose is risk predictions.

It appears that Ridge regression tuned with the AUC criterion achieves the best performance with respect to the stability of ranking the individual predicted risk probabilities. It must, however, be stressed that the sequential rank agreement in this application is only concerned with the agreement of the risk predictions across sub-samples and not with the actual accuracy of the

risk predictions. Thus, we also computed the AUC values for the different models based on the DACOVA data. The distributions across the 1000 sub-samples for a selection of the models is shown in the right panel of Figure 3. Here we see that PLS-DA attains the highest AUC values with a median value of 0.70 while the Ridge model with the AUC criterion attains a median AUC of 0.49. This implies that while Ridge regression optimized with respect to the AUC criterion achieves the best sequential rank agreement, it performs similar to a random coin toss with respect to classifying the DACOVA patients. In practice both concerns are of importance.

5. Simulation study — comparison of list agreements

We present results from a simulation study where we investigated the small sample properties of the sequential rank agreement and compared it to the topK method (Hall and Schimek, 2012) with respect to list depth agreement.

The purpose of the simulations were twofold: First we want to investigate the rank agreement as it changes with threshold q and number of lists L. Secondly, we want to compare the results (and conclusions) from sra with the topK for a realistic situation where the true, underlying agreement is not governed by a simple probability distribution. Thus, we are interested in two features of the methods: the depth until which they agree, and the number of unique predictors found in that set.

To define the depth of agreement for sra we set a constant threshold function to an integer q and report the first crossing point, i.e., the smallest list depth where sra exceeds q (as implemented in the `sra` function from package `SuperRanker` using the median absolute distance argument). For topK we use the function `j0.multi` which is implemented in the `R` package `TopKLists`. Specifically, we set the tuning parameter `v` of `j0.multi` to the value 6 and the window parameter `d` to q and report the output parameter `maxK` as the depth of agreement. Thus, to make this comparison we assume that the first crossing point of sra and the result of the topK method

measure the same underlying feature.

The simulations should mimic a data analysis situation where we have a single dataset and where important features are identified (and ranked) using marginal t tests. We want to used agreement to understand the stability of the observed feature selection. In each simulation run, we first generated an "original" dataset with 1000 predictors and 400 observations. The predictors were drawn independently from a standard Gaussian distribution with variance 1 such that

$$E(y_i) = \sum_{j=1}^{15} x_{ij},$$

where y_i is the ith response, and x_{ij} is the jth predictor for the ith measurement.

For each "original dataset" we obtained L ranked lists of the 1000 predictors by drawing L bootstrap samples (with replacement) of 400 observations and then ranking the 1000 predictors according to their marginal t test statistics. Thus, we assessed the depth of agreement among lists that are ranked with the same statistical method on bootstrap versions of the same dataset. We report results from two scenarios each based on 1000 simulated datasets:

Scenario I: Fix the number of lists $L = 8$ and vary the threshold $q \in \{3, 4, 5, 6, 7, 8, 9, 10\}$.

Scenario II: Fix the threshold $q = 5$ and vary the lists, $L \in \{3, 5, 10, 50\}$.

In both scenarios we summarized the distribution of the estimated depth of agreement as well as the average number of unique predictors found in the set of predictors which is selected by the estimated depth of agreement. The results from Scenario I are shown in the left panel of Figure 4. The violin plots (with rectangular kernel) show the distributions of the estimated depths of agreement for both methods. As expected the depth of agreement increased when the threshold for agreement/window increased.

We see that sra results in a substantially lower depth of agreement than the topK method. Also the average numbers of unique predictors (bold numbers inside the plots) which ideally should be 15 to reflect the number of true underlying predictors are markedly smaller — and close to

the true value — for sra. Even larger differences were found when we used the Euclidean distance instead of the median absolute distance for the sequential rank agreement (results not shown). The right panel of Figure 4 shows the results from Scenario II. The number of lists has little impact on the results, and again the sra is more conservative than topK and as a consequence sra includes fewer predictors in the selected set where the lists agree.

Note that the effect sizes of the 15 predictors in the model is the same and in practice we observe that the majority of the 15 predictors are generally picked in each sample but that their individual rankings vary substantially in the top 15 within each bootstrap sample. If the number of influential predictors is lessened then the variance in depth estimation and number of predictors is reduced.

6. Discussion

In this article we address the problem of comparing ranked lists of the same items. Our proposed method can handle both the situation where the underlying data to generate the ranked lists are available and the situation where the only available data is the actual ranked lists. In addition, incomplete ranked lists where only the ranks of the top k ranked items are known can be accommodated as well. The proposed agreement measure can be interpreted as the average distance between an item's rank and the average rank assigned to that item across lists.

The sequential rank agreement can be used to determine the depth at which the rank agreement becomes too large to be desirable based on prior requirements or acceptable differences, or it can be used to visually determine when the change in agreement becomes too large. In that regard the investigator can have prior limits on the level of agreement that is acceptable.

We have shown that sra is very versatile: it can be used not only to compare ranked lists of items produced from different samples/populations but that it also can be used to study the ranks obtained from different analysis methods on the same data, and to evaluate the stability

of the ranks by bootstrapping (or sub-sampling) the data repeatedly and comparing the ranks obtained from training the models on the bootstrapped data.

While the sequential rank agreement is primarily an exploratory tool we have suggested two null hypotheses that can be used to evaluate the sequential rank agreement obtained. Note that none of the two null hypotheses are concerned with the actual "true ranking" but are purely concerned with consistency/stability of the rankings among the lists, and consequently we cannot determine if the rankings are good but only whether they agree. The sequential rank agreement curve can be compared visually to the curves obtained under either of the null distributions and simple point-wise p-values can be obtained for each depth by counting the number of sequential rank agreements under the null hypothesis that is less than or equal to the observed rank agreement.

Finally, we have — whenever possible — used all available ranks from the lists. We could choose to restrict attention to the rank of items which show evidence for significance in their models. That would ensure that there would be put less emphasis on the agreement of the non-significant items and it would be easier to identify a change in agreement among the items that were deemed to be relevant. In our application section we have successfully introduced such an artificial censoring for the predictor rankings obtained with ridge regression.

We note that the sequential rank agreement is still marred by problems that generally apply to ranking of items and/or individuals. Collinearity in particular can be a huge problem when bootstrapping data or when comparing different analysis methods. For example, marginal analyses where each item is analyzed separately will assign similar ranks to two highly correlated predictors while methods that provide a sparse solution such as the Lasso will just rank one of the two predictors high while the other might have a very low rank. Thus in such a scenario we would expect low agreement of the rankings from Lasso and marginal analyses simply because of the way correlated predictors are handled. This is not a shortcoming of the sequential rank

agreement but is a problem general to all ranked lists.

Another caveat with the way the sequential rank agreement is defined is the use of the standard deviation to measure agreement. The standard deviation is an integral part of the limits-of-agreement as discussed by Altman and Bland (1983). However, the standard deviation can also be unstable when the number of observations is low and alternative measures such as the median absolute deviance may prove more stable in some situations. However, the current definition using the standard deviation is analogous to the approach used for agreement in method comparison studies so we have used that.

In conclusion we have introduced a method for evaluation of ranked (partial/censored) lists that can be easily interpreted and that can be applied to a large number of situations. The method presented here can be adapted further by using it to compare and classify statistical analysis methods that agree on the rankings they provide or by using the rank agreement to optimize a hyper-parameter in, say, elastic net regularized regression where the rank agreement is used to determine the mixing proportion between the L_1 and the L_2 penalty. Finally, the proposed method may be adapted (with some additional assumptions) to the situation where there are put equal emphasis on both ends of the lists and not just on the top of the lists. We will be investigating these extensions further in the future.

7. Supplementary Materials

The reader is referred to the on-line Supplementary Materials for technical appendices and proofs.

References

Altman, Douglas and Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *The Statistician* **32**, 307–317.

Bertelsen, K. (1991). Protocol allocation and exclusion in two Danish randomised trials in

ovarian cancer. *British journal of cancer* **64**(6), 1172.

Boulesteix, Anne-Laure. (2004). PLS dimension reduction for classification with microarray data. *Statistical applications in genetics and molecular biology* **3**(1), 1–30.

Boulesteix, Anne-Laure and Slawski, Martin. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics* **10**(5), 556–568.

Breiman, L. (2001). Random forests. *Machine learning* **45**(1), 5–32.

Carstensen, Bendix. (2010). *Comparing Clinical Measurement Methods: A Practical Guide*. Wiley.

Carterette, Ben. (2009). On rank correlation and the distance between rankings. In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09. New York, NY, USA: ACM. pp. 436–443.

Dudoit, Sandrine, Fridlyand, Jane and Speed, Terence P. (2002, March). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association* **97**(457), 77–87.

Fagin, Ronald, Kumar, Ravi and Sivakumar, D. (2003, January). Comparing Top k Lists. *SIAM Journal on Discrete Mathematics* **17**(1), 134–160.

Friedman, Jerome, Hastie, Trevor and Tibshirani, Rob. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* **33**(1), 1–22.

Golub, T. R. (1999, October). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286**(5439), 531–537.

Hall, Peter and Schimek, M. G. (2012). Moderate deviation-based inference for random degeneration in paired rank lists. *JASA* **107**, 661–672.

Hogdall, E. V., Ryan, A, Kjaer, S K, Blaakaer, J, Christensen, L, Bock, J E, Glud, E, Jacobs, I J and Hogdall, C K. (2004, Jun). Loss of heterozygosity on the X chromosome is an independent prognostic factor in ovarian carcinoma: from the Danish "MALOVA" ovarian carcinoma study. *Cancer* **100**(11), 2387–2395.

Kuhn, Max. (2014). *caret: Classification and Regression Training*. R package version 6.0-24.

Liaw, Andy and Wiener, Matthew. (2002). Classification and regression by randomForest. *R news* **2**(3), 18–22.

Reshef, David N, Reshef, Yakir A, Finucane, Hilary K, Grossman, Sharon R, McVean, Gilean, Turnbaugh, Peter J, Lander, Eric S, Mitzenmacher, Michael and Sabeti, Pardis C. (2011, December). Detecting novel associations in large data sets. *Science (New York, N.Y.)* **334**(6062), 1518–24.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H. and Cousens, D. R. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* **34**(3), 396–402.

Sampath, Srinath and Verducci, Joseph S. (2013). Detecting the end of agreement between two long ranked lists. *Statistical Analysis and Data Mining* **6**, 458–471.

Segerstedt, B. (1992). On ordinary ridge regression in generalized linear models. *Communications in Statistics-Theory and Methods* **21**(8), 2227–2246.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

Webber, William, Moffat, Alistair and Zobel, Justin. (2010, November). A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.* **28**(4), 20:1–20:38.

Algorithm 1 Sequential rank agreement algorithm for incomplete lists

procedure INCOMPLETE LIST RANK AGREEMENT

  Let B be the number of permutations to use

  for each $b \in B$ **do**

    for each list $l \in L$ **do**

      Permute the unassigned ranks, $\{d_l + 1, \ldots, P\}$, and assign them randomly to the items *not* found in the list, i.e., $\Lambda^{\complement}_l = X \setminus \Lambda_l$. Combine the result with Λ_l to fill out the list.

    end for

    Let $\text{sra}(b)$ be the sequential rank agreement computed from the filled out lists.

  end for

  Return element-wise averages across all B permutations of $\text{sra}(b)$.

end procedure

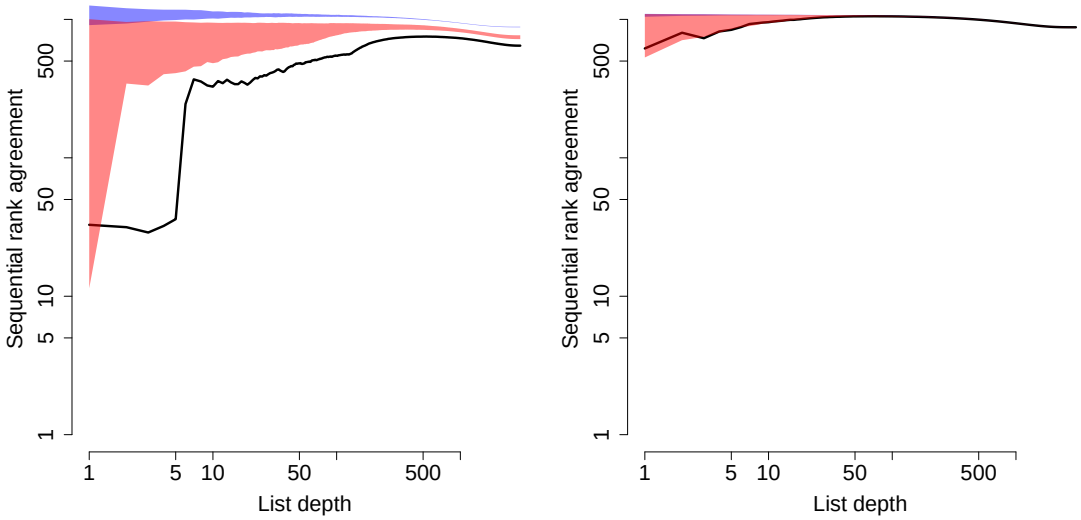

Fig. 1: Left panel: Sequential rank agreement for 4 different analysis methods applied to the 3051 genes in the Golub data (black line). Right panel: Corresponding sequential rank agreement for the same data but where only the top 20 ranked items are available and the rank of the remaining items are not available. The blue and red areas correspond to the independent and randomized reference hypothesis areas, respectively. Note that both the x and y axes are shown on the log scale to "zoom in" on the top of the lists.

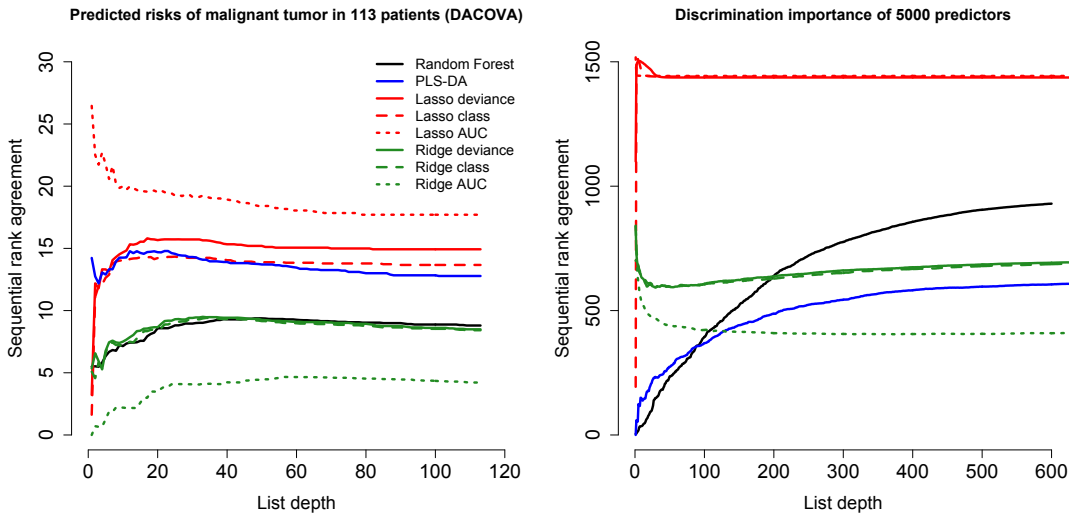

Fig. 2: Left panel: Sequential rank agreement of 1000 rankings of the predicted risks of malignant tumor. For each method the different rankings were obtained by first training models in 1000 random sub-samples of the MALOVA data and then predicting the risk of malignant tumor in the 113 DACOVA patients. Right panel: Sequential rank agreement of 1000 rankings of the 5000 predictors. The rankings were obtained from the same 1000 trained models.

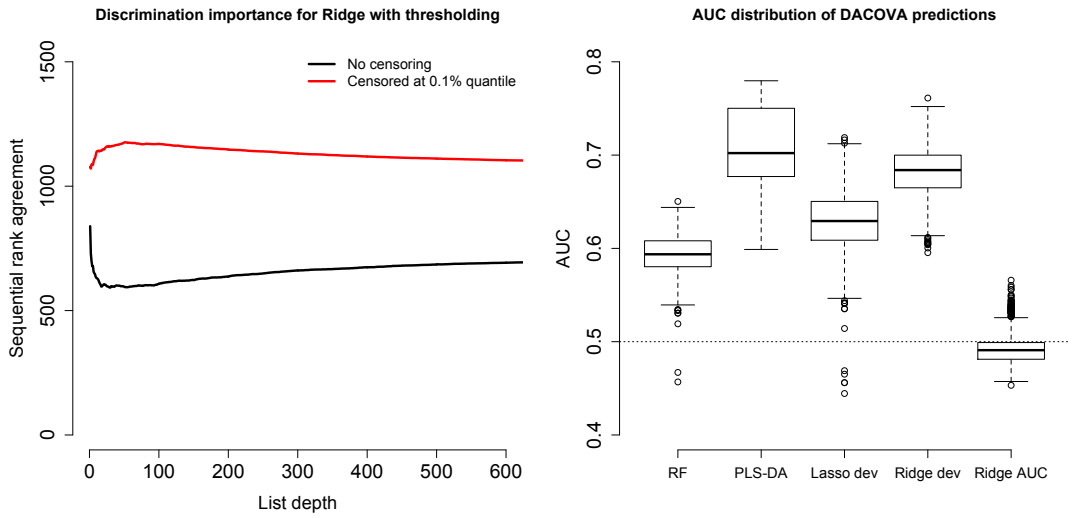

Fig. 3: Left panel: Sequential rank agreement for Ridge regression obtained by artificially censoring predictor ranks when their absolute coefficient values are lower than the 0.1% quantile. Right panel: Box plots of AUC values across the 1000 sub-samples with respect to the known class labels of the DACOVA data.

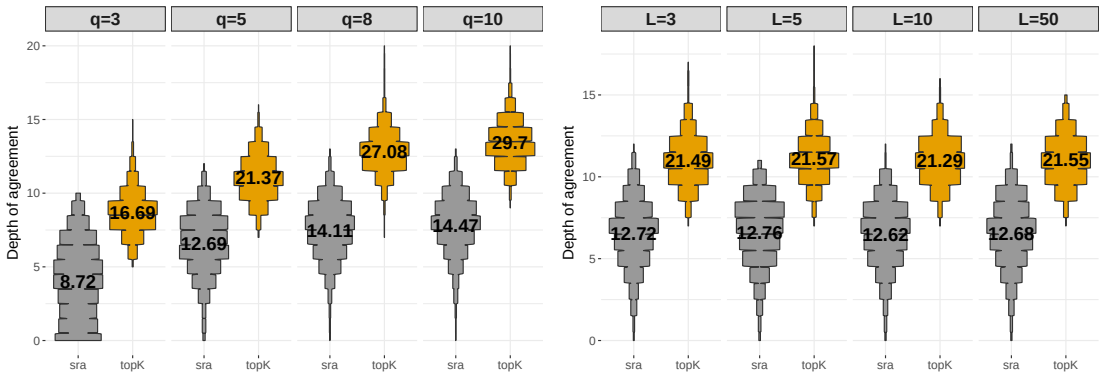

Fig. 4: Left panel: Simulation study showing distribution of estimated rank agreements for sra and topK for varying thresholds and fixed number of lists $L = 8$. The bold numbers are the average number of unique predictors included in the set where the lists agree. Right panel: Simulation results for varying number of lists and with fixed threshold of $q = 5$.

Table 1: Example set of ranked lists. (a) shows the ranked lists of items for each of three lists, (b) presents the ranks obtained by each item in each of the three lists and (c) shows the cumulative set of items up to a given depth in the three lists when $\varepsilon = 0$ (i.e., an item is added to $S(d)$ whenever it appears in at least one list).

(a)

Rank	R_1^{-1}	R_2^{-1}	R_3^{-1}
1	A	A	B
2	B	C	A
3	C	D	E
4	D	B	C
5	E	E	D

(b)

Item	R_1	R_2	R_3
A	1	1	2
B	2	4	1
C	3	2	4
D	4	3	5
E	5	5	3

(c)

Depth	S_d
1	{A, B}
2	{A, B, C}
3	{A, B, C, D, E}
4	{A, B, C, D, E}
5	{A, B, C, D, E}

Table 2: List of ranked results from the Golub data. Numbers indicate the predictor/gene for the given ranking and method. Only the top 10 ranks are shown in the table. The ranked lists appear to agree that genes 2124 and 829 are among the most interesting while the highest ranked gene from MIC, gene 378, is not found in the top 10 for two of the other methods.

Ranking	Welsh's t	LogReg	ElasticNet	MIC
1	2124	2124	829	378
2	896	896	2198	829
3	2600	829	2124	896
4	766	394	808	1037
5	829	766	1665	2124
6	2851	2670	1920	808
7	703	2939	1389	108
8	2386	2386	1767	515
9	2645	1834	1042	2670
10	2002	378	2600	2600