Algebraic stability indicators for ranked lists in molecular profiling

Giuseppe Jurman & MPBA

Centre for Bioinformation Sciences
Mathematical Science Institute
Australian National University
Canberra, 30 August 2007
Outline

1 Introduction

2 A set-theoretical solution

3 A group-theoretical solution

4 Application to high-throughput data

5 Odds ‘n’ ends

6 Bibliography
INTRODUCTION

A SET-THEORETICAL SOLUTION

A GROUP-THEORETICAL SOLUTION

APPLICATION TO HIGH-THROUGHPUT DATA

ODDS ‘N’ ENDS

BIBLIOGRAPHY
Outline

1 Introduction
2 A set-theoretical solution
3 A group-theoretical solution
4 Application to high-throughput data
5 Odds ‘n’ ends
6 Bibliography

G. Jurman & MPBA

Biomarkers’ lists stability
INTRODUCTION

A SET-THEORETICAL SOLUTION

A GROUP-THEORETICAL SOLUTION

APPLICATION TO HIGH-THROUGHPUT DATA

Odds ‘n’ ends

Bibliography
1. Introduction
2. A set-theoretical solution
3. A group-theoretical solution
4. Application to high-throughput data
5. Odds ‘n’ ends

Bibliography
1 INTRODUCTION
2 A SET-THEORETICAL SOLUTION
3 A GROUP-THEORETICAL SOLUTION
4 APPLICATION TO HIGH-THROUGHPUT DATA
5 ODDS ‘N’ ENDS
6 BIBLIOGRAPHY
**The Problem**

\[ M(n \times p, \mathbb{R}) \]

\[ n \approx 10^3 \] samples

\[ p \approx 10^5 \] genes

A result of a classification/ranking procedure is a list of genes, ranked according to their relevance for the classifier.

A complete validation experiment involves running several instances of the classification/ranking process, thus a whole set of lists is generated.

Feature ranking projects data from *gene expressions* to *gene lists*

**The Key Question**

How is it possible to derive information on the expression data by using the produced set of gene lists?

G. Jurman & MPBA

Biomarkers’ lists stability
THE PROBLEM

A result of a classification/ranking procedure is a list of genes, ranked according to their relevance for the classifier.

A complete validation experiment involves running several instances of the classification/ranking process, thus a whole set of lists is generated.

Feature ranking projects data from gene expressions to gene lists.

THE KEY QUESTION

How is it possible to derive information on the expression data by using the produced set of gene lists?
A result of a classification/ranking procedure is a list of genes, ranked according to their relevance for the classifier.

A complete validation experiment involves running several instances of the classification/ranking process, thus a whole set of lists is generated.

Feature ranking projects data from gene expressions to gene lists.

The key question: How is it possible to derive information on the expression data by using the produced set of gene lists?
THE PROBLEM

A result of a classification/ranking procedure is a list of genes, ranked according to their relevance for the classifier.

A complete validation experiment involves running several instances of the classification/ranking process, thus a whole set of lists is generated.

Feature ranking projects data from *gene expressions* to *gene lists*.

THE KEY QUESTION

How is it possible to derive information on the expression data by using the produced set of gene lists?
The methodology

A dataset of $n$ samples described by a set $\mathcal{F}$ of $p$ features indexed by integers from 1 to $p$: $\mathcal{F} = \{F_j\}_{j=1}^p$, with $n << p$.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>$F_j$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarray</td>
<td>gene expression</td>
<td>up to 50000</td>
</tr>
<tr>
<td>SNP</td>
<td>DNA CNVs</td>
<td>up to 500000</td>
</tr>
<tr>
<td>Proteomics</td>
<td>mass spectrometry intensities</td>
<td>$50 \leq p \leq 1000$</td>
</tr>
</tbody>
</table>

In classification and feature ranking tasks, methodology has to be carefully designed, not to incur in overfitting phenomena such as the selection bias.
The methodology

A dataset of $n$ samples described by a set $\mathcal{F}$ of $p$ features indexed by integers from 1 to $p$: $\mathcal{F} = \{F_j\}_{j=1}^p$, with $n << p$.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>$F_j$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarray</td>
<td>gene expression</td>
<td>up to 50000</td>
</tr>
<tr>
<td>SNP</td>
<td>DNA CNVs</td>
<td>up to 500000</td>
</tr>
<tr>
<td>Proteomics</td>
<td>mass spectrometry intensities</td>
<td>$50 \leq p \leq 1000$</td>
</tr>
</tbody>
</table>

In classification and feature ranking tasks, methodology has to be carefully designed, not to incur in overfitting phenomena such as the selection bias.

Error curve for the Colon cancer dataset (left) with true labels (blue) and random labels (black).
**The methodology**

A dataset of $n$ samples described by a set $\mathcal{F}$ of $p$ features indexed by integers from 1 to $p$: $\mathcal{F} = \{F_j\}_{j=1}^{p}$, with $n << p$.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>$F_j$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microarray</td>
<td>gene expression</td>
<td>up to 50000</td>
</tr>
<tr>
<td>SNP</td>
<td>DNA CNVs</td>
<td>up to 500000</td>
</tr>
<tr>
<td>Proteomics</td>
<td>mass spectrometry intensities</td>
<td>$50 \leq p \leq 1000$</td>
</tr>
</tbody>
</table>

In classification and feature ranking tasks, methodology has to be carefully designed, not to incur in overfitting phenomena such as the **selection bias**.

Error curve for the two synthetic datasets (right) $f_{1000} - 5000$, with 1000 discriminant features (black) and $f_0 - 5000$ (blue) with no discriminant features.
The methodology (ctd.)

Solution: separating the data used for feature selection from those used for evaluating accuracy.

How?
Adopting a resampling schema, where \( B \) different splits of data into training (for feature selection) and test (for error estimate) are realized.
Accuracy is evaluated by averaging over all the runs (Average Test Error, ATE)

At each run \( i=1 \ldots B \), a ranking process sorts the features according to their importance in building the \( i \)-th classifier and an ordered list \( L_i \) is produced.

The set of lists

\[ \mathcal{L} = \{L_i\}_{i=1}^B \] is the set of all lists and, for each list, \( L_i^k \) is its top-\( k \) list. i.e. the sublist consisting of the first ranked \( k \) elements from \( L_i \).
**The methodology (ctd.)**

Solution: separating the data used for feature selection from those used for evaluating accuracy.

How?
Adopting a **resampling** schema, where $B$ different splits of data into training (for feature selection) and test (for error estimate) are realized.

Accuracy is evaluated by averaging over all the runs (Average Test Error, ATE).

At each run $i=1 \ldots B$, a ranking process sorts the features according to their importance in building the $i$-th classifier and an ordered list $L_i$ is produced.

**The set of lists**

$\mathcal{L} = \{L_i\}_{i=1}^B$ is the set of all lists and, for each list, $L_i^k$ is its top-$k$ list. i.e. the sublist consisting of the first ranked $k$ elements from $L_i$. 
OUTLINE

1 INTRODUCTION
   • The setup
   • The group-theoretical tools

2 A SET-THEORETICAL SOLUTION

3 A GROUP-THEORETICAL SOLUTION

4 APPLICATION TO HIGH-THROUGHPUT DATA

5 ODDS ‘N’ ENDS

6 BIBLIOGRAPHY
A group is a set together with an associative operation which admits an identity element and such that every element has an inverse.

A permutation on a set $\Omega_n = \{s_1, s_2, \ldots, s_n\}$ of $n$ elements is a bijection

$$\tau : \Omega_n \to \Omega_n$$

$$s_i \mapsto \tau(s_i) = s_j$$

The number of different permutations on set of $n$ elements is $n!$.

The set of all permutations on a set $\Omega_n$ (usually $\Omega_n = \mathbb{N} \cap [1, n]$) becomes a group $S_n$, called permutation group or symmetric group when endowed with the operation

$$\sigma \tau : \Omega_n \to \Omega_n$$

$$s_i \mapsto \sigma(\tau(s_i)) = s_k$$
Permutation group theory

**Group**

A group is a set together with an associative operation which admits an identity element and such that every element has an inverse.

**Permutations**

A permutation on a set \( \Omega_n = \{s_1, s_2, \ldots, s_n\} \) of \( n \) elements is a bijection

\[
\tau : \Omega_n \rightarrow \Omega_n
\]

\[
s_i \mapsto \tau(s_i) = s_j
\]

The number of different permutations on set of \( n \) elements is \( n! \).

**Symmetric group**

The set of all permutations on a set \( \Omega_n \) (usually \( \Omega_n = \mathbb{N} \cap [1, n] \)) becomes a group \( S_n \), called permutation group or symmetric group when endowed with the operation

\[
\sigma \tau : \Omega_n \rightarrow \Omega_n
\]

\[
s_i \mapsto \sigma(\tau(s_i)) = s_k
\]
**Permutation Group Theory**

**Group**

A group is a set together with an associative operation which admits an identity element and such that every element has an inverse.

**Permutations**

A permutation on a set $\Omega_n = \{s_1, s_2, \ldots, s_n\}$ of $n$ elements is a bijection

$$\tau : \Omega_n \rightarrow \Omega_n$$

$$s_i \mapsto \tau(s_i) = s_j$$

The number of different permutations on set of $n$ elements is $n!$.

**Symmetric Group**

The set of all permutations on a set $\Omega_n$ (usually $\Omega_n = \mathbb{N} \cap [1, n]$) becomes a group $S_n$, called permutation group or symmetric group when endowed with the operation

$$\sigma \tau : \Omega_n \rightarrow \Omega_n$$

$$s_i \mapsto \sigma(\tau(s_i)) = s_k$$
List representation

E.g.: The symmetric group $S_3$ consists of

\[
\begin{align*}
\{1 \to 1, 2 \to 2, 3 \to 3\} & \quad \{1 \to 2, 2 \to 1, 3 \to 3\} \\
\{1 \to 3, 2 \to 2, 3 \to 1\} & \quad \{1 \to 1, 2 \to 3, 3 \to 2\} \\
\{1 \to 2, 2 \to 3, 3 \to 1\} & \quad \{1 \to 3, 2 \to 1, 3 \to 2\}
\end{align*}
\]

Representation

Labelling the $n$ genes involved in a problem by $\mathbb{N} \cap [1, n]$, every ranked list $L_t$ can be identified as a permutation $\tau_t \in S_n$, the image $\tau_t(i)$ of the $i$-th gene being its ranking inside the list $L_t$.

\[
\begin{array}{ccc}
\text{Gene} & \Omega_n & \tau_t(\Omega_n) \\
\text{AFFX-MurIL2_at} & \to & 1 \to 13347 \\
\ldots & \ldots & \ldots \\
\text{99764_at} & \to & n \to 175
\end{array}
\]

Thus $\tau$ can be written as

\[
\tau = (\tau(1), \ldots, \tau(n)) = (13347, \ldots, 175).
\]
**List representation**

*E.g.*: The symmetric group $S_3$ consists of

\[
\begin{align*}
\{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3\} & \quad \{1 \rightarrow 2, 2 \rightarrow 1, 3 \rightarrow 3\} \\
\{1 \rightarrow 3, 2 \rightarrow 2, 3 \rightarrow 1\} & \quad \{1 \rightarrow 1, 2 \rightarrow 3, 3 \rightarrow 2\} \\
\{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 1\} & \quad \{1 \rightarrow 3, 2 \rightarrow 1, 3 \rightarrow 2\}
\end{align*}
\]

Labelling the $n$ genes involved in a problem by $\mathbb{N} \cap [1, n]$, every ranked list $L_t$ can be identified as a permutation $\tau_t \in S_n$, the image $\tau_t(i)$ of the $i$-th gene being its ranking inside the list $L_t$.

<table>
<thead>
<tr>
<th>Gene</th>
<th>$\Omega_n$</th>
<th>$\tau_t(\Omega_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-MurIL2_at</td>
<td>$\rightarrow 1$</td>
<td>$\rightarrow 13347$</td>
</tr>
<tr>
<td>[ \ldots ]</td>
<td>[ \ldots ]</td>
<td>[ \ldots ]</td>
</tr>
<tr>
<td>99764_at</td>
<td>$\rightarrow n$</td>
<td>$\rightarrow 175$</td>
</tr>
</tbody>
</table>

Thus $\tau$ can be written as $\tau = (\tau(1), \ldots, \tau(n)) = (13347, \ldots, 175)$. 
**List Representation**

*E.g.:* The symmetric group $S_3$ consists of

\[
\begin{align*}
\{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3\} & \quad \{1 \rightarrow 2, 2 \rightarrow 1, 3 \rightarrow 3\} \\
\{1 \rightarrow 3, 2 \rightarrow 2, 3 \rightarrow 1\} & \quad \{1 \rightarrow 1, 2 \rightarrow 3, 3 \rightarrow 2\} \\
\{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 1\} & \quad \{1 \rightarrow 3, 2 \rightarrow 1, 3 \rightarrow 2\}
\end{align*}
\]

**Representation**

Labelling the $n$ genes involved in a problem by $\mathbb{N} \cap [1, n]$, every ranked list $L_t$ can be identified as a permutation $\tau_t \in S_n$, the image $\tau_t(i)$ of the $i$-th gene being its ranking inside the list $L_t$.

<table>
<thead>
<tr>
<th>Gene</th>
<th>$\Omega_n$</th>
<th>$\tau_t(\Omega_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-MurIL2_at</td>
<td>$\rightarrow 1$</td>
<td>$\rightarrow 13347$</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>99764_at</td>
<td>$\rightarrow n$</td>
<td>$\rightarrow 175$</td>
</tr>
</tbody>
</table>

Thus $\tau$ can be written as $\tau = (\tau(1), \ldots, \tau(n)) = (13347, \ldots, 175)$. 
List representation

*E.g.:* The symmetric group $S_3$ consists of

\[
\begin{align*}
\{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3\} & \quad \{1 \rightarrow 2, 2 \rightarrow 1, 3 \rightarrow 3\} \\
\{1 \rightarrow 3, 2 \rightarrow 2, 3 \rightarrow 1\} & \quad \{1 \rightarrow 1, 2 \rightarrow 3, 3 \rightarrow 2\} \\
\{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 1\} & \quad \{1 \rightarrow 3, 2 \rightarrow 1, 3 \rightarrow 2\}
\end{align*}
\]

**Representation**

Labelling the $n$ genes involved in a problem by $\mathbb{N} \cap [1, n]$, every ranked list $L_t$ can be identified as a permutation $\tau_t \in S_n$, the image $\tau_t(i)$ of the $i$-th gene being its ranking inside the list $L_t$.

<table>
<thead>
<tr>
<th>Gene</th>
<th>$\Omega_n$</th>
<th>$\tau_t(\Omega_n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFFX-MurIL2_at</td>
<td>$1$</td>
<td>$13347$</td>
</tr>
<tr>
<td>\ldots</td>
<td>\ldots</td>
<td>\ldots</td>
</tr>
<tr>
<td>99764_at</td>
<td>$n$</td>
<td>$175$</td>
</tr>
</tbody>
</table>

Thus $\tau$ can be written as $\tau = (\tau(1), \ldots, \tau(n)) = (13347, \ldots, 175)$. 
Outline

1. Introduction
2. A set-theoretical solution
   - A first indicator
   - A second indicator
3. A group-theoretical solution
4. Application to high-throughput data
5. Odds ‘n’ ends
6. Bibliography
**The top-k extraction set of** $F_j$

$$E_k(j) = \{ i \in \{1 \ldots B\} : \tau_i(j) \leq k \},$$

is the set of the indices of runs ranking each feature $F_j$ in the top-$k$ sublist.

**Extraction number**

$$e_k(j) = \text{Card}(E_k(j))$$

**Average position number**

$$a_k(j) = \frac{1}{e_k(j)} \sum_{i \in E_k(j)} \tau_i(j).$$

e_k(j) and a_k(j) induce a ranking of the features. First stability criterion: high e_k and low a_k indicate features extracted often in top positions.
**Extraction and position number**

**The top-k extraction set of** $F_j$

$$E_k(j) = \{i \in \{1 \ldots B\}: \tau_i(j) \leq k\},$$

is the set of the indices of runs ranking each feature $F_j$ in the top-$k$ sublist.

**Extraction number**

$$e_k(j) = \text{Card}(E_k(j))$$

**Average Position number**

$$a_k(j) = \frac{1}{e_k(j)} \sum_{i \in E_k(j)} \tau_i(j).$$

$e_k(j)$ and $a_k(j)$ induce a ranking of the features.

First stability criterion: high $e_k$ and low $a_k$ indicate features extracted often in top positions.

**Bibliography**

G. Jurman & MPBA

**Biomarkers’ lists stability**
Sese’s Liver Cancer dataset: binary classification by SVM/E–RFE on the complete (213 samples) and on the shaved (15 outliers removed) instances, both described by 1992 genes ($B = 400$ runs; shaving improved accuracy).

<table>
<thead>
<tr>
<th></th>
<th>k = 20 - Complete</th>
<th></th>
<th>k = 20 - Shaved</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>GS201</td>
<td>260</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>GS1324</td>
<td>212</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>GS1686</td>
<td>204</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>GS3244</td>
<td>198</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>GS6094</td>
<td>194</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>GS11601</td>
<td>194</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>GS11954</td>
<td>187</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>GS3097</td>
<td>163</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>GS2375</td>
<td>152</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>GS10424</td>
<td>145</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. A set-theoretical solution
   - A first indicator
   - A second indicator
3. A group-theoretical solution
4. Application to high-throughput data
5. Odds ‘n’ ends
6. Bibliography
A disarray measure

The union number

\[ \Sigma_k(\mathcal{L}) = \text{Card} \bigcup_{i=1}^{B} \{ F_j \in L_i: \tau_i(j) \leq k, 1 \leq j \leq p \} \]

is the number of different features occurring in all the \( B \) top-\( k \) lists.

The union number indicator

\[ l_U(\mathcal{L}) = \{(k, \Sigma_k): 1 \leq k \leq p\} \]

The higher the values of \( \Sigma_k \), the less stable is \( \mathcal{L} \): \( l_U = \{(k, p): 1 \leq k \leq p\} \) is the no-information curve.

Intersection instead of union leads to keeping track of the number of genes common to all top-\( k \) sublists, less informative for small values of \( k \) (e.g.: all but one identical lists).
**A DISARRAY MEASURE**

**THE UNION NUMBER**

\[ \Sigma_k(\mathcal{L}) = \text{Card} \bigcup_{i=1}^{B} \{F_j \in L_i: \tau_i(j) \leq k, 1 \leq j \leq p\} . \]

is the number of different features occurring in all the \( B \) top-\( k \) lists.

**THE UNION NUMBER INDICATOR**

\[ I_U(\mathcal{L}) = \{(k, \Sigma_k): 1 \leq k \leq p\} . \]

The higher the values of \( \Sigma_k \), the less stable is \( \mathcal{L} \): \( I_U = \{(k, p): 1 \leq k \leq p\} \) is the no-information curve.

Intersection instead of union leads to keeping track of the number of genes common to all top-\( k \) sublists, less informative for small values of \( k \) (e.g.: all but one identical lists).
INTRODUCTION

A SET-THEORETICAL SOLUTION

A GROUP-THEORETICAL SOLUTION

Definition and properties
- Partial lists
- Choosing the distance
- Bounds and approximations
- Indicators
- Formalization
- Consistency

APPLICATION TO HIGH-THROUGHPUT DATA

ODDS 'N' ENDS

BIBLIOGRAPHY
A map $d$ defined on a permutation group $S$

$$d : S \times S \rightarrow \mathbb{R}$$

is a distance function or a metric if it satisfies the following axioms:

- **Non negativity** $d(\sigma, \tau) \geq 0 \quad \forall \sigma, \tau \in S$
- **Separation** $d(\sigma, \tau) = 0 \iff \sigma = \tau \quad \forall \sigma, \tau \in S$
- **Symmetry** $d(\sigma, \tau) = d(\tau, \sigma) \quad \forall \sigma, \tau \in S$
- **Triangular** $d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma) \quad \forall \sigma, \tau, \rho \in S$
- **Right invariance** $d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S$

The first three properties are quite natural for a distance function in all environments, while the fourth one might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes $d$ a pseudometric.
A map \( d \) defined on a permutation group \( S \)

\[ d : S \times S \rightarrow \mathbb{R} \]

is a distance function or a metric if it satisfies the following axioms:

- **Non negativity** \( d(\sigma, \tau) \geq 0 \) \( \forall \sigma, \tau \in S \)
- **Separation** \( d(\sigma, \tau) = 0 \Leftrightarrow \sigma = \tau \) \( \forall \sigma, \tau \in S \)
- **Symmetry** \( d(\sigma, \tau) = d(\tau, \sigma) \) \( \forall \sigma, \tau \in S \)
- **Triangular** \( d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma) \) \( \forall \sigma, \tau, \rho \in S \)
- **Right invariance** \( d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \) \( \forall \sigma, \tau, \rho \in S \)

The first three properties are quite natural for a distance function in all environments, while the fourth one might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes \( d \) a pseudometric.
Metrics on permutation groups

**Definition**

A map $d$ defined on a permutation group $S$

$$d : S \times S \rightarrow \mathbb{R}$$

is a distance function or a metric if it satisfies the following axioms:

**Non negativity**  $d(\sigma, \tau) \geq 0 \quad \forall \sigma, \tau \in S$

**Separation**  $d(\sigma, \tau) = 0 \iff \sigma = \tau \quad \forall \sigma, \tau \in S$

**Symmetry**  $d(\sigma, \tau) = d(\tau, \sigma) \quad \forall \sigma, \tau \in S$

**Triangular**  $d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma) \quad \forall \sigma, \tau, \rho \in S$

**Right invariance**  $d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S$

The first three properties are quite natural for a distance function in all environments, while the fourth one might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes $d$ a pseudometric.
A map $d$ defined on a permutation group $S$

$$d : S \times S \rightarrow \mathbb{R}$$

is a distance function or a metric if it satisfies the following axioms:

**Non negativity** $d(\sigma, \tau) \geq 0 \quad \forall \sigma, \tau \in S$

**Separation** $d(\sigma, \tau) = 0 \iff \sigma = \tau \quad \forall \sigma, \tau \in S$

**Symmetry** $d(\sigma, \tau) = d(\tau, \sigma) \quad \forall \sigma, \tau \in S$

**Triangular** $d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma) \quad \forall \sigma, \tau, \rho \in S$

**Right invariance** $d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S$

The first three properties are quite natural for a distance function in all environments, while the fourth one might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes $d$ a pseudometric.
A map $d$ defined on a permutation group $S$

$$d : S \times S \rightarrow \mathbb{R}$$

is a distance function or a metric if it satisfies the following axioms:

- **Non negativity** $d(\sigma, \tau) \geq 0 \quad \forall \sigma, \tau \in S$
- **Separation** $d(\sigma, \tau) = 0 \iff \sigma = \tau \quad \forall \sigma, \tau \in S$
- **Symmetry** $d(\sigma, \tau) = d(\tau, \sigma) \quad \forall \sigma, \tau \in S$
- **Triangular** $d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma) \quad \forall \sigma, \tau, \rho \in S$
- **Right invariance** $d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S$

The first three properties are quite natural for a distance function in all environments, while the fourth one might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes $d$ a pseudometric.
A map $d$ defined on a permutation group $S$

$$d : S \times S \rightarrow \mathbb{R}$$

is a distance function or a metric if it satisfies the following axioms:

**Non negativity**  $d(\sigma, \tau) \geq 0$  $\forall \sigma, \tau \in S$

**Separation**  $d(\sigma, \tau) = 0 \iff \sigma = \tau$  $\forall \sigma, \tau \in S$

**Symmetry**  $d(\sigma, \tau) = d(\tau, \sigma)$  $\forall \sigma, \tau \in S$

**Triangular**  $d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma)$  $\forall \sigma, \tau, \rho \in S$

**Right invariance**  $d(\sigma, \tau) = d(\sigma \rho, \tau \rho)$  $\forall \sigma, \tau, \rho \in S$

The first three properties are quite natural for a distance function in all environments, while the fourth one might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes $d$ a pseudometric.
A map $d$ defined on a permutation group $S$

$$d : S \times S \rightarrow \mathbb{R}$$

is a distance function or a metric if it satisfies the following axioms:

**Non Negativity** $d(\sigma, \tau) \geq 0 \quad \forall \sigma, \tau \in S$

**Separation** $d(\sigma, \tau) = 0 \iff \sigma = \tau \quad \forall \sigma, \tau \in S$

**Symmetry** $d(\sigma, \tau) = d(\tau, \sigma) \quad \forall \sigma, \tau \in S$

**Triangular** $d(\sigma, \tau) + d(\tau, \rho) > d(\rho, \sigma) \quad \forall \sigma, \tau, \rho \in S$

**Right Invariance** $d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S$

The first three properties are quite natural for a distance function in all environments, while the fourth one might sometimes be relaxed to some less strict requirements (e.g. polygonal inequalities) that makes $d$ a pseudometric.
The fifth property
\[ d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S \]
is an essential requirements for distance functions in permutation group spaces.

**An example**
- \( L_1 = \{A, C, D, B, E\} \) and \( L_2 = \{B, C, D, E, A\} \) via the identification \( A1, B2, C3, D4, E5 \) correspond to \( \tau_1 = (1, 4, 2, 3, 5), \tau_2 = (5, 1, 2, 3, 4) \).
- Apply the permutation \( \rho \) corresponding to the relabeling \( A2, B5, C4, D1, E3 \).
- Then \( \tau_1 \rho = (3, 1, 5, 2, 4) \) and \( \tau_2 \rho = (3, 5, 4, 2, 1) \).

Right-invariance corresponds to relabeling independence.

The Euclidean distance \[ d(\tau, \sigma) = \sqrt{\sum_{i=1}^{n}(\tau^{-1}(i) - \sigma^{-1}(i))^2} \] is not right-invariant.
The fifth property

\[ d(\sigma, \tau) = d(\sigma \rho, \tau \rho) \quad \forall \sigma, \tau, \rho \in S \]

is an essential requirements for distance functions in permutation group spaces.

**An example**

- \( L_1 = \{A, C, D, B, E\} \) and \( L_2 = \{B, C, D, E, A\} \) via the identification
  - \( A1, B2, C3, D4, E5 \) correspond to \( \tau_1 = (1, 4, 2, 3, 5) \), \( \tau_2 = (5, 1, 2, 3, 4) \).
- Apply the permutation \( \rho \) corresponding to the relabeling \( A2, B5, C4, D1, E3 \).
- Then \( \tau_1 \rho = (3, 1, 5, 2, 4) \) and \( \tau_2 \rho = (3, 5, 4, 2, 1) \).

Right-invariance corresponds to relabeling independence.

The Euclidean distance \( d(\tau, \sigma) = \sqrt{\sum_{i=1}^{n}(\tau^{-1}(i) - \sigma^{-1}(i))^2} \) is not right-invariant.
**Statistical distances**

**Spearman’s Footrule**

It is the $L_1$ distance between permutations: $F(\sigma, \tau) = \sum_{i=1}^{n} |\sigma(i) - \tau(i)|$

**Spearman’s Rho**

It is the $L_2$ distance between permutations: $R(\sigma, \tau) = \sqrt{\sum_{i=1}^{n} (\sigma(i) - \tau(i))^2}$

**Kendall’s Tau**

It is the minimum number of pairwise adjacent transpositions (swaps) needed to transform $\sigma$ into $\tau$: $K(\sigma, \tau) = \text{Card}\{(i, j) \in \Omega_n^2 : \sigma(i) < \sigma(j) \text{ and } \tau(i) > \tau(j)\}$

All these measures are metric.
STATISTICAL DISTANCES

**Spearman’s footrule**

It is the $L_1$ distance between permutations: 

$$F(\sigma, \tau) = \sum_{i=1}^{n} ||\sigma(i) - \tau(i)||$$

**Spearman’s rho**

It is the $L_2$ distance between permutations: 

$$R(\sigma, \tau) = \sqrt{\sum_{i=1}^{n} (\sigma(i) - \tau(i))^2}$$

**Kendall’s tau**

It is the minimum number of pairwise adjacent transpositions (swaps) needed to transform $\sigma$ into $\tau$: 

$$K(\sigma, \tau) = \text{Card}\{(i, j) \in \Omega_n^2: \sigma(i) < \sigma(j) \text{ and } \tau(i) > \tau(j)\}$$

All these measures are metric.
**Statistical distances**

**Spearman’s footrule**

It is the $L_1$ distance between permutations: $F(\sigma, \tau) = \sum_{i=1}^{n} \|\sigma(i) - \tau(i)\|$.

**Spearman’s rho**

It is the $L_2$ distance between permutations: $R(\sigma, \tau) = \sqrt{\sum_{i=1}^{n} (\sigma(i) - \tau(i))^2}$.

**Kendall’s tau**

It is the minimum number of pairwise adjacent transpositions (swaps) needed to transform $\sigma$ into $\tau$: $K(\sigma, \tau) = \text{Card}\{(i, j) \in \Omega_n^2 : \sigma(i) < \sigma(j) \text{ and } \tau(i) > \tau(j)\}$.

All these measures are metric.
**Statistical distances**

**Spearman’s Footrule**

It is the $L_1$ distance between permutations: 
\[ F(\sigma, \tau) = \sum_{i=1}^{n} \| \sigma(i) - \tau(i) \| \]

**Spearman’s rho**

It is the $L_2$ distance between permutations: 
\[ R(\sigma, \tau) = \sqrt{\sum_{i=1}^{n} (\sigma(i) - \tau(i))^2} \]

**Kendall’s tau**

It is the minimum number of pairwise adjacent transpositions (swaps) needed to transform $\sigma$ into $\tau$: 
\[ K(\sigma, \tau) = \text{Card}\{(i, j) \in \Omega_n^2 : \sigma(i) < \sigma(j) \text{ and } \tau(i) > \tau(j)\} \]

All these measures are metric.
Since most important genes are located in the upper part of the lists, it is also important to be able to compare those more important portions of the lists. Managing partial lists is much harder than global lists, since they do not necessarily involve all the same elements.

**Top k-list**

A top-$k$ list is the sublist including the elements ranking from position 1 to $k$ of the original list, which corresponds to induce a partial ordering on $k$ out of $n$ objects.

This has a natural counterpart into the group-theoretical framework:

**The right coset**

Given a group $G$, a subgroup $H$ and a element $g \in G$, the set $Hg = \{ hg : h \in H \}$ is called a right coset of $H$ in $G$. The right cosets of $H$ form a partition of $G$, because the relation $g_1 \simeq g_2 \iff Hg_1 = Hg_2 \iff g_1g_2^{-1} \in H$ is an equivalence.
Distances on top-$k$ lists

Since most important genes are located in the upper part of the lists, it is also important to be able to compare those more important portions of the lists. Managing partial lists is much harder than global lists, since they do not necessarily involve all the same elements.

**TOP $k$-LIST**

A top-$k$ list is the sublist including the elements ranking from position 1 to $k$ of the original list, which corresponds to induce a partial ordering on $k$ out of $n$ objects.

This has a natural counterpart into the group-theoretical framework:

**THE RIGHT COSET**

Given a group $G$, a subgroup $H$ and a element $g \in G$, the set $Hg = \{ hg : h \in H \}$ is called a right coset of $H$ in $G$. The right cosets of $H$ form a partition of $G$, because the relation $g_1 \approx g_2 \iff Hg_1 = Hg_2 \iff g_1 g_2^{-1} \in H$ is an equivalence.
Since most important genes are located in the upper part of the lists, it is also important to be able to compare those more important portions of the lists. Managing partial lists is much harder than global lists, since they do not necessarily involve all the same elements.

**Top k-list**

A top-k list is the sublist including the elements ranking from position 1 to k of the original list, which corresponds to induce a partial ordering on k out of n objects.

**The right coset**

Given a group G, a subgroup H and a element $g \in G$, the set $Hg = \{hg : h \in H\}$ is called a right coset of H in G. The right cosets of H form a partition of G, because the relation $g_1 \sim g_2 \iff Hg_1 = Hg_2 \iff g_1g_2^{-1} \in H$ is an equivalence.
Thus $\sigma, \tau \in S_n$ belong to the same right coset of $S_{n-k}$ if and only if $\sigma^{-1}(i) = \tau^{-1}(i)$ for $1 \leq i \leq k$, that is, if they induce the same partial ranking on the indices $1 \ldots k$.

**The Hausdorff metric**

Given a group $G$ and $K$ a subgroup of $G$, if $d$ is a right-invariant metric on $G$, then $d$ induces a right-invariant metric $d_H$ on $G/K$, called the Hausdorff metric

$$d_H(\tau K, \sigma K) = \max\{ \max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta), \max_{\alpha \in \tau K} \min_{\beta \in \sigma K} d(\alpha, \beta), \}$$

**Notes**

- The number $\min_{\alpha \in \tau K} d(\alpha, \beta)$ is just the ordinary distance from the fixed permutation $\beta$ to the set $\tau K$, thus $\max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta)$ is the ordinary distance from the set $\tau K$ to the furthest away permutation in $\sigma K$.

- Thus $d_H(\tau K, \sigma K)$ represents how much $\tau K$ needs to be expanded in order to include the whole of $\sigma K$ and vice versa.

- The apparently unnatural choice of the candidates for the induced distance is motivated by the requirement of keeping the distance right-invariant.
The Hausdorff topology

Thus $\sigma, \tau \in S_n$ belong to the same right coset of $S_{n-k}$ if and only if $\sigma^{-1}(i) = \tau^{-1}(i)$ for $1 \leq i \leq k$, that is, if they induce the same partial ranking on the indices $1 \ldots k$.

The Hausdorff metric

Given a group $G$ and $K$ a subgroup of $G$, if $d$ is a right-invariant metric on $G$, then $d$ induces a right-invariant metric $d_H$ on $G/K$, called the Hausdorff metric

$$d_H(\tau K, \sigma K) = \max\{ \max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta), \max_{\alpha \in \tau K} \min_{\beta \in \sigma K} d(\alpha, \beta), \}$$

Notes

- The number $\min_{\alpha \in \tau K} d(\alpha, \beta)$ is just the ordinary distance from the fixed permutation $\beta$ to the set $\tau K$, thus $\max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta)$ is the ordinary distance from the set $\tau K$ to the furthest away permutation in $\sigma K$.

- Thus $d_H(\tau K, \sigma K)$ represents how much $\tau K$ needs to be expanded in order to include the whole of $\sigma K$ and vice versa.

- The apparently unnatural choice of the candidates for the induced distance is motivated by the requirement of keeping the distance right-invariant.
Thus $\sigma, \tau \in S_n$ belong to the same right coset of $S_{n-k}$ if and only if $\sigma^{-1}(i) = \tau^{-1}(i)$ for $1 \leq i \leq k$, that is, if they induce the same partial ranking on the indices $1 \ldots k$.

**The Hausdorff Metric**

Given a group $G$ and $K$ a subgroup of $G$, if $d$ is a right-invariant metric on $G$, then $d$ induces a right-invariant metric $d_H$ on $G/K$, called the Hausdorff metric

$$d_H(\tau K, \sigma K) = \max\{ \max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta), \max_{\alpha \in \tau K} \min_{\beta \in \sigma K} d(\alpha, \beta) \},$$

**Notes**

- The number $\min_{\alpha \in \tau K} d(\alpha, \beta)$ is just the ordinary distance from the fixed permutation $\beta$ to the set $\tau K$, thus $\max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta)$ is the ordinary distance from the set $\tau K$ to the furthest away permutation in $\sigma K$.

- Thus $d_H(\tau K, \sigma K)$ represents how much $\tau K$ needs to be expanded in order to include the whole of $\sigma K$ and vice versa.

- The apparently unnatural choice of the candidates for the induced distance is motivated by the requirement of keeping the distance right-invariant.
Thus $\sigma, \tau \in S_n$ belong to the same right coset of $S_{n-k}$ if and only if $\sigma^{-1}(i) = \tau^{-1}(i)$ for $1 \leq i \leq k$, that is, if they induce the same partial ranking on the indices $1 \ldots k$.

The Hausdorff metric

Given a group $G$ and $K$ a subgroup of $G$, if $d$ is a right-invariant metric on $G$, then $d$ induces a right-invariant metric $d_H$ on $G/K$, called the Hausdorff metric

$$d_H(\tau K, \sigma K) = \max \{ \max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta), \max_{\alpha \in \tau K} \min_{\beta \in \sigma K} d(\alpha, \beta) \}$$

Notes

- The number $\min_{\alpha \in \tau K} d(\alpha, \beta)$ is just the ordinary distance from the fixed permutation $\beta$ to the set $\tau K$, thus $\max_{\beta \in \sigma K} \min_{\alpha \in \tau K} d(\alpha, \beta)$ is the ordinary distance from the set $\tau K$ to the furthest away permutation in $\sigma K$.
- Thus $d_H(\tau K, \sigma K)$ represents how much $\tau K$ needs to be expanded in order to include the whole of $\sigma K$ and vice versa.
- The apparently unnatural choice of the candidates for the induced distance is motivated by the requirement of keeping the distance right-invariant.
Other methods are known in literature to derive a distance on partial lists from a distance on global lists. For instance

**Distance with location parameter**

- Consider a global distance $D$
  - Consider two top-$k$ lists $\tau_m$ for $m = 1, 2$, involving respectively the subsets $D_{\tau_m}^{-1} = \tau_m^{-1}(\mathbb{N} \cap [1, k])$ of $\Omega_n$
  - Define the global list $\tau'_m(i) = \begin{cases} \tau_m(i) & \text{for } i \in D_{\tau_m}^{-1} \\ k + 1 & \text{otherwise} \end{cases}$
  - Define $D_H(\tau_1, \tau_2) = D(\tau'_1, \tau'_2)$.

Anyway, equivalence theorems and experimental results point out that $F^{(k+1)}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_1}^{-1} \cup D_{\tau_2}^{-1}} \|\tau'_1(i) - \tau'_2(i)\|$ the Spearman’s footrule with location parameter $k + 1$ is a good candidate as a distance for gene list equivalent to $F_H$.  

G. Jurman & MPBA

Biomarkers’ lists stability
Other methods are known in literature to derive a distance on partial lists from a distance on global lists. For instance

**Distance with location parameter**

- Consider a global distance $D$
- Consider two top-$k$ lists $\tau_m$ for $m = 1, 2$, involving respectively the subsets $D_{\tau_m} = \tau_m^{-1}(N \cap [1, k])$ of $\Omega_n$.
- Define the global list $\tau_m'(i) = \begin{cases} \tau_m(i) & \text{for } i \in D_{\tau_m}^{-1} \\ k + 1 & \text{otherwise} \end{cases}$
- Define $D_H(\tau_1, \tau_2) = D(\tau_1', \tau_2')$.

Anyway, equivalence theorems and experimental results point out that

$$F^{(k+1)}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_1}^{-1} \cup D_{\tau_2}^{-1}} \|\tau_1'(i) - \tau_2'(i)\|$$

the Spearman’s footrule with location parameter $k + 1$ is a good candidate as a distance for gene list equivalent to $F_H$. 

---

**EQUIVALENCE RESULTS**

G. Jurman & MPBA

**Biomarkers’ lists stability**
EQUIVALENCE RESULTS

Other methods are known in literature to derive a distance on partial lists from a distance on global lists. For instance

**Distance with location parameter**

- Consider a global distance $D$
- Consider two top-$k$ lists $\tau_m$ for $m = 1, 2$, involving respectively the subsets $D_{\tau_{m-1}} = \tau_{m-1}^{-1}(\mathbb{N} \cap [1, k])$ of $\Omega_n$;
- Define the global list $\tau'_m(i) = \begin{cases} \tau_m(i) & \text{for } i \in D_{\tau_{m-1}} \\ k+1 & \text{otherwise} \end{cases}$
- Define $D_H(\tau_1, \tau_2) = D(\tau'_1, \tau'_2)$.

Anyway, equivalence theorems and experimental results point out that

$$F^{(k+1)}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_{1}} \cup D_{\tau_{2}}} ||\tau'_1(i) - \tau'_2(i)||$$

the Spearman’s footrule with location parameter $k + 1$ is a good candidate as a distance for gene list equivalent to $F_H$. 
Other methods are known in literature to derive a distance on partial lists from a distance on global lists. For instance:

**Distance with location parameter**

- Consider a global distance $D$
- Consider two top-$k$ lists $\tau_m$ for $m = 1, 2$, involving respectively the subsets $D_{\tau_m^{-1}} = \tau_m^{-1}(\mathbb{N} \cap [1, k])$ of $\Omega_n$.
- Define the global list $\tau'_m(i) = \begin{cases} \tau_m(i) & \text{for } i \in D_{\tau_m^{-1}} \\ k + 1 & \text{otherwise} \end{cases}$
- Define $D_H(\tau_1, \tau_2) = D(\tau'_1, \tau'_2)$.

Anyway, equivalence theorems and experimental results point out that

$$F^{(k+1)}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_1^{-1}} \cup D_{\tau_2^{-1}}} \|\tau'_1(i) - \tau'_2(i)\|$$

the Spearman’s footrule with location parameter $k + 1$ is a good candidate as a distance for gene list equivalent to $F_H$. 

---

**G. Jurman & MPBA**

**Biomarkers’ lists stability**
Other methods are known in literature to derive a distance on partial lists from a distance on global lists. For instance

**Distance with location parameter**

- Consider a global distance $D$
- Consider two top-$k$ lists $\tau_m$ for $m = 1, 2$, involving respectively the subsets $D_{\tau_m}^{-1} = \tau_m^{-1}(\mathbb{N} \cap [1, k])$ of $\Omega_n$;
- Define the global list $\tau'_m(i) = \begin{cases} \tau_m(i) & \text{for } i \in D_{\tau_m}^{-1} \\ k + 1 & \text{otherwise} \end{cases}$
- Define $D_H(\tau_1, \tau_2) = D(\tau'_1, \tau'_2)$.

Anyway, equivalence theorems and experimental results point out that

$$F^{(k+1)}(\tau_1, \tau_2) = \sum_{\substack{i \in D_{\tau_1}^{-1} \cup D_{\tau_2}^{-1} \\ i \in D_{\tau_1}^{-1} \cup D_{\tau_2}^{-1}}} \|\tau'_1(i) - \tau'_2(i)\|$$

the Spearman’s footrule with location parameter $k + 1$ is a good candidate as a distance for gene list equivalent to $F_H$. 

G. Jurman & MPBA

Biomarkers’ lists stability
Other methods are known in literature to derive a distance on partial lists from a distance on global lists. For instance

**Distance with location parameter**

- Consider a global distance $D$
- Consider two top-$k$ lists $\tau_m$ for $m = 1, 2$, involving respectively the subsets $D_\tau_{\tau_m^{-1}} = \tau_m^{-1}(\mathbb{N} \cap [1, k])$ of $\Omega_n$;
- Define the global list $\tau_m'(i) = \begin{cases} \tau_m(i) & \text{for } i \in D_\tau_{\tau_m^{-1}} \\ k + 1 & \text{otherwise} \end{cases}$
- Define $D_H(\tau_1, \tau_2) = D(\tau_1', \tau_2')$.

Anyway, equivalence theorems and experimental results point out that

$$F^{(k+1)}(\tau_1, \tau_2) = \sum_{i \in D_{\tau_1^{-1}} \cup D_{\tau_2^{-1}}} ||\tau_1'(i) - \tau_2'(i)||$$

the Spearman’s footrule with location parameter $k + 1$ is a good candidate as a distance for gene list equivalent to $F_H$. 

G. Jurman & MPBA

Biomarkers’ lists stability
Moreover, the existence of feature modules, for instance highly correlated genes (e.g. clones), or genes belonging to the same pathway must be taken into account.

Classifiers as SVM tend to swap the relative importance of correlated features during different ranking processes, and those swaps should be less penalized than movements between uncorrelated genes.

**Modules-aware distances**

Let $M = \{g_1, \ldots, g_m\} \subset \Omega_p = \{1, \ldots, p\}$ be a module consisting of $m$ features. For each permutation $\tau \in \mathcal{L}$, define the permutation $\eta_M \in S_m$ by the property $\eta_M(i) < \eta_M(j)$ if $\tau(g_{\eta_M(i)}) < \tau(g_{\eta_M(j)})$.

Then define the new permutation $\tau_M$ as $\tau_M(g_i) = \tau(g_{\eta_M(i)})$, leaving $\tau_M|_{\Omega_p \setminus M} = \tau$ out of $M$.

The distance is then computed on $\mathcal{L}_M = \{\tau_M : \tau \in \mathcal{L}\}$.
Moreover, the existence of feature modules, for instance highly correlated genes (e.g. clones), or genes belonging to the same pathway must be taken into account.

Classifiers as SVM tend to swap the relative importance of correlated features during different ranking processes, and those swaps should be less penalized than movements between uncorrelated genes.

**Modules-aware distances**

Let $M = \{g_1, \ldots, g_m\} \subset \Omega_p = \{1, \ldots, p\}$ be a module consisting of $m$ features. For each permutation $\tau \in \mathcal{L}$, define the permutation $\eta_M \in S_m$ by the property

$$\eta_M(i) < \eta_M(j) \text{ if } \tau(g_{\eta_M(i)}) < \tau(g_{\eta_M(j)})\,.$$  

Then define the new permutation $\tau_M$ as $\tau_M(g_i) = \tau(g_{\eta_M(i)})$, leaving $\tau_M|_{\Omega_p \setminus M} = \tau$ out of $M$.

The distance is then computed on $\mathcal{L}_M = \{\tau_M : \tau \in \mathcal{L}\}$. 
Moreover, the existence of feature modules, for instance highly correlated genes (e.g. clones), or genes belonging to the same pathway must be taken into account.

Classifiers as SVM tend to swap the relative importance of correlated features during different ranking processes, and those swaps should be less penalized than movements between uncorrelated genes.

**Modules-aware distances**

Let \( M = \{g_1, \ldots, g_m\} \subset \Omega_p = \{1, \ldots, p\} \) be a module consisting of \( m \) features. For each permutation \( \tau \in \mathcal{L} \), define the permutation \( \eta_M \in S_m \) by the property

\[
\eta_M(i) < \eta_M(j) \text{ if } \tau(g_{\eta_M(i)}) < \tau(g_{\eta_M(j)}).
\]

Then define the new permutation \( \tau_M \) as \( \tau_M(g_i) = \tau(g_{\eta_M(i)}) \), leaving \( \tau_M|_{\Omega_p \setminus M} = \tau \) out of \( M \).

The distance is then computed on \( \mathcal{L}_M = \{\tau_M: \tau \in \mathcal{L}\} \).
INTRODUCTION

A SET-THEORETICAL SOLUTION

A GROUP-THEORETICAL SOLUTION

Definition and properties
Partial lists
Choosing the distance
Bounds and approximations
Indicators
Formalization
Consistency

APPLICATION TO HIGH-THROUGHPUT DATA

ODDS ’N’ ENDS

BIBLIOGRAPHY
The Canberra distance

A key fact in gene lists is that variations in the bottom part of the lists are much less relevant than differences in the top part.

Thus a weight factor is needed to take care of such requirement.

A natural candidate to substitute Spearman’s footrule is its weighted version.

Thus, in general,

\[ Ca^{(k+1)}(\tau, \sigma) = \sum_{i=1}^{p} \frac{\min\{\tau(i), k+1\} - \min\{\sigma(i), k+1\}}{\min\{\tau(i), k+1\} + \min\{\sigma(i), k+1\}}, \]

with \( Ca^{(p+1)} = Ca \).
A key fact in gene lists is that variations in the bottom part of the lists are much less relevant than differences in the top part.

Thus a weight factor is needed to take care of such requirement.

A natural candidate to substitute Spearman’s footrule is its weighted version.

Thus, in general,

\[
Ca^{(k+1)}(\tau, \sigma) = \sum_{i=1}^{n} \frac{|\tau(i) - \sigma(i)|}{\tau(i) + \sigma(i)}
\]

\[
Ca^{(k+1)} defined by the above procedure.
\]

with \(Ca^{(p+1)}=Ca\).
THE CANBERRA DISTANCE

- A key fact in gene lists is that variations in the bottom part of the lists are much less relevant than differences in the top part.
- Thus a weight factor is needed to take care of such requirement.
- A natural candidate to substitute Spearman’s footrule is its weighted version.

\[
\text{Ca}^{(k+1)}(\tau, \sigma) = \sum_{i=1}^{n} \frac{|\tau(i) - \sigma(i)|}{\tau(i) + \sigma(i)}
\]

Thus, in general,

\[
\text{Ca}^{(k+1)}(\tau, \sigma) = \sum_{i=1}^{p} \frac{|\min\{\tau(i), k+1\} - \min\{\sigma(i), k+1\}|}{\min\{\tau(i), k+1\} + \min\{\sigma(i), k+1\}},
\]

with \(\text{Ca}^{(p+1)} = \text{Ca}\).
INTRODUCTION

A SET-THEORETICAL SOLUTION

A GROUP-THEORETICAL SOLUTION
Definition and properties
Partial lists
Choosing the distance
Bounds and approximations
Indicators
Formalization
Consistency

APPLICATION TO HIGH-THROUGHPUT DATA

ODDS ’N’ ENDS

BIBLIOGRAPHY
The expected value

Define $A_s = \begin{cases} \frac{1}{2} H^s_2 & \text{for even } s \\ H^s - H^{s-1}_2 & \text{for odd } s \end{cases}$, $(H_s = \sum_{j=1}^{s} (1/j)$ is the $s$-th harmonic number).

**Theorem**

The expected (average) value of the Canberra metric on $S_n$ is

$$E\{Ca^{(k+1)}\}_{S_p} = \frac{2}{p} \sum_{t=1}^{k-1} t(A_{2k-t} - A_t) + 2p - k \left(2(k+1)(H_{2k+1} - H_{k+1}) - k\right),$$

which can be approximated up to terms $o(1)$ as

$$\hat{E}\{Ca^{(k+1)}\}_{S_p} = \frac{(k+1)(2p-k)}{p} \log(4) - \frac{2kp + 3p - k - k^2}{p}.$$ 

For complete lists $\hat{E}\{Ca\}_{S_p} = (p+1) \log(4) - (p+2)$.

For top-$k$ lists, up to terms converging to zero with $p$, the approximation $\hat{E}\{Ca^{(k+1)}\}_{S_p}$ as a function of $k$ and fixed $p$ is a parabola.

For $p \approx 10^4$, $|E\{Ca^{(k+1)}\}_{S_p} - \hat{E}\{Ca^{(k+1)}\}_{S_p}| \approx 10^{-3}$. 
INTRODUCTION

A SET-THEORETICAL SOLUTION

A GROUP-THEORETICAL SOLUTION

APPLICATION TO HIGH-THROUGHPUT DATA

ODDS ‘N’ ENDS

BIBLIOGRAPHY
Given a set of lists $\mathcal{L} = \{L_t\}_{t=1}^b$ of $p$ genes and an integer computation of all mutual top-$k$ distances leads to the construction of a symmetric distance matrix $M_k \in \mathcal{M}(b \times b, \mathbb{R}^+)$.

Then the corresponding histogram can be built.

Often the histogram can be approximated by a gaussian distribution (exp).

It is thus possible to use the mean (and the variance if gaussian-like) of the matrix $M_k$ as measures of the intrinsic distance of the top-$k$ lists of the set $\mathcal{L}$. 
Distance matrix and distribution

- Given a set of lists \( \mathcal{L} = \{ L_t \}_{t=1}^{b} \) of \( p \) genes and an integer computation of all mutual top-\( k \) distances leads to the construction of a symmetric distance matrix
  \( M_k \in \mathcal{M}(b \times b, \mathbb{R}^+) \).
- Then the corresponding histogram can be built
- Often the histogram can be approximated by a gaussian distribution (exp)

\[ n = 100, p = 500, b = 100, SVM-RFE, D = C \]

Indicators

It is thus possible to use the mean (and the variance if gaussian-like) of the matrix \( M_k \) as measures of the intrinsic distance of the top-\( k \) lists of the set \( \mathcal{L} \).
Given a set of lists \( \mathcal{L} = \{ L_t \}_{t=1}^b \) of \( p \) genes and an integer computation of all mutual top-\( k \) distances leads to the construction of a symmetric distance matrix \( M_k \in \mathcal{M}(b \times b, \mathbb{R}^+) \).

Then the corresponding histogram can be built.

Often the histogram can be approximated by a gaussian distribution (exp).
Stability

Intuitively, the more mutually different the lists, the more unstable the problem (either because of the data or because of the employed classification procedure).

Stability curve

- Given a set of list $\mathcal{L}$ on $p$ genes, decide a sequence $K$ of relevant gene subset dimensions.
- For each $k \in K$, compute the mean and the variance of $M_k$ (possibly normalized by $k$).
- Plot those values versus the sequence $K$: this may be computationally heavy, depending on $p$ and on the number of lists.

The above procedure is independent from the source generating the lists and from the dimensions of the gene expression data.
Stability

Intuitively, the more mutually different the lists, the more unstable the problem (either because of the data or because of the employed classification procedure).

Stability curve

- Given a set of list $\mathcal{L}$ on $p$ genes, decide a sequence $K$ of relevant gene subset dimensions.
- For each $k \in K$, compute the mean and the variance of $M_k$ (possibly normalized by $k$).
- Plot those values versus the sequence $K$: this may be computationally heavy, depending on $p$ and on the number of lists.

The above procedure is independent from the source generating the lists and from the dimensions of the gene expression data.
Intuitively, the more mutually different the lists, the more unstable the problem (either because of the data or because of the employed classification procedure).

**Stability Curve**

- Given a set of list \( \mathcal{L} \) on \( p \) genes, decide a sequence \( K \) of relevant gene subset dimensions.
- For each \( k \in K \), compute the mean and the variance of \( M_k \) (possibly normalized by \( k \)).
- Plot those values versus the sequence \( K \): this may be computationally heavy, depending on \( p \) and on the number of lists.

The above procedure is independent from the source generating the lists and from the dimensions of the gene expression data.
INTRODUCTION

A SET-THEORETICAL SOLUTION

A GROUP-THEORETICAL SOLUTION

Definition and properties
Partial lists
Choosing the distance
Bounds and approximations
Indicators
Formalization
Consistency

APPLICATION TO HIGH-THROUGHPUT DATA

ODDS 'N' ENDS

BIBLIOGRAPHY
Let $\mu_k = \frac{2}{B(B-1)} \sum_{2 \leq i < j \leq B} (M_k)_{ij}$ be the mean of all the $\frac{B(B-1)}{2}$ non-trivial values of the distance matrix $M_k \in \mathcal{M}(B \times B, \mathbb{R}^+)$. 

**The mean list distance indicator** is the sequence

$$I_{D_{\mu}}(\mathcal{L}) = \{(k, \mu_k) : 1 \leq k \leq p\}.$$ 

**The no-information curve** is the sequence

$$I_{D_{\mu}}(S_p) = \{(k, E\{C(a^{(k+1)}_{S_p})\}) : 1 \leq k \leq p\}$$

associated to the group $S_p$ of all possible dual lists with $p$ features.
Let $\mu_k = \frac{2}{B(B-1)} \sum_{2 \leq i < j \leq B} (M_k)_{ij}$ be the mean of all the $\frac{B(B-1)}{2}$ non-trivial values of the distance matrix $M_k \in \mathcal{M}(B \times B, \mathbb{R}^+)$. 

**The mean list distance indicator**

is the sequence

$$I_{D_{\mu}}(\mathcal{L}) = \{(k, \mu_k): 1 \leq k \leq p\}.$$ 

**The no-information curve**

is the sequence

$$I_{D_{\mu}}(S_p) = \{(k, E\{Ca^{(k+1)}\}_{S_p}): 1 \leq k \leq p\}$$

associated to the group $S_p$ of all possible dual lists with $p$ features.
An indication of the stability of a list set \( \mathcal{L} \) is given by comparing \( I_{D,\mu}(\mathcal{L}) \) with \( I_{D,\mu}(S_p) \), the latter representing the situation of maximal instability.

The defined indicators are independent from the particular classifier and feature ranking algorithms.

The pointwise ratios of \( I_U \) and \( I_{D,\mu} \) with their no-information curves (\( \Sigma_k = p \) and \( I_{D,\mu}(S_p) \), respectively) are independent from \( p \) and \( k \).

\[
\hat{I}_U = \left\{ \left( k, \frac{\Sigma_k}{p} \right) : 1 \leq k \leq p \right\}
\]
\[
\hat{I}_{D,\mu} = \left\{ \left( k, \frac{\mu_k}{E\{Ca(k+1)\}_{S_p}} \right) : 1 \leq k \leq p \right\} .
\]

The resulting normalized indicators \( \hat{I}_U \) and \( \hat{I}_{D,\mu} \) can be then used to compare experiments on different datasets.
An indication of the stability of a list set $\mathcal{L}$ is given by comparing $\hat{I}_{D,\mu}(\mathcal{L})$ with $\hat{I}_{D,\mu}(S_p)$, the latter representing the situation of maximal instability.

The defined indicators are independent from the particular classifier and feature ranking algorithms.

The pointwise ratios of $I_U$ and $I_{D,\mu}$ with their no-information curves ($\Sigma_k = p$ and $I_{D,\mu}(S_p)$, respectively) are independent from $p$ and $k$.

$$\hat{I}_U = \left\{ \left( k, \frac{\Sigma_k}{p} \right) : 1 \leq k \leq p \right\}$$

$$\hat{I}_{D,\mu} = \left\{ \left( k, \frac{\mu_k}{E\{Ca(k+1)\}_{S_p}} \right) : 1 \leq k \leq p \right\}$$

The resulting normalized indicators $\hat{I}_U$ and $\hat{I}_{D,\mu}$ can be then used to compare experiments on different datasets.
An indication of the stability of a list set $\mathcal{L}$ is given by comparing $I_{D,\mu}(\mathcal{L})$ with $I_{D,\mu}(\mathcal{S}_p)$, the latter representing the situation of maximal instability.

The defined indicators are independent from the particular classifier and feature ranking algorithms.

The pointwise ratios of $I_U$ and $I_{D,\mu}$ with their no-information curves ($\Sigma_k = p$ and $I_{D,\mu}(\mathcal{S}_p)$, respectively) are independent from $p$ and $k$.

\[
\hat{I}_U = \left\{ \left( k, \frac{\Sigma_k}{p} \right) : 1 \leq k \leq p \right\}
\]

\[
\hat{I}_{D,\mu} = \left\{ \left( k, \frac{\mu_k}{E\{C_{a(k+1)}\}|_{\mathcal{S}_p}} \right) : 1 \leq k \leq p \right\}.
\]

The resulting normalized indicators $\hat{I}_U$ and $\hat{I}_{D,\mu}$ can be then used to compare experiments on different datasets.
An indication of the stability of a list set $\mathcal{L}$ is given by comparing $I_{D\mu}(\mathcal{L})$ with $I_{D\mu}(S_p)$, the latter representing the situation of maximal instability.

The defined indicators are independent from the particular classifier and feature ranking algorithms.

The pointwise ratios of $I_U$ and $I_{D\mu}$ with their no-information curves ($\Sigma_k = p$ and $I_{D\mu}(S_p)$, respectively) are independent from $p$ and $k$.

\[
\hat{I}_U = \left\{ \left( k, \frac{\Sigma_k}{p} \right) : 1 \leq k \leq p \right\}
\]

\[
\hat{I}_{D\mu} = \left\{ \left( k, \frac{\mu_k}{E\{Ca(k+1)\}_S} \right) : 1 \leq k \leq p \right\}.
\]

The resulting normalized indicators $\hat{I}_U$ and $\hat{I}_{D\mu}$ can be then used to compare experiments on different datasets.
**No-information cases**

**Random lists**

Mean distance indicator and no-information curve, with 95% bootstrap Student’s t confidence intervals for 10 sets of 1000 random lists on 100 genes ($\text{Card}(S_{100}) = 100! \approx 10^{158}$)

**Random labels**

Huntington’s Disease dataset: $n = 14 + 17$, $p = 5186$, Affy U133A

True labels (TL, solid line), 1000 × 2 (R1, dashed) and 100 × 20 random. (R2, dotted).
**No-information cases**

**Random lists**

Mean distance indicator and no-information curve, with 95% bootstrap Student’s t confidence intervals for 10 sets of 1000 random lists on 100 genes ($\text{Card}(S_{100}) = 100! \approx 10^{158}$)

**Random labels**

Huntington’s Disease dataset: $n = 14 + 17$, $p = 5186$, Affy U133A
True labels (TL, solid line), $1000 \times 2$ (R1, dashed) and $100 \times 20$ random. (R2, dotted).
**Predictive profiling on synthetic data**

**Synthetic datasets** \( fX/Y: 50 + 50 \) binary labelled samples described by \( X \) discriminant features from \( \mathcal{N}(\mu, \sigma) (\mu \text{ depending on the sample label}) \) and by \( Y - X \) random features from the uniform \( \mathcal{U}[a, b] \).

**Complete validation**: \( B = 400 \) LinSVM/RFE exps, with correlation modules M1 (>0.8) and M2 (>0.7). On both datasets, less than five features are sufficient to reach perfect classification.

Stability indicators in profiling synthetic data (solid lines: \( f10/100 \); dotted lines: \( f30/100 \)) and in presence of feature modules.
Classifiers comparison

- Classifiers: Linear SVM / Terminated ramp SVM
- Feature ranking methods: (E−)RFE / 1–RFE
- Datasets: Microarray Breast Cancer - 183 cases described by 22215 gene expression / Proteomic Ovarian Cancer - 160 samples described by 123 mass spec peaks.

Breast Cancer

![Breast Cancer Test Error Graph](image)

Average Test Error
LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)

Ovarian Cancer

![Ovarian Cancer Test Error Graph](image)

Average Test Error
LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)
Classifiers: Linear SVM / Terminated ramp SVM

Feature ranking methods: (E–)RFE / 1–RFE

Datasets: Microarray Breast Cancer - 183 cases described by 22215 gene expression / Proteomic Ovarian Cancer - 160 samples described by 123 mass spec peaks.

**Classifiers Comparison**

### Breast Cancer

- Union number $I_U$
- LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted)

### Ovarian Cancer

- Union number $I_U$
- TRSVM/1RFE (T1: solid)

---

**G. Jurman & MPBA**

**Biomarkers’ lists stability**
Classifiers comparison

- Classifiers: Linear SVM / Terminated ramp SVM
- Feature ranking methods: (E–)RFE / 1–RFE
- Datasets: Microarray Breast Cancer - 183 cases described by 22215 gene expression / Proteomic Ovarian Cancer - 160 samples described by 123 mass spec peaks.

Breast Cancer

Mean distance $I_{D_{\mu}}$
LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)

Ovarian Cancer

Mean distance $I_{D_{\mu}}$
Classifiers comparison

- Classifiers: Linear SVM / Terminated ramp SVM
- Feature ranking methods: (E–)RFE / 1–RFE
- Datasets: Microarray Breast Cancer - 183 cases described by 22215 gene expression / Proteomic Ovarian Cancer - 160 samples described by 123 mass spec peaks.

Breast Cancer

\[ I_{D_{\mu}} \text{ top-25} \]
LinearSVM/RFE (LR: dashed) / LinearSVM/1RFE (L1: dotted) / TRSVM/1RFE (T1: solid)

Ovarian Cancer

\[ I_{D_{\mu}} \text{ top-25} \]
Accuracy and stability

- Although they share a common trend in many cases, accuracy and stability are independent measures.
- Thus we can analyze a dataset in the accuracy vs. stability space.
- Left-down direction indicates better performance.
- This diagnostic plot allows the comparison of different datasets, different profiling methods (classifiers/feature ranking algorithm) and different models.

ATE vs. $I_{D_{\mu}}$ for different profiling methods (LE,L1,T1) and cancer datasets (Breast, Ovarian); each point corresponds to a feature subset size, indicated for extremal models.
Outline

1. Introduction
2. A set-theoretical solution
3. A group-theoretical solution
4. Application to high-throughput data
5. Odds 'n' ends
   - Optimal list
   - Enrichment
6. Bibliography
How is it possible to derive a single list from an entire set such that it is representative of the whole set?

**Borda count**

For each $k$, consider as the ranking criterion the extraction number $e_k(j)$ in decreasing order, and, when ties occur, the position number $a_k(j)$ in increasing order. The criterion defines a dual list $\tau^k_O$, which correspond to a list of features called the *optimal top-k list* of $\mathcal{L}$.

When $k = p$, the complete lists are considered and trivially it is $e_p(j) = B$ for all $F_j$; thus $\tau_O = \tau^p_O$ is determined only by the $a_p(j)$ average values. Other choices are determining the list having minimal average distance w.r.t. the whole set, or the list yielding the best accuracy curve.
**Optimal list**

**Deriving a single list**

How is it possible to derive a single list from an entire set such that it is representative of the whole set?

**Borda count**

For each $k$, consider as the ranking criterion the extraction number $e_k(j)$ in decreasing order, and, when ties occur, the position number $a_k(j)$ in increasing order. The criterion defines a dual list $\tau^k_O$, which correspond to a list of features called the *optimal top-k list* of $L$.

When $k = p$, the complete lists are considered and trivially it is $e_p(j) = B$ for all $F_j$; thus $\tau_O = \tau^p_O$ is determined only by the $a_p(j)$ average values.

Other choices are determining the list having minimal average distance w.r.t. the whole set, or the list yielding the best accuracy curve.
**Optimal list**

**Deriving a single list**

How is it possible to derive a single list from an entire set such that it is representative of the whole set?

**Borda count**

For each $k$, consider as the ranking criterion the extraction number $e_k(j)$ in decreasing order, and, when ties occur, the position number $a_k(j)$ in increasing order. The criterion defines a dual list $\tau^k_O$, which correspond to a list of features called the *optimal top-k list* of $L$.

When $k = p$, the complete lists are considered and trivially it is $e_p(j) = B$ for all $F_j$; thus $\tau_O = \tau^p_O$ is determined only by the $a_p(j)$ average values. Other choices are determining the list having minimal average distance w.r.t. the whole set, or the list yielding the best accuracy curve.
INTRODUCTION

A set-theoretical solution

A group-theoretical solution

Application to high-throughput data

ODDS ‘N’ ENDS

Bibliography

Optimal list

Enrichment

G. Jurman & MPBA

Biomarkers’ lists stability
If a library of functionally correlated gene sets (e.g. pathways) is available, the stability indicator can be used to compare the stability of the original lists with the stability of the ranked gene sets produced by their enrichment.

A dishomogeneous set of lists may be shown to correspond to a much stabler lists of pathways.

Example on synthetic data with 100 lists of 100 genes and 124 pathways, where the lists are created by random permutations fixing the gene sets.
REFERENCES

[Borda81] Borda, J.
Mémoire sur les élections au scrutin.

[Critchlow85] Critchlow, D.
Metric methods for analyzing partially ranked data.

[Diaconis88] Diaconis, P.
Group representations in probability and statistics.

[Dwork01] Dwork, C., Kumar, R., Naor, M. & Sivakumar, D.
In WWW10.

[Fagin04] Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D. & Vee, E.
Comparing and Aggregating Rankings with Ties.

[Saari01] Saari, D.
Chaotic Elections! A Mathematician Looks at Voting.

Development and Evaluation of Therapeutically Relevant Predictive Classifiers Using Gene Expression Profiling.