Imprecise predictive selection based on low structure assumptions

Citation for published version (APA):

Coolen, F. P. A., & Laan, van der, P. (1998). *Imprecise predictive selection based on low structure assumptions*. (Memorandum COSOR; Vol. 9808). Technische Universiteit Eindhoven.

Document status and date:

Published: 01/01/1998

Document Version:

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Eindhoven University
of Technology

Department of Mathematics and Computing Sciences

Memorandum COSOR 98-08

Imprecise predictive selection based on low structure assumptions

F.P.A. Coolen
P. van der Laan

Eindhoven, May 1998
The Netherlands

Imprecise predictive selection based on low structure assumptions

F.P.A. Coolen

University of Durham

Department of Mathematical Sciences

Science Laboratories, South Road

Durham, DH1 3LE, England

Frank.Coolen@durham.ac.uk

P. van der Laan

Eindhoven University of Technology

Department of Mathematics and Computing Science

P.O. Box 513, 5600 MB Eindhoven, The Netherlands

pvdlaan@win.tue.nl

Abstract

New methods for statistical selection are presented, where the inferences have a nonparametric predictive nature. The basic assumption is Hill's $A_{(n)}$. Assuming that values of some random quantities from $k \geq 2$ independent sources are observed, $A_{(n)}$ provides a predictive probabilistic inference for further unknown random quantities from each source. Selection is mainly based on imprecise probabilities for the event that the next observation from the selected source will be greater than the next observation from every non-selected source. Both selection of a single source and selection of m ($m \leq k - 1$) sources are presented, considering the probability that all selected sources are better than all non-selected sources. A second related approach, using imprecise previsions with an interpretation as bounds for expected values for future observations, is presented briefly.

AMS Subject Classifications: 62-07, 62A25, 62C05, 62G99.

Keywords: $A_{(n)}$; Direct conditional probabilities; Exchangeability; Imprecise probabilities; Nonparametric methods; Predictive inference; Selection.

1 Introduction

Problems of comparing several treatments or sources are often formulated as which treatment or source can be considered to be the best. This paper presents a new approach to such problems, where comparison of different sources has a predictive nature in terms of probabilistic inferences for a next observation, based on data given in the form of past observations. (We prefer to use the term ‘source’ to the more common term ‘population’ in this paper, as the common view for observations from a single population is that these are independent and identically distributed, an assumption that we do not make for the method suggested in this paper.) The inferences at the heart of our method are nonparametric, and succeed in adding only few structural assumptions to the data. Methods designed specifically for the problem of choosing the best treatment or source, and to answer questions regarding the best in an adequate way, are known as ‘selection procedures’, our method is an alternative to existing selection procedures. Two important researchers in the field of selection are Bechhofer and Gupta. Bechhofer’s method [1] is known as the ‘Indifference Zone Approach’, the approach of Gupta [10] as ‘Subset Selection’. These approaches are well developed by now, and are mostly restricted to parametric models, with differences between sources often in terms of location parameters.

Bechhofer’s indifference zone approach assumes a minimal difference in the unknown parameter values, from which a value for the common sample size is developed as needed to meet certain requirements on the probability of correct selection. This method is useful for design of selection experiments, a recent overview is provided by Bechhofer, Santner and Goldsman [2]. One possible generalization is selection of a subset of the sources, such that all selected sources are better than all not selected sources, a problem that we address in section 3 of this paper.

Gupta’s subset selection approach aims at selection of a subset of all sources such that the probability that the best source is included exceeds a chosen value. The size of the subset is random and depends on the data observed. Subset selection can be used as a screening procedure, even when the ultimate goal of the experimenter is to choose the best, the subset selection approach can be applied to eliminate inferior sources. The size of the selected subset reflects the confidence of the experimenter in choosing the best. A small subset would mean that either the populations are not close together or that the sample sizes are large, or both. Moreover, the subset selection method can be used to analyze the data after the experiment has been realized, but is not useful to design experiments. Recently, Verheijen, Coolen and van der Laan [21] suggested a generalized approach for selection, which shares advantages with both indifference zone and subset selection, and has these two classical methods as special cases.

Nonparametric approaches to selection have also been considered, in particular within the framework of subset selection [11, 19]. These methods typically include sources in the selected subset if a sufficient statistic of the source is close to the optimal value of this sufficient statistic as observed for

any of the sources, where ‘optimal’ is related to the criterion chosen to describe what is considered to be the best source, and ‘close’ is determined using the distribution of the Wilcoxon statistic [17].

The approaches mentioned above all have a frequentist nature. Selection has also been studied from a Bayesian point of view, see Gupta and Yang [12] for a presentation of some Bayesian subset selection procedures and Miescke [20] for a discussion of Bayes sample designs for selection procedures. These two papers also give useful further references to the literature on Bayesian selection methods.

In this paper, we consider a nonparametric approach for selection, where comparison of different sources is via predictive inferences for future observations, based on past observations, while adding only few additional structural assumptions. Our method is based on Hill’s assumption $A_{(n)}$ [13], which gives a direct conditional probability [8] for a future observable random quantity, conditioned on observed values of related random quantities. In fact [15], this conditional probability can be used as a predictive posterior probability in a general Bayesian framework [4, 9]. With regard to the Bayesian approach it seems sufficient to remark that $A_{(n)}$ is a De Finetti coherent procedure [7, 14, 15, 16].

We are interested in one-dimensional finite real-valued random quantities from $k \geq 2$ different independent sources. In fact, we consider $n_j + 1$ random quantities from source j , $j = 1, \dots, k$, denoted by X_{j,i_j} , $i_j = 1, \dots, n_j, n_j + 1$. We will observe $X_{j,1}, \dots, X_{j,n_j}$, $j = 1, \dots, k$. In this paper, we assume that ties do not occur with positive probability. Our results are easily generalized to allow ties [15]. Denoting observations by lower case letters, we assume that $-\infty < x_{j,1} < x_{j,2} < \dots < x_{j,n_j} < \infty$ are the observations from source j . This assumption implies that the specific order in which the observations became available is considered to be irrelevant. For ease of notation, we define $x_{j,0} = -\infty$ and $x_{j,n_j+1} = \infty$, where care should be taken in particular to the fact that x_{j,n_j+1} is not an observed value of the random quantity of interest in our inferences, X_{j,n_j+1} . For each source, the open intervals created by the observations are denoted by

$$I_{j,i_j} = (x_{j,i_j-1}, x_{j,i_j}), \quad i_j = 1, \dots, n_j + 1.$$

Predictive inferences about the unknown quantity X_{j,n_j+1} are our goal. For a given source j , the assumption $A_{(n_j)}$ [13] is that

$$P(X_{j,n_j+1} \in I_{j,i_j}) = \frac{1}{n_j + 1}, \quad \text{for } i_j = 1, \dots, n_j + 1.$$

Although we develop our method for a single future observation per source only, results for several future observations are achievable [6], but care should be taken that such observations are not independent. It should be remarked that these assumptions $A_{(n)}$ do not assume anything else, and are clearly post-data assumptions related to (finite) exchangeability (see De Finetti [7, ch. 11]). Hill [14] gives a detailed presentation and discussion of $A_{(n)}$.

De Finetti's [7] representation theorem uses a similar setting to justify a Bayesian framework to learn about an underlying parameter, and a probability distribution for that parameter, but he relies on the assumption that indeed there is an infinite sequence of random quantities involved, whereas our interest here (as in many practical situations) is in a finite number of future observations, mostly restricting attention to a single next observation. Even more, the Bayesian approach, as justified by De Finetti's [7] important results, explicitly needs a specified prior distribution, and together with the conditional independence of future observations (conditional on an unknown parameter) this adds quite a bit more structure to the data than needed for our results. Our approach seems suitable if there is hardly any knowledge about the random quantities of interest, other than the first n_j observations per source j ($j = 1, \dots, k$), or, which may be more realistic, if one explicitly does not want to use any such information. This may occur, for example, if one wants to study the (often hidden) effects of additional structural assumptions underlying statistical models or methods. Inferences based on such restricted knowledge have been called low structure inferences [9, sect. 2.1.2] and black-box inferences [16].

The assumption $A_{(n)}$ is not sufficient to derive precise probability results for many problems of interest. However, it does provide bounds for probabilities and expectations, as presented in this paper, and this is essentially an application of De Finetti's 'fundamental theorem of probability' [7, sect. 3.10]. The bounds that we derive are imprecise probabilities and imprecise previsions (expectations) in the sense of Walley [22]. Adopting a subjective interpretation of probability and prevision, suppose that we are interested in an uncertain quantity A . In a subjective framework [22] that is a generalization of De Finetti's theory [7], your lower prevision $\underline{E}(A)$ for A is the supremum of all 'prices' you want to pay to get the uncertain quantity A , and your upper prevision $\overline{E}(A)$ for A is the infimum of all 'prices' for which you want to sell A (some unit of linear utility is needed for the prices, see Walley [22, sect. 2.2]). If one is not familiar with these concepts, $\underline{E}(A)$ and $\overline{E}(A)$ can be considered as lower and upper bounds for the expected value of A . Imprecise probabilities are simply imprecise previsions for events, so with A an indicator function that is 1 if the event occurs and 0 else. We denote a lower probability for A by $\underline{P}(A)$ and an upper probability for A by $\overline{P}(A)$.

The results in this paper also have a possible interpretation (and justification) as bounds of confidence statements in a nonparametric predictive frequentist setting, see Geisser [9, sect. 2.1.2] for more details about basic results related to this interpretation. Most statistical concepts exploit (finite) exchangeability or stronger assumptions, and agree with the $A_{(n)}$ -type assumptions before data are actually observed. Once data are observed, however, an assumed parametric model in effect introduces dependence between the numerical information from the data, and information about the ranks of possible future data related to the current data. This then undermines the validity of predictive statements as based on assumed exchangeability only. This dependence between numerical information and ranks is explicitly absent when using a nonparametric method, and is excellently shown by

our inferences based on $A_{(n)}$ alone. If good reasons for a certain (family of) parametric model(s) are present, indeed one may want to use such for inferences. However, if it is purely done for mathematical convenience or necessity, one should be careful as the model assumed cancels out the weak exchangeability assumption after the data are observed, and what is the justification of the model? Often, in case of few data many models seem justifiable, whereas in case of many data no simple model seems to fit anymore. In the mean time, $A_{(n)}$ -based inferences are entirely flexible, valid for few data, although high imprecision may be the fair price of only little information, and valid for many data as its asymptotics are obviously closely related to those for the empirical distribution function. The strength of the assumption $A_{(n)}$ can best be indicated by citing the final paragraph of Hill [14]: “Let me conclude by observing that $A_{(n)}$ is supported by all of the serious approaches to statistical inference. It is Bayesian, fiducial, and even a confidence/tolerance procedure. It is simple, coherent, and plausible. It can even be argued, I believe, that $A_{(n)}$ constitutes the fundamental solution to the problem of induction”.

In section 2 of this paper, imprecise probabilities are derived for selection of one source, in section 3 this is generalized to selection of m ($m \leq k-1$) sources, with interest only in the event that all selected sources are better than all non-selected sources. Section 4 briefly discusses imprecise previsions. In section 5 we present three examples to illustrate possible application of our results and to highlight some further aspects of this approach.

2 Selection of one source based on imprecise probabilities

For selection of one source out of k independent sources, we are interested in the probability that a specific X_{j,n_j+1} is the maximum of all next observations X_{t,n_t+1} , $t = 1, \dots, k$. In this section, we derive lower and upper bounds for this probability based on the assumptions $A_{(n_j)}$ per source j , these bounds are imprecise probabilities [22].

Comparing the next observation from each of k sources based on lower and upper probabilities is effectively a rank-based procedure, simplified by the following statistics for comparing source j with source l ($j \neq l$; $j, l = 1, \dots, k$)

$$c_{j,i_j}^l = \#\{x_{l,i_l} \mid x_{l,i_l} < x_{j,i_j}, \ i_l = 1, \dots, n_l\}, \ i_j = 1, \dots, n_j + 1.$$

Obviously, c_{j,i_j}^l is the number of observations from source l smaller than x_{j,i_j} , with $c_{j,n_j+1}^l = n_l$.

A lower bound for the probability for the event $X_{j,n_j+1} = \max_{t=1,\dots,k} X_{t,n_t+1}$, for given j , can be attained

$$\begin{aligned} P\left(X_{j,n_j+1} = \max_{t=1,\dots,k} X_{t,n_t+1}\right) &= P\left(\bigcap_{l=1, l \neq j}^k \{X_{j,n_j+1} > X_{l,n_l+1}\}\right) = \\ &\sum_{i_j=1}^{n_j+1} P\left(\bigcap_{l=1, l \neq j}^k \{X_{l,n_l+1} < X_{j,n_j+1}\} \mid X_{j,n_j+1} \in I_{j,i_j}\right) P(X_{j,n_j+1} \in I_{j,i_j}) \geq \\ &\frac{1}{n_j+1} \sum_{i_j=1}^{n_j+1} P\left(\bigcap_{l=1, l \neq j}^k \{X_{l,n_l+1} < x_{j,i_j-1}\}\right) = \frac{1}{n_j+1} \sum_{i_j=1}^{n_j} \left[\prod_{l=1, l \neq j}^k P(X_{l,n_l+1} < x_{j,i_j}) \right] \geq \\ &\frac{1}{n_j+1} \sum_{i_j=1}^{n_j} \left[\prod_{l=1, l \neq j}^k \frac{c_{j,i_j}^l}{n_l+1} \right] = \frac{1}{\prod_{t=1}^k (n_t+1)} \left[\sum_{i_j=1}^{n_j} \prod_{l=1, l \neq j}^k c_{j,i_j}^l \right]. \end{aligned}$$

In the derivation of this lower bound, we use the appropriate assumptions $A_{(n_j)}$ for all sources, the assumed independence of the k sources implying that X_{j,i_j} and X_{l,i_l} are independent for $j \neq l$, and the assumption that X_{l,n_l+1} is greater than $-\infty$ with probability 1. In effect, the lower bound is attained by putting the probability mass per interval, as assigned by the assumptions $A_{(n_j)}$ for all sources, at end points; for the selected source j at the left end point (infimum), and for all non-selected sources at the right end point (supremum). Without further assumptions, it is clear that no greater lower bound can be justified, and our derivation can be regarded as an application of De Finetti's 'fundamental theorem of probability' [7]. The above lower bound is a lower probability [22] and we use the notation (for given j)

$$\underline{P}\left(X_{j,n_j+1} = \max_{t=1,\dots,k} X_{t,n_t+1}\right) = \frac{1}{\prod_{t=1}^k (n_t+1)} \left[\sum_{i_j=1}^{n_j} \prod_{l=1, l \neq j}^k c_{j,i_j}^l \right].$$

If confusion is unlikely, we will also use the short notation

$$\underline{P}_j = \underline{P} \left(X_{j,n_j+1} = \max_{t=1,\dots,k} X_{t,n_t+1} \right),$$

and refer to it as the lower probability that source j is the best of all sources, where 'best source' is clearly to be interpreted in terms of a single next observation per source.

We attain an upper bound for the probability for the event $X_{j,n_j+1} = \max_{t=1,\dots,k} X_{t,n_t+1}$, for given j , as follows

$$\begin{aligned} P \left(X_{j,n_j+1} = \max_{t=1,\dots,k} X_{t,n_t+1} \right) &= P \left(\bigcap_{l=1, l \neq j}^k \{X_{j,n_j+1} > X_{l,n_l+1}\} \right) = \\ &\sum_{i_j=1}^{n_j+1} P \left(\bigcap_{l=1, l \neq j}^k \{X_{l,n_l+1} < X_{j,n_j+1}\} \mid X_{j,n_j+1} \in I_{j,i_j} \right) P \left(X_{j,n_j+1} \in I_{j,i_j} \right) \leq \\ &\frac{1}{n_j+1} \sum_{i_j=1}^{n_j} P \left(\bigcap_{l=1, l \neq j}^k \{X_{l,n_l+1} < x_{j,i_j}\} \right) + \frac{1}{n_j+1} = \\ &\frac{1}{n_j+1} \sum_{i_j=1}^{n_j} \left[\prod_{l=1, l \neq j}^k P \left(X_{l,n_l+1} < x_{j,i_j} \right) \right] + \frac{1}{n_j+1} \leq \\ &\frac{1}{n_j+1} \sum_{i_j=1}^{n_j} \left[\prod_{l=1, l \neq j}^k \frac{c_{j,i_j}^l + 1}{n_l + 1} \right] + \frac{1}{n_j+1} = \\ &\frac{1}{\prod_{t=1}^k (n_t + 1)} \left[\sum_{i_j=1}^{n_j} \prod_{l=1, l \neq j}^k (c_{j,i_j}^l + 1) \right] + \frac{1}{n_j+1}. \end{aligned}$$

This upper bound is derived similarly to the above lower bound. The term $\frac{1}{n_j+1}$ is a consequence of the assumption that X_{l,n_l+1} is smaller than ∞ with probability 1, and relates to the probability mass for source j in interval I_{j,n_j+1} . The upper bound is also attained by putting the probability mass per interval at end points; for the selected source j at the right end point (supremum), and for all non-selected sources at the left end point (infimum), and without further assumptions no smaller upper bound can be justified. This upper bound is an upper probability [22] for which we introduce the notation (for given j)

$$\overline{P}_j = \overline{P} \left(X_{j,n_j+1} = \max_{t=1,\dots,k} X_{t,n_t+1} \right) = \frac{1}{\prod_{t=1}^k (n_t + 1)} \left[\sum_{i_j=1}^{n_j} \prod_{l=1, l \neq j}^k (c_{j,i_j}^l + 1) \right] + \frac{1}{n_j+1},$$

and we refer to it as the upper probability that source j is the best of all sources.

These lower and upper probabilities are only based on the sufficient statistics n_j, c_{j,i_j}^l ($j \neq l; j, l = 1, \dots, k$), further information on location as contained in the observations x_{j,i_j} is not used. As such, this approach can be regarded as a fully predictive alternative to standard rank-based methods [17, 18], which are usually justified by theory of asymptotics. The method in section 4, using imprecise previsions, is not based on ranks and uses location information from the observations.

General bounds for these lower and upper probabilities $\underline{P}_j$ and $\bar{P}_j$, for given j , are easily derived by use of $0 \leq c_{j,i_j}^l \leq n_l$,

$$0 \leq \underline{P}_j \leq \prod_{t=1}^k \left(\frac{n_t}{n_t + 1} \right)$$

$$\frac{n_j}{\prod_{t=1}^k (n_t + 1)} + \frac{1}{n_j + 1} \leq \bar{P}_j \leq 1.$$

If $n_t = n$ for all $t = 1, \dots, k$, the upper bound for the lower probability is $\left(\frac{n}{n+1}\right)^k$ and the lower bound for the upper probability is $\frac{n}{(n+1)^k} + \frac{1}{n+1}$. Both theoretical lower bounds for these imprecise probabilities are actually achieved if all $c_{j,i_j}^l = 0$, so x_{j,n_j} is smaller than all observations from all other sources. Both theoretical upper bounds are actually achieved if all $c_{j,i_j}^l = n_l$, so $x_{j,1}$ is greater than all observations from all other sources. It is clear from the construction of our lower and upper probabilities that $\underline{P}_j \leq \bar{P}_j$. We cannot use the results of this section to determine sample sizes, but when assuming that $n_t = n$ for all $t = 1, \dots, k$, we can easily derive a lower bound for the necessary number of observations n per population, needed to get a lower probability $\underline{P}_j$, for one $j \in \{1, \dots, k\}$, of at least a certain required value, say $\tilde{p} \in (0, 1)$. From the upper bound for this lower probability we easily derive that

$$\underline{P} \left(X_{j,n_j+1} = \max_{t=1, \dots, k} X_{t,n_t+1} \right) \geq \tilde{p}$$

for one $j \in \{1, \dots, k\}$ can only be achieved for

$$n \geq \frac{\tilde{p}^{1/k}}{1 - \tilde{p}^{1/k}}.$$

This result only indicates that we can never get such a single lower probability $\underline{P}_j$ exceeding $\tilde{p}$ for smaller values of n , not even in the most favourable case for the selected source j where all observations for this source exceed all observations from each of the other sources.

3 Selection of m sources

In this section the results of section 2 are generalized. Instead of selecting a single source, we consider selection of a subset of the k independent sources, with the subset containing m ($2 \leq m \leq k - 1$) sources. Let $S = \{j_1, \dots, j_m\} \subset \{1, \dots, k\}$ be the subset of indices of the m selected sources, and let $NS = \{1, \dots, k\} \setminus S$ be the subset of indices of the $k - m$ non-selected sources. Using the appropriate assumptions $A_{(n_j)}$ for all sources, we derive lower and upper probabilities for the event

$$\min_{j \in S} X_{j, n_j+1} > \max_{l \in NS} X_{l, n_l+1}.$$

This event is true if the next observation for each selected source is greater than the next observation from all non-selected sources, we refer to this event as selection of the m best sources.

A lower bound for the probability for the event $\min_{j \in S} X_{j, n_j+1} > \max_{l \in NS} X_{l, n_l+1}$, for given S , is attained as follows

$$\begin{aligned} P\left(\min_{j \in S} X_{j, n_j+1} > \max_{l \in NS} X_{l, n_l+1}\right) &= P\left(\bigcap_{l \in NS} \left\{X_{l, n_l+1} < \min_{j \in S} X_{j, n_j+1}\right\}\right) = \\ &\sum_{i_{j_1}=1}^{n_{j_1}+1} \dots \sum_{i_{j_m}=1}^{n_{j_m}+1} P\left(\bigcap_{l \in NS} \left\{X_{l, n_l+1} < \min_{j \in S} X_{j, n_j+1}\right\} \mid X_{j_1, n_{j_1}+1} \in I_{j_1, i_{j_1}}, \dots, X_{j_m, n_{j_m}+1} \in I_{j_m, i_{j_m}}\right) \\ &\quad \times P\left(X_{j_1, n_{j_1}+1} \in I_{j_1, i_{j_1}}, \dots, X_{j_m, n_{j_m}+1} \in I_{j_m, i_{j_m}}\right) \geq \\ &\frac{1}{\prod_{j \in S} (n_j + 1)} \sum_{i_{j_1}=1}^{n_{j_1}+1} \dots \sum_{i_{j_m}=1}^{n_{j_m}+1} P\left(\bigcap_{l \in NS} \left\{X_{l, n_l+1} < \min_{s=1, \dots, m} x_{j_s, i_{j_s}-1}\right\}\right) = \\ &\frac{1}{\prod_{j \in S} (n_j + 1)} \sum_{i_{j_1}=1}^{n_{j_1}+1} \dots \sum_{i_{j_m}=1}^{n_{j_m}+1} \prod_{l \in NS} P\left(X_{l, n_l+1} < \min_{s=1, \dots, m} x_{j_s, i_{j_s}-1}\right) \geq \\ &\frac{1}{\prod_{j \in S} (n_j + 1)} \sum_{i_{j_1}=1}^{n_{j_1}} \dots \sum_{i_{j_m}=1}^{n_{j_m}} \left[\prod_{l \in NS} \frac{1}{n_l + 1} \min_{s=1, \dots, m} c_{j_s, i_{j_s}}^l \right] = \\ &\frac{1}{\prod_{t=1}^k (n_t + 1)} \sum_{i_{j_1}=1}^{n_{j_1}} \dots \sum_{i_{j_m}=1}^{n_{j_m}} \left[\prod_{l \in NS} \min_{s=1, \dots, m} c_{j_s, i_{j_s}}^l \right]. \end{aligned}$$

The detailed steps in the derivation of this lower bound are similar as in the derivation of the lower bound in section 2, with the second inequality a straightforward consequence of the definition of c_{j, i_j}^l . As before, this is the best achievable lower bound based on the assumptions $A_{(n_j)}$ per source alone, and it is attained by putting the probability mass per interval at end points; for all selected sources at the left end point, for all non-selected sources at the right end point. This lower probability will be

denoted by (for given S)

$$\underline{P}_S = \underline{P} \left(\min_{j \in S} X_{j,n_j+1} > \max_{l \in NS} X_{l,n_l+1} \right) = \frac{1}{\prod_{t=1}^k (n_t + 1)} \sum_{i_{j_1}=1}^{n_{j_1}} \cdots \sum_{i_{j_m}=1}^{n_{j_m}} \left[\prod_{l \in NS} \min_{s=1, \dots, m} c_{j_s, i_{j_s}}^l \right].$$

The corresponding upper bound for the event $\min_{j \in S} X_{j,n_j+1} > \max_{l \in NS} X_{l,n_l+1}$ is again derived similarly, also generalizing the upper bound derived in section 2, and is an upper probability that equals

$$\overline{P}_S = \overline{P} \left(\min_{j \in S} X_{j,n_j+1} > \max_{l \in NS} X_{l,n_l+1} \right) = \frac{1}{\prod_{t=1}^k (n_t + 1)} \sum_{i_{j_1}=1}^{n_{j_1}+1} \cdots \sum_{i_{j_m}=1}^{n_{j_m}+1} \left[\prod_{l \in NS} \left(\min_{s=1, \dots, m} c_{j_s, i_{j_s}}^l + 1 \right) \right].$$

In comparison with the upper bound derived in section 2, it could be remarked that the contribution to this upper probability from the term in this summation corresponding to $i_j = n_j + 1$ for all $j \in S$ equals

$$\frac{\prod_{l \in NS} (n_l + 1)}{\prod_{t=1}^k (n_t + 1)} = \frac{1}{\prod_{j \in S} (n_j + 1)}.$$

For $m = 1$ this indeed gives the term $\frac{1}{n_j+1}$ in the upper probability in section 2. As before, this term relates to the probability masses for all selected sources $j \in S$ in the respective intervals I_{j,n_j+1} .

The upper probability $\overline{P}_S$ is again the best achievable upper bound based on the assumptions $A_{(n_j)}$ per source alone, and it is attained by putting the probability mass per interval at end points; for all selected sources at the right end point, for all non-selected sources at the left end point.

Bounds for these imprecise probabilities are easily derived, in a similar way as for selection of a single population,

$$0 \leq \underline{P}_S \leq \prod_{t=1}^k \left(\frac{n_t}{n_t + 1} \right)$$

and

$$\frac{1}{\prod_{l \in NS} (n_l + 1)} + \frac{1}{\prod_{j \in S} (n_j + 1)} - \frac{1}{\prod_{t=1}^k (n_t + 1)} \leq \overline{P}_S \leq 1.$$

These bounds can again be used to derive minimum required sample sizes.

4 Selection based on imprecise previsions

Imprecise previsions [22] can be regarded as bounds for expected values for unknown quantities, generalizing De Finetti's concept of previsions [7]. Coolen [5] discusses imprecise previsions, based on the assumptions $A_{(n_j)}$ per source alone, for comparison of two populations. As these bounds for expected values for X_{j,n_j+1} , given observations $x_{j,1} < \dots < x_{j,n_j}$, are derived per source j individually, comparison of k sources based on such bounds is a straightforward generalization of the results for two populations. When dealing with expected values we must avoid possible probability mass in $-\infty$ or ∞ , as in effect used in deriving the imprecise probabilities before. We assume finite bounds for the observations per source, $\lambda_j < X_{j,i_j} < \rho_j$, $j = 1, \dots, k$; $i_j = 1, \dots, n_j + 1$. With observations $x_{j,1} < \dots < x_{j,n_j}$, it is obvious that $\lambda_j < x_{j,1}$ and $x_{j,n_j} < \rho_j$ should hold, and the assumption $A_{(n_j)}$ for source j is now slightly modified by redefining $I_{j,1} = (\lambda_j, x_{j,1})$ and $I_{j,n_j+1} = (x_{j,n_j}, \rho_j)$, leading to $P(\lambda_j < X_{j,n_j+1} < x_{j,1}) = P(x_{j,n_j} < X_{j,n_j+1} < \rho_j) = \frac{1}{n_j+1}$, and no probability mass to the left of λ_j or to the right of ρ_j .

Lower and upper bounds for the expected value of X_{j,n_j+1} are easily derived by putting all probability mass per interval to the left (infimum) for the lower bound (lower prevision $\underline{E}$), and to the right (supremum) for the upper bound (upper prevision $\overline{E}$)

$$\begin{aligned}\underline{E}(X_{j,n_j+1}) &= \frac{1}{n_j+1} \left(\lambda_j + \sum_{i_j=1}^{n_j} x_{j,i_j} \right) \\ \overline{E}(X_{j,n_j+1}) &= \frac{1}{n_j+1} \left(\rho_j + \sum_{i_j=1}^{n_j} x_{j,i_j} \right).\end{aligned}$$

With interpretations for lower and upper previsions as discussed in section 1, and aiming to maximise the next observation, we could say that one has a strong preference for source j over source l if $\underline{E}(X_{j,n_j+1}) > \overline{E}(X_{l,n_l+1})$. Similarly, one could say that a strong preference for source j over all other sources $l = 1, \dots, k$, $l \neq j$, exists when this inequality holds for all those l simultaneously. If all sample sizes and bounds on observations are equal, $n_j = n$, $\lambda_j = \lambda$, $\rho_j = \rho$ ($j = 1, \dots, k$), then this strong preference for source j holds if $\sum_{i_j=1}^n x_{j,i_j} - \sum_{i_l=1}^n x_{l,i_l} > \rho - \lambda$, for all $l \neq j$.

A weaker form of preference for source j over source l could be said to hold if both $\underline{E}(X_{j,n_j+1}) > \underline{E}(X_{l,n_l+1})$ and $\overline{E}(X_{j,n_j+1}) > \overline{E}(X_{l,n_l+1})$ hold simultaneously. With $n_j = n$, $\lambda_j = \lambda$, $\rho_j = \rho$, for all j , this reduces to weak preference for source j if $\sum_{i_j=1}^n x_{j,i_j} > \sum_{i_l=1}^n x_{l,i_l}$.

These imprecise previsions use the detailed information on locations from the observations, but may seem less appealing than the imprecise probabilities attained before. We suggest that both methods are used together to get a more complete picture of the situation. It is quite well possible, that one population is preferred over all other populations (either strongly or weakly) when using imprecise

previsions, while at the same time that population only has small lower and upper probabilities of giving the maximum value on the next observation. This is a situation that could appear, for example, if there are some outliers in the data.

5 Examples

Example 1

The data for this example (Table 1) are given in Bechhofer, Santner and Goldsman [2, p.84], and were earlier discussed by Box, Hunter and Hunter [3, p.277]. The data result from an experiment conducted to determine the weight loss, under a mechanical test, for seven types of cloth, and represent the measured weight loss, in tenths of a milligram, for each cloth following 1000 revolutions on a wear tester. By the nature of the experiment there is a block effect, but we do not consider this here. Development of our method to include such effects is an important and interesting topic for future research, for this example use of our method is justified if we consider this blocking information not to be available to us (even though we know that such an effect influences the data, if we would not know the actual information about this effect, the appropriate $A_{(n_j)}$ assumptions remain justifiable). Based on the data collected, suppose that we want to select the cloth with the *smallest* weight loss, and use our predictive method. In this paper, our method has been presented for selection related to a maximum, but the results could easily be transformed aiming at selection of a minimum. However, to stay with the presentation in this paper, we transform the data by replacing original weight loss w_{j,i_j} , representing the i_j -th ordered weight loss observation for cloth type j , $j = 1, \dots, 7$, $i_j = 1, \dots, 4$, by $x_{j,5-i_j} = 1000 - w_{j,i_j}$, and we aim at selection of the cloth type corresponding to the maximum $X_{j,5}$, $j = 1, \dots, 7$. Table 2 gives the transformed data suitable for the methods presented in this paper.

Cloth j	$w_{j,1}$	$w_{j,2}$	$w_{j,3}$	$w_{j,4}$
1	337	344	369	396
2	520	537	602	627
3	233	240	251	278
4	196	211	248	273
5	160	185	195	199
6	442	563	595	606
7	226	252	297	300

Table 1: Ordered weight losses (in 0.1 mg) per cloth (example 1)

Cloth j	$x_{j,1}$	$x_{j,2}$	$x_{j,3}$	$x_{j,4}$
1	604	631	656	663
2	373	398	463	480
3	722	749	760	767
4	727	752	789	804
5	801	805	815	840
6	394	405	437	558
7	700	703	748	774

Table 2: *Transformed data per cloth (example 1)*

Our predictive method compares the cloths via unknown next observations per cloth, $X_{j,5}$, $j = 1, \dots, 7$, using the assumptions $A_{(4)}$ to assign probabilities for these random quantities per cloth. For selection of a single cloth, or a subset of cloths, the sufficient statistics c_{j,i_j}^l are easily derived, for example, Table 3 gives the c_{j,i_j}^l values for $j = 4$, as would be needed to calculate the imprecise probabilities for cloth 4 to give the best next observation.

l	$c_{4,1}^l$	$c_{4,2}^l$	$c_{4,3}^l$	$c_{4,4}^l$
1	4	4	4	4
2	4	4	4	4
3	1	2	4	4
5	0	0	0	1
6	4	4	4	4
7	2	3	4	4

Table 3: *Values c_{4,i_4}^l (example 1)*

The imprecise probabilities, $\underline{P}_j$ and $\overline{P}_j$, that cloth j is the best source (using terminology from section 2) are given in Table 4.

j	$\underline{P}_j$	$\overline{P}_j$
1	0	0.2013
2	0	0.2001
3	0	0.2560
4	0.0131	0.3488
5	0.1966	0.9600
6	0	0.2002
7	0	0.2336

Table 4: *Imprecise probabilities for selection of one cloth (example 1)*

For these data, the theoretical upper bound for the lower probability is equal to 0.2097 and the lower bound for the upper probability is $\frac{15629}{78125} = 0.2001$. The upper probability for cloth 2, which is in fact $\overline{P}_2 = \frac{15636}{78125}$, is closest to this theoretical lower bound, and only slightly larger because the observations for cloth 6 were not all greater than those for cloth 2. The lower probability for cloth 5, $\underline{P}_5 = 0.1966$, is close to its theoretical upper bound, with the difference caused by the fact that one observation for cloth 4 exceeds the smallest observation for cloth 5. Obviously, in this example there is quite a strong suggestion that cloth 5 is the best one, and this is clearly shown by the results in Table 4, where the rather large differences between lower and upper probabilities reflect the fact that there are only few observations available. All the lower probabilities that equal 0 in Table 4 are caused by all the related observations being smaller than all observations for cloth 5. The upper probability $\overline{P}_5 = \frac{24}{25} = 0.96$ is less than 1 because, even when spreading the probability masses, as based on the assumptions $A_{(4)}$ per cloth, most favourably for cloth 5, a probability mass $\frac{1}{5}$ for $X_{4,5}$ assigned to the interval $(804, \infty)$ remains to the right of a probability mass $\frac{1}{5}$ for $X_{5,5}$ assigned to the interval $(-\infty, 801)$.

The imprecise probabilities in Table 4 show, for example, that cloths 2 and 6 are unlikely to lead to the best next observation, with the upper probability for cloth 6 slightly greater than for cloth 2. This is consistent with the related imprecise probabilities for pairwise comparisons, easily derived from the earlier results by deleting all but one alternative [5]. In this example, pairwise comparison of cloth j to cloth l leads to

$$\begin{aligned}\underline{P}(X_{j,5} > X_{l,5}) &= \frac{1}{25} \sum_{i_j=1}^4 c_{j,i_j}^l \\ \overline{P}(X_{j,5} > X_{l,5}) &= \frac{1}{25} \sum_{i_j=1}^4 (c_{j,i_j}^l + 1) + \frac{1}{5}.\end{aligned}$$

For these imprecise probabilities, we have $\bar{P}(X_{j,5} > X_{l,5}) = \underline{P}(X_{j,5} > X_{l,5}) + \frac{9}{25}$, and Table 5 gives all lower probabilities. The imprecise probabilities again satisfy $\bar{P}(X_{j,5} > X_{l,5}) = 1 - \underline{P}(X_{l,5} > X_{j,5})$.

$(\times \frac{1}{25}) \quad l =$	1	2	3	4	5	6	7
$j = 1$	-	16	0	0	0	16	0
2	0	-	0	0	0	7	0
3	16	16	-	5	0	16	11
4	16	16	11	-	1	16	13
5	16	16	16	15	-	16	16
6	0	9	0	0	0	-	0
7	16	16	5	3	0	16	-

Table 5: $\underline{P}(X_{j,5} > X_{l,5})$ for pairwise comparison (example 1)

Table 5 indicates that cloths 2 and 6 are very unlikely to be preferred to any other cloth in pairwise comparison, and that cloth 6 has a slight advantage over cloth 2 when compared with each other. As we saw before, cloth 4 is the only one to have positive lower probability of a greater next observation than cloth 5, and this lower probability is again $\frac{1}{25}$, so $\bar{P}(X_{5,5} > X_{4,5}) = \frac{24}{25}$. This value is equal to $\bar{P}\left[\bigcap_{l=1, l \neq 5}^7 \{X_{5,5} > X_{l,5}\}\right]$, reflecting the fact that, with probability masses put at the end points per interval such as to favour cloth 5 most, only cloth 4 has positive probability to lead to a greater next observation than cloth 5.

One should be careful with pairwise comparisons, and not use them for selection as an alternative to joint comparison to all other alternatives, since transitivity does not need to hold under pairwise comparisons. We pay some attention to this in example 2, but first we continue this first example by presenting some imprecise probabilities related to selected subsets, according to the theory of section 3 (Table 6).

S	$\underline{P}_S$	$\bar{P}_S$
4,5	0.1311	0.7440
3,4,5	0.1343	0.7520
4,5,7	0.0492	0.4800
3,4,5,7	0.2097	1
1,3,4,5	0	0.3104

Table 6: Some imprecise probabilities for selection of several cloths (example 1)

Selection of cloths 3, 4, 5 and 7 (Table 6) is interesting, as all observations for these four cloths are greater than all observations for cloths 1, 2 and 6, leading to upper probability 1. If we include cloth 1 in the subset instead of 7, the lower probability becomes 0 as all observations for cloth 7 exceed those for cloth 1. The large differences between the upper and lower probabilities in Table 6 again reflect the fact that there are only few observations, not enough to strongly support a claim that the next observations for some of these cloths will all be greater than those for the other cloths (a large value for a lower probability would support such a strong claim).

To finish this first example, we briefly consider imprecise previsions according to section 4. We choose the same lower and upper bounds, λ and ρ , for all cloths. Strong preference for cloth j over all other cloths, as modelled via imprecise previsions, is achieved if (since we have four observations for each cloth)

$$\sum_{i_j=1}^4 x_{j,i_j} - \sum_{i_l=1}^4 x_{l,i_l} > \rho - \lambda,$$

for all $l \neq j$. The sums of observations per cloth are (for cloth 1, ..., 7): 2554, 1714, 2998, 3072, 3261, 1794, 2925. The only candidate for strong preference over all cloths would be cloth 5, but $\sum_{i_5=1}^4 x_{5,i_5} - \sum_{i_4=1}^4 x_{4,i_4} = 189$ will not exceed $\rho - \lambda$, since $\rho - \lambda$ has to be greater than $840 - 373 = 467$, the difference between the maximum and minimum observation for all cloths. We can use the same method for pairwise comparison between cloths too [5], for example, if we use $l = 200$ and $r = 1000$, a difference between the sample sums for two cloths of more than 800 represents strong preference for the cloth with the largest sample sum over the other cloth. This easily leads to the following strong preferences for cloths: cloth 3, 4, 5 and 7 are all strongly preferred to cloths 2 and 6, and cloth 1 is strongly preferred to cloth 2, there are no further strong preferences in this pairwise comparison, when using $r - l = 800$.

Example 2

This is a brief example to show that pairwise comparison, based on imprecise probabilities, is not a useful alternative to the selection results in this paper, as transitivity does not hold. Consider a problem to select the best (maximum value) of three sources, based on predictive inferences as presented in this paper, with available ordered data per source as given in Table 7.

j	$x_{j,1}$	$x_{j,2}$	$x_{j,3}$
1	3.05	3.18	3.23
2	1.60	3.92	4.04
3	2.23	2.44	4.58

Table 7: Ordered data (example 2)

In this example, pairwise comparisons via our imprecise probabilities, as based on the assumption $A_{(3)}$ per source, imply $\bar{P}(X_{j,4} > X_{l,4}) = \underline{P}(X_{j,4} > X_{l,4}) + \frac{7}{16}$. The lower probabilities are given in Table 8.

$\left(\times \frac{1}{16}\right)$ $l =$	1	2	3
$j = 1$	-	3	6
2	6	-	4
3	3	5	-

Table 8: $\underline{P}(X_{j,4} > X_{l,4})$ for pairwise comparison (example 2)

Based on these imprecise probabilities, one would prefer (if choices must be made by pairwise comparison) source 1 to 3, source 3 to 2, but source 2 to 1, so there is no transitivity and making a choice might be difficult. Using our selection results of section 2, we compare one source with all other sources at the same time, and this leads to the imprecise probabilities given in Table 9.

j	$\underline{P}_j$	$\bar{P}_j$
1	0.0938	0.5313
2	0.1875	0.6406
3	0.1406	0.5625

Table 9: Imprecise probabilities for selection of one source (example 2)

On the basis of these imprecise probabilities, source 2 seems the logical candidate if a single source needs to be selected.

Example 3

The approach presented in this paper is quite generally applicable, without any restrictions on the sample sizes. Suppose that we have four sources, with sample sizes $n_1 = 20, n_2 = 18, n_3 = 15, n_4 = 3$, and ordered data as given in Table 10.

j	data
1	5.01 5.04 5.60 5.78 6.43 6.53 6.96 7.00 7.21 7.58 8.12 8.26 8.27 8.34 8.62 8.66 8.91 8.94 9.05 9.16
2	4.50 4.86 5.10 5.15 5.17 5.34 5.99 6.18 6.72 7.39 7.44 7.46 7.47 7.76 8.38 8.42 8.52 8.81
3	6.84 6.91 7.22 7.24 7.25 7.35 7.55 7.62 7.69 7.98 7.99 8.04 8.08 8.18 8.97
4	4.71 8.20 9.03

Table 10: Ordered data (example 3)

Using the results in section 2, assuming the appropriate assumption $A_{(n_j)}$ per source, these data lead to the imprecise probabilities, for each source to be best, as given in Table 11.

j	$\underline{P}_j$	$\overline{P}_j$
1	0.1991	0.3878
2	0.0756	0.1948
3	0.0850	0.2481
4	0.2747	0.5820

Table 11: Imprecise probabilities for selection of one source (example 3)

Although we only have three observations from source 4, two of these values are relatively large (when compared to the observations from all other sources), so this source would indeed not be unlikely to give the maximum value of a next observation, if one more observation were taken from every source. However, the particularly small sample size for source 4 is reflected by the large imprecision $\overline{P}_4 - \underline{P}_4$.

Suppose that it turns out that the lack of further data from source 4 is caused by this source not being available anymore. We delete it from our study and want to select one of the first three sources. Our approach now leads to the imprecise probabilities in Table 12.

j	$\underline{P}_j$	$\overline{P}_j$
1	0.3814	0.4930
2	0.1751	0.2643
3	0.2893	0.4054

Table 12: *Imprecise probabilities, source 4 deleted (example 3)*

Leaving source 4 out of consideration not only increases the lower and upper probabilities for the other sources, as expected, but also reduces the imprecision for those sources. This reflects the fact that inclusion of a source for which only few observations are available obscures the selection of a best source, particularly if some of those few observations are relatively large which provides some evidence that source 4 might be quite good, but this evidence is rather weak at the moment.

Acknowledgement

The authors are grateful to Mark van de Wiel (Eindhoven University of Technology) for developing a Mathematica program used for the calculations in this paper. Part of this research was carried out while the first named author visited Eindhoven University of Technology, with financial support from The British Council and NWO.

References

- [1] Bechhofer, R.E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics*, **25**, 16-39.
- [2] Bechhofer, R.E., Santner, T.J. and Goldsman, D.M. (1995). *Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons*. Wiley, New York.
- [3] Box, G.E.P., Hunter, W.G. and Hunter, J.S. (1978). *Statistics for Experimenters*. Wiley, New York.
- [4] Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*. Wiley, Chichester.
- [5] Coolen, F.P.A. (1996) Comparing two populations based on low stochastic structure assumptions. *Statistics & Probability Letters*, **29**, 297-305.
- [6] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, **36**, 349-357.
- [7] De Finetti, B. (1974). *Theory of Probability* (2 volumes). Wiley, London.
- [8] Dempster, A.P. (1963). On direct probabilities. *Journal of the Royal Statistical Society, Series B*, **25**, 100-110.
- [9] Geisser, S. (1993). *Predictive Inference: an Introduction*. Chapman and Hall, London.
- [10] Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. *Technometrics*, **7**, 225-245.
- [11] Gupta, S.S. and McDonald, G.C. (1970). On some classes of selection procedures based on ranks. In: *Nonparametric Theory in Statistical Inference* (M. Puri, ed.). Cambridge University Press, London, 491-514.
- [12] Gupta, S.S. and Yang, H.M. (1985). Bayes- P^* subset selection procedures for the best population. *Journal of Statistical Planning and Inference*, **12**, 213-233.
- [13] Hill, B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677-691.
- [14] Hill, B.M. (1988). De Finetti's Theorem, Induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference. In: *Bayesian Statistics 3* (J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds.). University Press, Oxford, 211-241 (with discussion).

- [15] Hill, B.M. (1993). Parametric models for A_n : splitting processes and mixtures. *Journal of the Royal Statistical Society, Series B*, **55**, 423-433.
- [16] Lane, D.A. and Sudderth, W.D. (1984). Coherent predictive inference. *Sankhyā: The Indian Journal of Statistics, Series A*, **46**, 166-185.
- [17] Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.
- [18] Marden, J.I. (1995). *Analyzing and Modeling Rank Data*. Chapman and Hall, London.
- [19] McDonald, G.C. (1979). Nonparametric selection procedures applied to state traffic fatality rates. *Technometrics*, **21**, 515-523.
- [20] Miescke, K.J. (1998). Bayes sampling designs for selection procedures. In: *Multivariate, Design, and Sampling* (S. Gosh, ed.). M. Dekker, New York, to appear.
- [21] Verheijen, J.H.M., Coolen, F.P.A. and Van der Laan, P. (1997). Combining two classical approaches for statistical selection. *Communications in Statistics - Theory and Methods*, **26**, 1291-1312.
- [22] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.