Enhancing Multiple Testing: Two Applications of the Probability of Correct Selection Statistic

Erin Tao, *student*
Jason Wilson, *supervisor*
Biola University

August 24, 2011

Abstract

The calculation of PCS shows how likely it is that the populations chosen as "best" truly are the top populations, according to a well-defined standard. PCS is useful for the researcher with limited resources or the statisticians attempting to test the quality of two different statistics. This paper explores the theory behind two selection goals for PCS, G-best and d-best, and how they improve previous definitions of PCS for massive datasets. This paper also calculates PCS for two applications that have already been analyzed by multiple testing procedures in the literature. The two applications are in Neuroimaging and Econometrics. It is shown through these applications that PCS not only supports the multiple testing conclusions but also provides further information about the statistics used.

1 Introduction

Because of the advancements in technology and science, a new development in statistics involves correctly and usefully analyzing massive datasets. With internet applications and financial data, there can be as many ten million populations to analyze, and sometimes more. Statisticians have developed methods such as family-wide error control and the false discovery rate to deal with the multiple testing problem, or the problem of finding too many false positives when testing k hypotheses simultaneously. This paper deals instead with ranking and selection methodology, which is a separate branch of statistics that has also been expanded to apply to massive datasets.

Ranking and Selection Methodology (RSM) is a well-defined system of ranking a set of populations based on sample data and selecting those that are "best." In laboratory research, resources are always limited. A scientist may want to know which of 10,000 genes available will provide the most information, to avoid studying all of them. Similarly, no one can invest in every company on the stock market, and so an investor only wants to know which ones will make the most money. In these two cases, "best" can be defined as the highest expression levels or the highest average returns on investment, respectively. Traditional hypothesis tests are not meant for ranking and selection purposes. Instead, one can calculate the probability of correct selection (PCS) to evaluate a chosen set of populations and see if the "best" have actually been chosen.

As with multiple testing procedueres, PCS has evolved in the last century from being accurate with large datasets (≈ 10) to being accurate with massive datasets. The two previous methods of ranking and/or selecting the "best" populations are the Indifference Zone (IZ) method, originated by Robert Bechhofer [1] and the Subset Selction (SS) method, originated by Shanti S. Gupta [4]. More recently there have been improvements in PCS for massive datasets by Cui and Wilson [2] in the form of G-best and d-best selection. In this paper, we explore both the theory and some applications of this improved method of calculating PCS.

Specifically, we look at the definitions of G-best and d-best selection, the use of index sets in those definitions, and the use of each selection goal. We also apply PCS to a Neuroimaging dataset and an Econometrics dataset. We find that PCS supports the results found using multiple testing procedures with the Neuroimaging application. In addition, PCS provides us with a measure of how accurate our choice of the "best" populations were. We were able to find the probability that the populations we chose as "best" based on sample data actually were the "best" populations. The same information was found for the Econometrics data, which is measured by two statistics. PCS was easily adapted for both statistics. These applications show the usefulness of PCS and how it can be applied generally.

The remainder of this paper is organized as follows. Section 2 gives account of the theory behind PCS and G-best and d-best selection. Two applications of PCS to datasets already analyzed by multiple testing procedures are given in Section 3. Finally, we draw concludions in Section 4.

2 G-best and d-best selection

2.1 Introduction

This section describes the mathematical theory behind G-best and d-best selection. The purpose of ranking and selection methodology is ultimately to choose the top t populations for some specified t. To do this, we first look at how to notate the ranking of both population parameters and sample statistics. We also use sets of indices to compact the definitions of G-best and d-best selection. The notation is somewhat subtle, but necessary, and is covered in Section 2.2. Furthermore, in Section 2.3, we define both G-best and d-best selection in terms of index sets, and describe how they each meet different needs of a researcher. We formally state how to calculate PCS in Section 2.4. Finally, we note the improvements these selection goals make for analyzing massive datasets in Section 2.5.

2.2 Notation

For clarification, we will use the following example throughout this section. Suppose we know the true means from five populations of interest. We also have taken samples from each population, and have calculated each sample mean. Our findings can be described in Table 1.

Population i	i =1	i =2	i =3	i =4	i =5
Sample Mean, Y_i	2.97	1.26	2.90	3.58	1.36
True Mean, θ_i	2.30	1.70	2.50	4.20	2.50

Table 1: *Each column of this table of example statistics shows information about a hypothetical population. Each population is numbered i=1,...,5. The sample mean is also recorded, above the true mean of the population. All of this information will be used to illustrate the notation for PCS.*

With this in mind, consider k populations, each with the same cumulative distribution function (CDF), except for location parameter θ_i, i=1,...,k. In the example, we have k=5. Let the population parameters of interest be denoted $\theta_1, ..., \theta_k$, and let $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ be the vector of those parameters. In Table 1, θ_1 would then be equal to 2.30. We are really interested in the order of the parameters, and so also have a numbering system for rank. Let $\theta_{(1)} \leq ... \leq \theta_{(k)}$ be the ordered parameters of $\boldsymbol{\theta}$. For example, in the table, $\theta_{(1)} = 1.70$, while $\theta_{(5)} = 4.20$.

Sometimes a researcher is interested in the largest statistics, and sometimes the smallest. The definition of the "best" populations must be defined explicitly for each application. Without loss of generality we assume in this paper that the "best" population has the largest statistic. What we are ultimately trying to find, then, is the population with the top t parameters, $\theta_{(k-t+1)}, ..., \theta_{(k)}$, or the top t parameters. For example, if we want the top $t = 3$ means from our example, we would want $\theta_{(5-3+1)}, ..., \theta_{(5)}$, or $\theta_{(3)}, \theta_{(4)}$, and $\theta_{(5)}$.

Because the top parameters are assumed to be unknown, we must pick a statistic Y to estimate the unkown population parameter. Each statistic will have a continuous CDF. Y_i denotes the particular statistic of the ith population. If we are interested in the usual mean, let the statistic Y denote the mean so that

$$Y_2 = Y(X_{2,1}, X_{2,2}, X_{2,3}) = (.75 + 1.78 + 1.25)/3 = 1.26$$

where $X_{2,1}, X_{2,2}, ...$ denote particular observations from the 2nd population.

To order the sample statistics, we use the notation $Y_{[i]}$, in that $Y_{[1]} \leq Y_{[2]} \leq ... \leq Y_{[k]}$. On the other hand, we denote the sample statistic that is drawn from the same population as the ordered parameter $\theta_{(i)}$ with $Y_{(i)}$. With Table 1, then, $Y_{[4]} = Y_{(2)} = 2.97$, because 2.97 is ranked four as far as the largest statistics, but it was from the population with $\theta_{(2)}$. With this notation, we can label our data, which is illustrated in Table 2.

To choose which populations we should assert as the top t, we use the top t statistics. The way we will notate correct selection is using index notation. Let s be the set of indices

Population i	i =2	i =5	i =3	i =1	i =4
Sample Mean	$Y_{[1]} = 1.26$	$Y_{[2]} = 1.36$	$Y_{[3]} = 2.90$	$Y_{[4]} = 2.97$	$Y_{[5]} = 3.58$
True Mean	$\theta_{(1)} = 1.70$	$\theta_{(3)} = 2.50$	$\theta_{(3)} = 2.50$	$\theta_{(2)} = 2.30$	$\theta_{(4)} = 4.20$

Table 2: *This table contains the same information as Table 1, but with the technical notation for PCS added. It is also now sorted by sample mean from lowest to highest.*

of the top t statistics. For example, if $t = 1$, then $Y_{(5)}$ is the top statistic, but it comes from population $k = 4$. So s in this case would be $s = \{4\}$. Then let A_t be the set of indices of the top t population parameters. In our case, $\theta_{(4)}$ is the highest, and also comes from population $k = 4$. Therefore, $A_t = \{4\}$ in this case as well. Rule R resulting in a correct selection is denoted by

$$CS_t = \{s = A_t\}.$$

Our example would result in a correct selection, then, because the sets s and A_t are equal.

It is important to note here that if two population parameters are equal ($\theta_i = \theta_j$ for some $i \neq j$), both are ranked equally, as we have done in Table 2. In past selection methods, if more than one parameter was equal in value, only one would be randomly chosen and asserted as the correct selection. This may significantly reduce the value of PCS in an unneccessary manner. The way we will handle this situation in this paper is as Cui and Wilson [2] define it. If population parameters θ_i and θ_j are equal, then $A_t = \{i\}$ or $\{j\}$. In other words, if $t = 1$ and $\theta_i = \theta_j$ are the top ranked populations, then either $s = \{i\}$ or $s = \{j\}$ will result in a correct selection. In our example, if $t = 2$, then $A_t = \{4, 3\}$ or $A_t = \{4, 5\}$ because 4.20 and 2.50 are the two largest values among the parameters. Therefore $s = \{4, 3\}$ or $s = \{4, 5\}$ would both result in a correct selection. This generalizes to handle more than two populations with the same values.

2.3 G-best and d-best selection

All of this notation is the set-up for the two selection goals we discuss in this paper: G-best and d-best selection, as defined by Cui and Wilson [2]. These are two different ways to define whether or not the populations that we choose as "best" really are the "best." One can use these methods before an experiment to determine how many subjects to study in order to ensure finding the top populations. After an experiment, one can use these goals to calculate the probability that the researcher has found the actual top t populations.

If our previous toy example was an actual experiment, we might have a need to find the top, say, one statistic, but we have the resources to study two. We would then use G-best selection, where we choose a fixed amount of populations, $t+G$, that contain the top t statistics. On the other hand, if we simply needed the populations to be within a certain threshold of quality, we would use d-best selection. In d-best selection we are finding a random number of populations, say r, which contains populations that are within a certain distance d from the top t populations. The number r is determined by an interval of pre-specified length d.

Defintion 2.1: Let s be the set of the indices corresponding to the top $t + G$ statistics for some pre-specified G. Let A_t be the set of indices of the top t parameters. Then

$$CS_{G,t} = \{A_t \subseteq s\}.$$

A set s that satisfies $CS_{G,t}$ is called G-best, and the probability that we have chosen a G-best set is denoted by $P(CS_{G,t})$.

For example, let $t = 2$ and let $G = 1$. We will choose $t+G = 3$ populations that we assert to contain the top two populations. We would then choose, from Table 2, $Y_4 = Y_{[5]} = 3.58$ and $Y_1 = Y_{[4]} = 2.97$ and $Y_3 = Y_{[3]} = 2.90$ as our top three statistics, to make $s = \{1, 3, 4\}$. A_t would be $A_t = \{4, 3\}$ or $\{4, 5\}$. In this case, since $A_t \subseteq s$, we have chosen correctly.

With this definition, a set is not G-best unless we have actually chosen the top t statistics. On the other hand, instead of only choosing t statistics to work with, we are choosing $t + G$ for some pre-specified G. Thus, the G parameter allows one to control the minimum proportion of "best" populations in the correct selection. For example, selecting the top 20 out of 20 voxel clusters in a neuroimaging scan might be highly unlikely. In this scenario we would have $t = 20$ and $G = 0$. This does not allow for any of the chosen populations to be wrong. However, suppose we can determine that having 90% of the populations actually being the "best" is allowable. In this case, $t = 18$ and $G = 2$, and the top 18 out of 20 might have a reasonable chance of actually being correct. It may be a low $P(CS_{G,t})$, but a reasonable chance is still an improvement. The point of ranking and selection procedures is to narrow down the populations to the "best" ones, and controlling the proportion of top populations among a group of populations does this.

Definition 2.2: Let s be set of the indices corresponding to the top t statistics. Let A_{t1} be the set of indices of the parameters in the interval $(\theta_{(k-t+1)} + d, \theta_{(k)}]$. Also let A_{t2} be the set of indices of the parameters in the interval $[\theta_{(k-t+1)} - d, \theta_{(k-t+1)} + d]$. See Figure 1 for a graphical representation of these intervals. A correct selection occurs when

$${}_dCS_t = \{A_{t1} \subseteq s \text{ and } s \backslash A_{t1} \subseteq A_{t2}\}.$$

where the operator $\backslash$ denotes the set difference operator, $B \backslash C = \{x : x \in B \text{ and } x \notin C\}$. If a set s satisfies ${}_dCS_t$, then it is said to be a d-best set. The probability of selecting a d-best set is $P({}_dCS_t)$.

This selection goal is more complex. To illustrate, let $t = 3$ and $d = .5$. Our s remains the same because our top three statistics are still the top three. So $s = \{1, 3, 4\}$. Referring to Table 2, we see that A_{t1} would be the set of indices of the parameters in the interval $(\theta_{(3)} + .5, \theta_{(5)}] = (3.00, 4.20)$. So $A_{t1} = \{4\}$. Then A_{t2} would be the set of indices of the parameters in the interval $[\theta_{(3)} - .5, \theta_{(3)} + .5] = [2.00, 3.00]$. $A_{t2} = \{1, 3, 5\}$. Our result is that $A_{t1} = \{4\} \subseteq s$ and $s \backslash A_{t1} = \{1, 3\} \subseteq A_{t2} = \{1, 3, 5\}$, resulting in a correct selection.

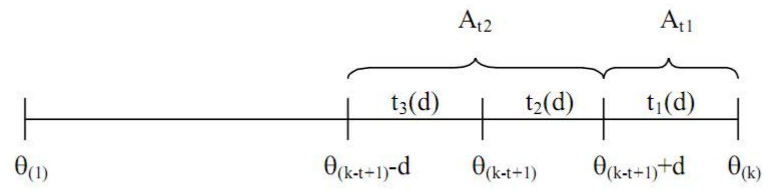

Figure 1: *The labels $t_1(d)$, $t_2(d)$, and $t_3(d)$ denote the number of population parameters in their respective intervals. Note: $t_1(d) + t_2(d) = t$. A_{t1} and A_{t1} are the sets of indices of the populations with values in their respective intervals.*

This selection goal has different advantages and disadvantages than G-best selection. Unlike a G-best set, a d-best set could contain indices of populations that are not actually in the top t, and exclude some that are in the top t. The population with the highest parameter must be chosen for the set to be considered a correct selection though. Furthermore, d-best selection ensures that the populations deemed a correct selection are within d of the best parameters. The situation for using d-best selection would be when a selection of t populations is desired, and the difference of d units between parameters is unimportant. For example, selecting the absolute top 10 "best" performing stocks might be virtually impossible (low $P({}_dCS_t)$), but the 10 best within \$0.50 might have a reasonable chance of success. Because fifty cents is negligable, the margin of error is acceptable.

2.4 Calculation of G-best and d-best selection

The formula for $P(CS_{G,t})$ and $P({}_dCS_t)$ is the same, but the calculations will differ based on the definitions of G-best and d-best sets. The formula is:

$$\sum_{g=1}^{|S|}\int_{-\infty}^{\infty}\prod_{j=k-t+1}^{k}[1-F(y-\theta_{s_{g,j}})]d\{\prod_{j=1}^{k-t}F(y-\theta_{\bar{s}_{g,j}}\} \tag{1}$$

We use S to denote the set of all G-best or d-best sets, or all the sets s that will result in a correct selection. Furthermore, let $\boldsymbol{\theta}_g$ be the g-th combination of ordered parameters. This will be a set of sets, each of which will contain the highest parameter $\theta_{(k)}$ and be of size t. Then $\boldsymbol{\theta}_g = \{\boldsymbol{\theta}_{s_{g,j}}, \boldsymbol{\theta}_{\bar{s}_{g,j}}\}$ where $\boldsymbol{\theta}_{s_{g,j}}$ denotes the combinations of parameters that satisfy the specific sets $s_{g,j} \in S$ and $\boldsymbol{\theta}_{\bar{s}_{g,j}}$ contains the combinations of parameters that do NOT satisfy G-best or d-best sets. That is, $\bar{s}_{g,j} \in \bar{S}$. $F(y - \theta_{s_{g,j}})$ is the continuous cumulative distribution function of the statistic y, adjusted to center around 0. See Cui and Wilson [2] for the derivation of Equation (1).

These are extremely difficult integrals to integrate, analytically and even numerically. There are expansions that simplify the expression in order to make a numerical solution possible (specifically, Gauss-Hermite quadrature, Cui and Wilson [2]). To calculate the PCS for our datasets we use an R package that uses a parametric bootsrapping method.

2.5 The Improvements for Massive Datasets

The use of G-best and d-best selection is very practical with massive datasets. Using ranking and selection methods to reduce the number of populations to study can be used along with multiple testing methods, but it may even suit the needs of the researcher better than these methods. Furthermore, G-best and d-best selection are an improvement on previous definitions of PCS with respect to massive datasets. First of all, because G-best and d-best sets are defined in terms of index sets, they deal with the problem of having two equal parameters effectively. Also, d-best selection is especially useful for a dataset with high density, which is characteristic of massive datasets. The more the population parameters are approximately equal to other parameters, the more dense the data is. A researcher may not actually be interested in the absolute top t populations, but rather which populations will be most worthwhile to study. With d-best selection, the researcher can choose an interval around the true top parameters that is allowably close to find populations that may not be the "best," but will be worth spending time on.

3 Application

3.1 Introduction

Although G-best and d-best selection are fully generalizable, to date they have only been applied to microarray data in the literature. To test and illustrate the applicability and usefulness of $P(CS_{G,t})$ and $P({}_dCS_t)$, we will calculate the probability of correct selection to Neuroimaging [5] and Econometrics [6] data. To calculate the probability of correct selection for these applications, we use an R package called "PCS," which can be found at www.r-project.org.

3.2 Neuroimaging

First of all, we look at brain scans from a test on verbal fluency. Scientists conducted the study on five people who both listened passively and said words aloud. They then studied whether areas of the brain were activated more in listening to or in generating words. To study the brain, they used 3D scans comprised of voxels, which are three-dimensional pixels. Figure 3.2 shows the shaded voxels that represent activated areas of the brain common to all five subjects.

In 2003, Nichols and Hayasaka took this study and measured each of the 55,027 voxels of the brain scans to see if any part of the brain was more active for word generation as opposed to passive listening [5]. With this particular experiment, no voxels were found to be significant. To analyze these results with PCS, we used a program called SPM8 [3], to find the possible clusters of voxels that might be significant in the conjunction of all five brains. We then calculated the probability that one, two, three or four of these clusters may actually be the "best" clusters of voxels, or show the most brain activity. The calculation of PCS for

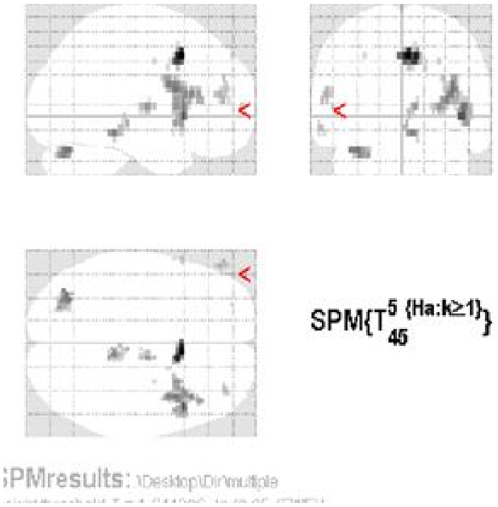

Figure 2: *This scan shows the three planes that represent the location of an area of the brain. For example, the arrow in the image is indicating an area of the brain in the front left of the brain, a little less than halfway from the bottom.*

this data supports Nichols' and Hayasaka's conclusion that there were no significant voxels (see Table 3).

The very low probabilities in Table 3 show that if one were to choose even one cluster as significant, it would not likely be the "best" cluster of voxels. Consider the probability of correct selection of t=1 for both G=2 and d=.5. To understand the meaning of the probability of a G-best selection, refer to Defintion 2.1. For t=1 and G=2, the probability of correct selection is .33. In the table, this is notated as $P(CS_{2,1}) = .33$. If one chooses the top three clusters of voxels, there is only a .33 probability that the "best" cluster is among them. For d-best selection, refer to Definition 2.2. With t=1 and d=.5, the probability of correct selection is .19. That is, $P({}_{.5}CS_1) = .19$. The probability that the top cluster of voxels is even within a margin of .5 of the top clusters is only .19 .This complements the multiple testing result that none of the voxels are significantly different from any of the others. The highest probability found was for t = 1 and G =6. These parameters result in a probability of more than half. This supports Nichols' and Hayasaka's findings but adds the information that we would have to choose seven clusters just to find one that stands out.

3.3 Econometrics

The Economics data we chose comes from the Center for International Securities and Derivatives Market (CISDM) from January 1992 to March 2004. There are 105 hedge funds, and each fund has 147 recorded returns, one from each month in the time period. Instead of simply recording the return on investment for each month, the data records the amount the return above (or below) a certain benchmark. In this case, the benchmark is the risk-free rate, or rate of return on an investment with zero risk. Romano and Wolf [6] used stepwise

$t =$	1	2	3	4
$P(CS_{0,t}) = P({}_0CS_t) =$	.11	.02	.00	.00
$P(CS_{2,t}) =$	.33	.11	.03	.01
$P(CS_{4,t}) =$	.49	.22	.08	.03
$P(CS_{6,t}) =$	.60	.33	.15	.06
$P({}_{.5}CS_t) =$	.19	.04	.05	.03
$P({}_1CS_t) =$	.24	.19	.35	.29

Table 3: *These low probabilities support Nichols' and Hayasaka's findings. For example, when $t = 3$ and G=4, the probability that the clusters selected actually are "best" is only .08.*

multiple testing procedures to find the top ten absolute best performing funds. They defined "best" as the fund with the largest return in excess of the benchmark. We have calculated the probability that the ten funds that Romano and Wolf chose are actually the top ten using different parameters of G-best and d-best selection, but always choosing ten funds. The index set of the ten chosen funds is

$$s = \{31, 105, 16, 8, 25, 101, 38, 4, 82, 57\},$$

and the chosen funds according to Romano and Wolf [6] are shown in Table 4. Table 5 shows the probabilities of correct selection.

Excess	Fund
1.70	Libra Fund
1.41	Private Investment Fund
1.36	Aggressive Appreciation
1.27	Gamut Investments
1.26	Turnberry Capital
1.14	FBR Weston
1.11	Berkshire Partnership
1.09	Eagle Capital
1.07	York Capital
1.07	Cabelli Intl.

Table 4: *These are the names of the top ten performing funds from January 1992 to March 2004 according to the average return in excess of the risk free rate.*

From these probabilities, one can see that it is not very certain that all ten funds chosen are truly the top ten funds. $P(CS_{0,10}) = P({}_0CS_{10})$=.13 shows that these ten funds only have a 13% chance of being the top ten. Using G-best selection, we can be confident that these top ten statistics contain the top five funds, but that only accounts for half of the funds chosen. Furthermore, by looking at d-best selection, there is an .86 probability that the top ten funds found with the absolute statistic are within a margin of 2 of the top ten funds. This is actually a very large margin. The reason the probabilities are not that certain is because there is a large amount of variability in this statistic. To address this issue, Romano and Wolf propose standardizing the statistic.

Absolute	Statistic			Studentized	Statistic		
$P(CS_{G,t})$		$P({}_dCS_t)$		$P(CS_{G,t})$		$P({}_dCS_t)$	
$P(CS_{0,10})$	$= 0.13$	$P({}_0CS_{10})$	$= 0.13$	$P(CS_{0,10})$	$= 0.71$	$P({}_0CS_{10})$	$= 0.71$
$P(CS_{1,9})$	$= 0.29$	$P({}_{.5}CS_{10})$	$= 0.32$	$P(CS_{1,9})$	$= 0.98$	$P({}_{.5}CS_{10})$	$= 0.71$
$P(CS_{2,8})$	$= 0.53$	$P({}_1CS_{10})$	$= 0.52$	$P(CS_{2,8})$	$= 1.00$	$P({}_1CS_{10})$	$= 0.95$
$P(CS_{3,7})$	$= 0.74$	$P({}_{1.5}CS_{10})$	$= 0.78$	$P(CS_{3,7})$	$= 1.00$	$P({}_{1.5}CS_{10})$	$= 0.95$
$P(CS_{4,6})$	$= 0.92$	$P({}_2CS_{10})$	$= 0.86$	$P(CS_{4,6})$	$= 1.00$	$P({}_2CS_{10})$	$= 0.97$
$P(CS_{5,5})$	$= 1.00$	$P({}_{2.5}CS_{10})$	$= 0.93$	$P(CS_{5,5})$	$= 1.00$	$P({}_{2.5}CS_{10})$	$= 1.00$
$P(CS_{6,4})$	$= 1.00$	$P({}_3CS_{10})$	$= 0.99$	$P(CS_{6,4})$	$= 1.00$	$P({}_3CS_{10})$	$= 1.00$
$P(CS_{7,3})$	$= 1.00$	$P({}_{3.5}CS_{10})$	$= 1.00$	$P(CS_{7,3})$	$= 1.00$	$P({}_{3.5}CS_{10})$	$= 1.00$
$P(CS_{8,2})$	$= 1.00$	$P({}_4CS_{10})$	$= 1.00$	$P(CS_{8,2})$	$= 1.00$	$P({}_4CS_{10})$	$= 1.00$
$P(CS_{9,1})$	$= 1.00$	$P({}_{4.5}CS_{10})$	$= 1.00$	$P(CS_{9,1})$	$= 1.00$	$P({}_{4.5}CS_{10})$	$= 1.00$

Table 5: *The first two columns show PCS for the absolute statistic. These probabilities are significantly lower than those from the studentized statistic, reflected in the third and fourth columns. The difference is due to the studentized statistic accounting for the variability in the data.*

Romano and Wolf studentize the absolute statistic, that is, they used the usual t-statistic, which is calculated by dividing the statistic used above by the standard error. Romano and Wolf estimate variance using a sophisticated method involving a time series bootstrap, whose code is unavailable. Thus, for this paper we simply divide the first statistic by the usual standard error (standard deviation/$\sqrt{n}$) of each fund. The top ten funds Romano and Wolf chose using the t-statistics were a completely disjoint set from the non-studentized set of "best" funds. The index set of the top ten funds found using the usual standard error is

$$s = \{61, 60, 102, 100, 23, 30, 18, 22, 63, 46\}$$

Just as in the article, this is a completely different set of ten funds chosen as "best." Table 5 shows the probability that the studentized statistic shows the true top ten studentized funds.

As one can see, the probability that these ten chosen funds are in actuality the best funds is significantly more than with the non-studentized statistic. From these results, we can see that the studentized statistic, even using the usual standard error for each fund, is much more likely to identify the true "best" hedge funds according to the t-statistic.

Romano and Wolf show that the studentized statistic is a better measure of the performance of a fund because it takes into account the amount of risk involved. As one can see in Figure 3, the magnitude of the top fund chosen according to the absolute statistic is much larger than the top fund chosen according to the studentized statistic. However, the second graph in Figure 3 shows that the top fund according to the studentized statistic is in positive excess of the risk-free rate for the vast majority of the months recorded. The high probabilities found with PCS further cement that the standardized statistic is superior to the absolute statistic. It is also important to note that the PCS of the studentized statistics may change with the more sophisticated estimate of variance. Still, even in this application, PCS provides useful information on both the absolute and studentized statistics.

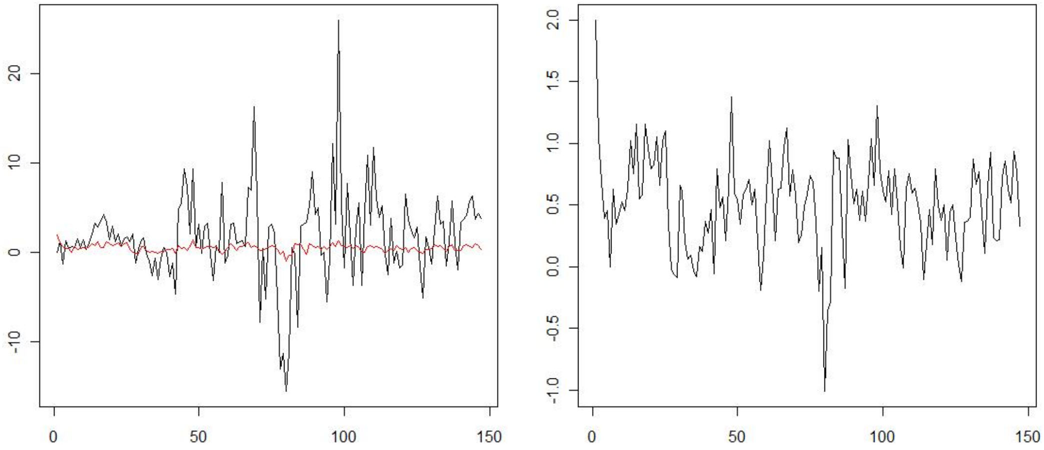

Figure 3: *The graph on the left shows the fund chosen as "best" according to the absolute statistics (black) and the fund chosen as "best" according to the studentized statistic (red). The graph on the right shows the top fund chosen using the studentized statistic alone.*

3.4 Results

Applying PCS to these areas shows how useful Ranking and Selection methodology can be. The probability of correct selection has thus far been consistent with the latest multiple testing procedures, as it is in the Neuroimaging application. However, PCS provides different information than a multiple hypothesis test. In each application we found a measure of how accurate our chosen "best" populations actually are. Instead of simply choosing the "best" statistics, one can have a better idea of how close they are to actually being "best." If a neuroimaging scientist actually found a significant cluster of voxels, he or she would know how unlikely it is for that cluster to be "best." With the Econometrics application, investors can see the probability that the ten funds chosen either contain the actual top t funds, or that they are within a certain margin of the actual top ten funds. This is valuable information that can help drive the development of a new hypothesis.

4 Conclusions

The probability of correct selection is a useful tool in statistics, and we have striven to illustrate this through both the theory behind G-best and d-best selection and its application to differing areas. The use of PCS deals with the problem of massive datasets by accommodating dense datasets that may have many parameters in common or close enough to study. Furthermore, G-best and d-best selection are useful tools for a researcher with limited resources. Instead of having a list of significant populations too large to adequately study, one can actually find the populations that are most likely to be the "best". Depending on the needs of the researcher, G-best or d-best selection may be more useful. Both selection goals were found to be consistent with previous claims of significance in the Neuroimaging application, which supported their validity. In the Econometrics application, PCS provided information on two separate statistics in the same study, which showed its adaptability. In both applications, PCS provided additional information that was not available through hypothesis testing. With PCS, we gain an insight into the quality of the populations chosen as "best" by seeing how likely it is that they truly are "best." Clearly, PCS is a powerful tool.

For further research, we would like to find the variance estimator for a time series regression bootstrap in order to find the PCS of the top ten funds actually chosen by Romano and Wolf. We would also like to apply PCS to mass spectrometry and other large k populations found in the literature.

References

[1] Bechhofer, R. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist 25*, 1 (1954), 16–39.

[2] Cui, X., and Wilson, J. On the probability of correct selection for large k populations with application to microarray data. *Biometrical Journal 50*, 5 (2008), 870–883.

[3] Guillaume. Statistical parametric mapping. `http://www.fil.ion.ucl.ac.uk/spm/`.

[4] Gupta, S. *On a Decision Rule for a Problem in Ranking Means.* PhD thesis, University of North Carolina, 1956.

[5] Nichohls, T., and Hayasaka, S. Controlling the familwise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research 12*, 5 (2003), 419–446.

[6] Romano, J. P., and Wolf, M. Stepwise multiple testing as formalized data snooping. *Econometrica 73*, 4 (2005), 1237–1282.