Efficient Fully Sequential Indifference-Zone Procedures Using Properties of Multidimensional Brownian Motion Exiting a Sphere

A.B. Dieker
Columbia University
New York, NY 10027

Seong-Hee Kim
Georgia Institute of Technology
Atlanta, GA 30332-0205

June 21, 2017

Abstract

We consider a ranking and selection (R&S) problem with the goal to select a system with the largest or smallest expected performance measure among a number of simulated systems with a pre-specified probability of correct selection. Fully sequential procedures take one observation from each survived system and eliminate inferior systems when there is clear statistical evidence that they are inferior. Most fully sequential procedures make elimination decisions based on sample performances of each possible pair of survived systems and exploit the bound crossing properties of a univariate Brownian motion. In this paper, we present new fully sequential procedures with elimination decisions that are based on sample performances of all competing systems. Using properties of a multidimensional Brownian motion exiting a sphere, we derive heuristics that aim to achieve a given target probability of correct selection. We show that in practice the new procedures significantly outperform a widely used fully sequential procedure. Compared to BIZ, a recent fully-sequential procedure that uses statistics inspired by Bayes posterior probabilities, our procedures have better performance under difficult mean or variance configurations but similar performance under easy mean configurations.

Subject classification: Simulation, Ranking and Selection, Fully Sequential, Multidimensional Brownian Motion, Sphere

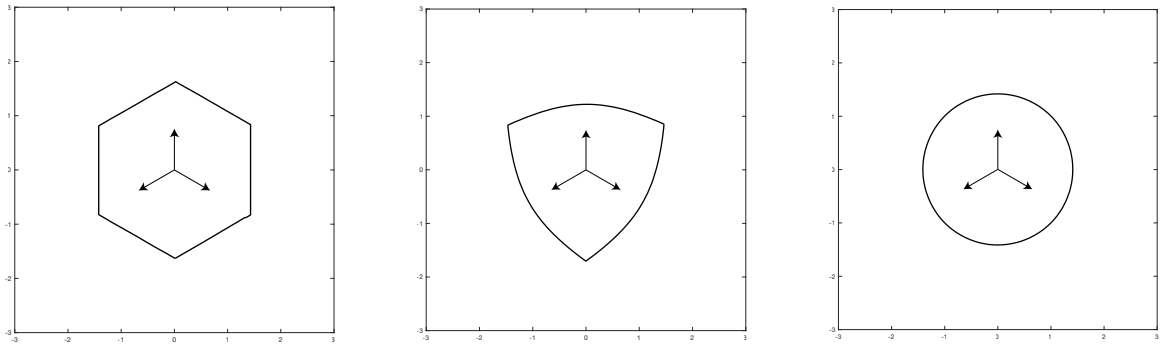
1. Introduction

Ranking and selection (R&S) is one of the classical and well-studied problems in the operations research literature. It aims to find the best system among a number of systems for which noisy performance information is accessible through simulation. In this paper, we assume that the best system is one with the largest or smallest expected performance, which is known as the finding-the-best problem. There are at least three approaches for the finding-the-best problem: the indifference-zone (IZ) approach, the Bayesian approach, and the optimal computing budget allocation (OCBA) approach. Hong et al. (2014) and Kim and Nelson

(2011) provide a brief review of each approach. For more information on the Bayesian approach, see Chick (2006) and Chen et al. (2014). When there is a fixed computing budget until a decision is made, the OCBA approach provides an efficient way to find the best system, see for example Chen and Lee (2011). In this paper we study an indifference-zone (IZ) procedure, where the decision maker specifies a difference worth detecting called the IZ parameter.

Among procedures that take the IZ approach, Rinott (1978) is one of the earliest procedures. It is a two-stage procedure and does not have any elimination step for clearly inferior systems. Nelson et al. (2001) also propose a two-stage procedure but their procedures can eliminate systems after the first stage if there is statistical evidence that they are inferior. Therefore the latter procedure is more efficient than Rinott's procedure in terms of the number of observations needed until a decision is made. On the other hand, fully-sequential IZ procedures take one observation from competing systems and eliminate inferior systems as additional observations become available. They carry the risk of incorrectly eliminating the best system due to stochastic noise in the performance measurements. Examples of fully-sequential IZ procedures are the KN procedures from Kim and Nelson (2001), which are widely used as they are available in leading commercial simulation software. KN's parameters are chosen to control the probability of eliminating the best system. Since this probability is intractable, the procedures instead rely on a Bonferroni-type lower bound on the worst-case probability of incorrect selection, which corresponds to the best system having a mean performance that exceeds the means of the other systems by exactly the IZ parameter; this setup is known as the slippage configuration (SC). Particularly when the number of systems is large, this lower bound tends to be a poor approximation for the worst-case probability of correct selection. As discussed in Wang and Kim (2011), the result is that KN procedures tend to take many more observations than necessary to control the probability of incorrect selection, and are thus inefficient in that sense.

The primary contribution of this paper is to develop a new family of IZ procedures that does not suffer from the inefficiencies caused by the use of the Bonferroni bound. The screening statistic used in an IZ procedure gives rise to contours, and a system is eliminated when the vector of cumulative sums of performance measurements hits such a contour, see Figure 1. Controlling PICS is done by analyzing hitting behavior of Brownian motion to set the 'radius' of the contour. It is the second ingredient where the Bonferroni bound



(a) KN; $\min_{i<j} |x_i - x_j|$ (b) BIZ; $\min_i e^{x_i}/(e^{x_1} + e^{x_2} + e^{x_3})$ (c) This paper; $\sum_{i<j}(x_i - x_j)^2$

Figure 1: Contours of the screening statistics with $k = 3$ competing systems for three procedures. The form of the screening statistic is given below each figure. Also depicted are each of the three possible drifts (mean sample paths) under the slippage configuration (SC). Since the screening statistics do not change when adding a constant to each coordinate, we have plotted the plane $\{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0\}$ with its intersections of the contours.

is invoked for KN procedures, since it is analytically intractable to study properties of a Brownian motion hitting KN contours, see Figure 1(a).

An important recently developed family of IZ ranking and selection procedures is the Bayes-inspired indifference zone (BIZ) procedures from Frazier (2014). An example of the resulting elimination contours is given in Figure 1(b). BIZ procedures circumvent the use of the Bonferroni bound, which results in dramatic improvements over the KN procedure especially when the number of competing systems is large. We see in Figure 1(b) that the three possible drifts under SC (one for each of the systems being the best) hit the elimination contour at its closest point to the origin. Thus, possibly sample paths that deviate significantly from their mean sample paths require a larger number of observations from the competing systems than those that are close to their mean sample paths.

This paper is a first investigation towards efficient procedures with spherical elimination contours, see Figure 1(c). Using path length as a proxy for the number of observations needed, with spherical elimination contours all points on the contour are equally close to the origin, and this could perhaps lead to faster elimination. Setting the radius of the contours given a target probability of correct selection is facilitated by some analytic results about a multidimensional Brownian motion hitting hyperspheres. Like BIZ, we do not need to appeal to the Bonferroni bound to exert control over the probability of correct selection. However,

our procedure is ultimately heuristic since we found it intractable to control the multi-stage probability of correct selection; we leave this as an open problem.

Experimental results show that estimated probability of correct selection is all higher or close to the target confidence level for all cases tested including SC. Our procedures also significantly outperform KN. On the other hand, our procedures perform better than BIZ under difficult mean or variance configurations while they perform similarly compared to BIZ under easy mean configurations. More specifically, when variances are unknown and unequal across systems with a slippage mean configuration, our procedures show up to 30% savings compared to the BIZ procedures in terms of the number of replications needed until a decision is made. Under easier scenarios where means spread out over systems unlike SC, our procedures perform similar to the BIZ procedures.

To extend our procedures to unknown and unequal variances, we use multiple tricks. These tricks are standard but render any statistically valid approach (including BIZ) heuristic. To handle unknown variances, we update variance estimates as the procedures advance. Kim and Nelson (2006) show that variance update enables a procedure to be treated as if variances are known in an appropriate limit. To handle unequal variances we use a heuristic approach which essentially changes the sampling frequency of each system hoping to approximately equalize variances across systems.

Preliminary work related to this work is published in the Winter Simulation Conference proceedings which include Kim and Dieker (2011) and Dieker and Kim (2012, 2014). The first two papers consider only three systems with known variances. Dieker and Kim (2014) give a procedure for a general number of systems but require known and equal variances. Moreover, the spheres that play a crucial role in the procedure all have the same radius and the procedure performs worse than KN when the means of the systems are spread out evenly. In the procedures presented in the present paper, the radii of the spheres vary as the number of survived systems decreases, outperforming KN in all scenarios; and a version of our procedure can handle unknown and unequal variances.

When there exists a finite simulation budget or a tight deadline in time, OCBA and Bayesian procedures are shown to be highly efficient and very useful in practice. Branke et al. (2007), Chen and Lee (2011) and Powell and Ryzhov (2012) provide a good review of OCBA and Bayesian ranking and selection procedures

and provide extensive empirical results. As our primary goal is to investigate the impact of different shapes of continuation regions in IZ ranking and selection procedures, we only compare our procedures with IZ procedures. Specifically, two state-of-art IZ procedures, KN and BIZ, are considered.

The paper is organized as follows. Section 2 defines our problem and introduces notation. Section 3 proposes new fully-sequential procedures. Section 4 explains the statistics that we use for elimination decisions and the properties of our statistics. In Section 5, we provide justifications for our procedures and approximations in order to set the parameter values of the procedures. Experimental results are presented in Section 6, followed by conclusions in Section 7.

2. Problem and Notation

This section introduces our notation and assumptions and defines the problem. We assume there are k systems ($k \geq 2$). Let X_{ij} represent the jth observation from system i for $i = 1, \ldots, k$ and $j = 1, 2, \ldots$. Then the mean and variance of the outputs from system i are defined as $\mu_i = \mathrm{E}[X_{ij}]$ and $\sigma_i^2 = \mathrm{Var}[X_{ij}]$, respectively. We want to find the system with the largest mean μ_i.

Throughout the paper, we assume that the following assumptions hold:

Assumption 1.

$$X_{ij} \overset{iid}{\sim} \boldsymbol{N}(\mu_i, \sigma_i^2), \quad j = 1, 2, \ldots,$$

where $\overset{iid}{\sim}$ *represents 'are independent and identically distributed as' and* $\boldsymbol{N}(\mu_i, \sigma_i^2)$ *denotes the normal distribution with mean* μ_i *and variance* σ_i^2. *Moreover,* X_{ij} *and* $X_{i'j}$ *are independent for any* $i \neq i'$ *and* $j = 1, 2, \ldots$.

Assumption 2. $\mu_1 \leq \mu_2 \leq \ldots \leq \mu_{k-1} \leq \mu_k - \delta$ *for* $\delta \in \mathbb{R}^+$.

Assumption 1 implies that observations from each system are marginally iid normally distributed and systems are simulated independently (note that this rules out common random numbers). Without loss of generality, we assume that system k is the true best system. Assumption 2 assumes that the mean of the true

best system k is at least δ better than any alternative system. The parameter δ is a user-specified parameter known as the IZ parameter.

We aim to devise a method that observes systems sequentially and eliminates clearly inferior systems from further consideration. The method stops once only one system remains, and this system is declared as the best system.

Additional notation is needed for later sections:

$$
\begin{aligned}
n &\equiv \text{the current number of observations or the current stage number;} \\
I &\equiv \text{set of competing systems at the } n\text{th stage;} \\
\bar{X}_i(n) &\equiv \frac{1}{n}\sum_{j=1}^{n} X_{ij}, \text{ the sample mean of system } i \text{ based on the first } n \text{ observations;} \\
\boldsymbol{X}_I(n) &\equiv |I|\times 1 \text{ vector of } \textstyle\sum_{j=1}^{n} X_{ij} \text{ for } i \in I; \\
\hat{\sigma}_i^2(n) &\equiv \text{sample variance of system } i \text{ from } X_{i1},\ldots,X_{in} \text{ which is } \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i(n))^2; \\
A^T &\equiv \text{the transpose of a matrix } A; \\
\delta_{|I|}^2 &\equiv \delta^2\frac{|I|-1}{|I|}.
\end{aligned}
$$

3. $\mathcal{DK}$ Procedures

In this section, we provide the descriptions of our new procedures. We present $\mathcal{DK}_1$ for known and equal variances and extend it to unknown but equal variances, resulting in $\mathcal{DK}_2$. Then $\mathcal{DK}_3$ is presented for unknown and unequal variances.

3.1 Equal and Known Variances

We first consider a case where variances are equal across all systems and known so that $\sigma_i^2 = \sigma^2$ for any system i. Suppose $x \in \mathbb{R}^s$ and $I \subset \{1,\ldots,k\}$ and define a function $\mathcal{S}_I(x)$ as follows:

$$\mathcal{S}_I(x) = \frac{1}{\sigma^2}\sum_{i\in I}(x_i - \bar{x})^2$$

where $\bar{x} = \frac{1}{s}\sum_{i\in I} x_i$.

The $\mathcal{DK}_1$ procedure for equal and known variances is as follows:

The $\mathcal{DK}_1$ Procedure

Setup: Select the nominal level $1-\alpha$ and the IZ parameter δ. Set $I = \{1, 2, \ldots, k\}$ and choose $\eta_{|I|}$ (which will be discussed in Section 5). Take one observation from each system. Set $n = 1$ and go to **Calculation**.

Calculation: Calculate $\mathcal{S}_I(\boldsymbol{X}_I(n))$.

Screening: If $\mathcal{S}_I(\boldsymbol{X}_I(n)) \geq \left(\frac{\sigma \cdot \eta_{|I|}}{\delta_{|I|}}\right)^2$, then eliminate the system with the smallest $\bar{X}_i(n)$ among $i \in I$. Update I by removing the eliminated system and go back to **Calculation**. Otherwise, go to **Stopping Rule**.

Stopping Rule: If $|I| = 1$, stop and declare the surviving system as the best. Otherwise, take one more observation for all $i \in I$, set $n = n + 1$, and go to **Calculation**.

In Section 4, we show that $\mathcal{S}_I(x)$ calculates the squared distance of the point x orthogonally projected onto a hyperplane $\{x : \sum_{i \in I} x_i = 0\}$ and that the screening rule in $\mathcal{DK}_1$ implies that we have an open (infinite) cylinder as our continuation region with a radius depending on $\eta_{|I|}$ and $\delta_{|I|}$. When x is located outside the cylinder, elimination occurs and the screening rule is checked again with updated parameters (without obtaining additional observations), i.e., a lower dimensional cylinder. We only obtain new observations (i.e., move to the next stage) if no more elimination occurs for a given number of observations.

3.2 Unknown but Equal Variances

We present a straightforward variant of $\mathcal{DK}_1$ for unknown but equal variances, σ^2. As the variance parameter σ^2 is unknown, it needs to be estimated. Let $\hat{\sigma}_i^2(n)$ represent sample variance of system i. The pooled variance estimator $\hat{\sigma}_p^2(n)$ is defined as follows:

$$\hat{\sigma}_p^2(n) = \frac{1}{|I|} \sum_{i \in I} \hat{\sigma}_i^2(n).$$

Then our statistic is modified to

$$\mathcal{S}'_I(x) = \frac{1}{\hat{\sigma}_p^2(n)} \sum_{i \in I} (x_i - \bar{x})^2$$

and $\hat{\sigma}_i^2(n)$ and $\hat{\sigma}_p^2(n)$ need to be updated in the [Stopping Rule] step after additional observations are obtained. Then the $\mathcal{DK}_2$ procedure is defined below.

The $\mathcal{DK}_2$ Procedure

Setup: Select the nominal level $1-\alpha$ and the IZ parameter δ. Set $I = \{1, 2, \ldots, k\}$ and choose $\eta_{|I|}$. Take $n_0 \geq 2$ observations from each system and calculate $\hat{\sigma}_i^2(n_0)$ and $\hat{\sigma}_p^2(n_0)$. Set $n = n_0$ and go to **Calculation**.

Calculation: Calculate $\mathcal{S}'_I(\boldsymbol{X}_I(n))$.

Screening: If $\mathcal{S}'_I(\boldsymbol{X}_I(n)) \geq \left(\frac{\hat{\sigma}_p(n)\cdot\eta_{|I|}}{\delta_{|I|}}\right)^2$, then eliminate the system with the smallest $\bar{X}_i(n)$ among $i \in I$. Update I by removing the eliminated system and go back to **Calculation**. Otherwise, go to **Stopping Rule**.

Stopping Rule: If $|I| = 1$, stop and declare the surviving system as the best. Otherwise, take one more observation for all $i \in I$; set $n = n + 1$; and update $\hat{\sigma}_i^2(n)$ for $i \in I$ and $\hat{\sigma}_p^2(n)$. Then go to **Calculation**.

3.3 Unknown and Unequal Variances

This subsection extends the $\mathcal{DK}_1$ procedure to handle unknown and unequal variances, resulting in $\mathcal{DK}_3$. The main idea is to make the sampling frequency of each system proportional to the variance parameter of the system, which eventually leads to equal variances. This approach is similar to the one in Frazier (2014).

Let n_i denote the number of observations system i have received so far. In $\mathcal{DK}_1$, $n_i = n$ for any system $i \in I$ but in $\mathcal{DK}_3$, $n_i \leq n$. Also let $W_i(n) = \sum_{j=1}^{n_i} X_{ij}/n_i$ and $\boldsymbol{W}_I(n)$ represent a $|I| \times 1$ vector of $W_i(n)$ for $i \in I$.

Then

$$\mathcal{S}''_I(x) = \frac{1}{\hat{\lambda}^2} \sum_{i \in I} (x_i - \bar{x})^2$$

where

$$\hat{\lambda}^2 = \frac{\sum_{i \in I} \hat{\sigma}_i^2(n_i)}{\sum_{i \in I} n_i}.$$

We can now describe Procedure $\mathcal{DK}_3$.

The $\mathcal{DK}_3$ Procedure

Setup: Select the nominal level $1-\alpha$ and the IZ parameter δ. Also select a constant B_z. Set $I = \{1, 2, \ldots, k\}$ and choose $\eta_{|I|}$. Take n_0 observations from each system and calculate $W_i(n_0)$, $\hat{\sigma}_i^2(n_0)$ and $\hat{\lambda}^2$. Set $n = n_0$ and $n_i = n_0$ for $i \in I$, and go to **Calculation**.

Calculation: Calculate $\mathcal{S}_I''(\boldsymbol{W}_I(n))$.

Screening: If $\mathcal{S}_I''(\boldsymbol{W}_I(n)) \geq \left(\frac{\hat{\lambda}\cdot\eta_{|I|}}{\delta_{|I|}}\right)^2$, then eliminate the system with the smallest $\bar{X}_i(n)$ among $i \in I$. Update I by removing the eliminated system and go back to **Calculation**. Otherwise, go to **Stopping Rule**.

Stopping Rule: If $|I| = 1$, stop and declare the surviving system as the best. Otherwise, let $z = \arg\min \frac{n_i}{\hat{\sigma}_i^2(n_i)}$ for $i \in I$.
For each $i \in I$,

- calculate
$$\Delta_i = \left\lceil \hat{\sigma}_i^2(n_i) \cdot \frac{n_z + B_z}{\hat{\sigma}_z^2(n_z)} \right\rceil;$$
- if $\Delta_i > n_i$, then take $(\Delta_i - n_i)$ observations.

Set $n = n + 1$ and $n_i = \max(n_i, \Delta_i)$; and update $\hat{\sigma}_i^2(n_i)$ for all $i \in I$ and $\hat{\lambda}^2$. Then go to **Calculation**.

Frazier (2014) recommends $B_z = 1$. The parameter $\eta_{|I|}$ needs to be chosen carefully so that the actual probability of correct selection is at least $1-\alpha$. In the next section, we derive some analytical results for the $\mathcal{DK}_1$ procedure and then discuss how to choose $\eta_{|I|}$.

4. Statistics for Screening

The canonical choice for fully sequential procedures is to use $\sum_{j=1}^{n}(X_{ij} - X_{\ell j})$ for every $i \neq \ell$ as observed statistics and to eliminate a system whenever the statistics exit a so-called continuation region defined by two parallel lines such as $(-a, a)$ for a constant $a > 0$ or a function $h(n) > 0$ such as $(-h(n), h(n))$. Kim and Nelson (2014) use a triangular shaped continuation region defined by a decreasing linear function $h(n)$. Note that traditional continuation regions are defined in a two-dimensional space. Our procedures use different statistics based on a quadratic form and our continuation region is an open cylinder.

Consider $x \in \mathbb{R}^s$ and $I \subset \{1, \ldots, k\}$ with $I = \{i_1, \ldots, i_s\}$. Furthermore let Γ represent the covariance matrix

of $(X_{i_1 j}, X_{i_2 j}, \ldots, X_{i_s j})^T$,

$$\Gamma = \begin{bmatrix} \sigma_{i_1}^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_{i_2}^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_{i_3}^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \sigma_{i_s}^2 \end{bmatrix}$$

and let V represent an $s-1$ by s matrix given by

$$V = \begin{bmatrix} 1 & 0 & \cdots & 0 & -1 \\ 0 & 1 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix}.$$

Then our statistic $\mathcal{S}_I(x)$ is defined as

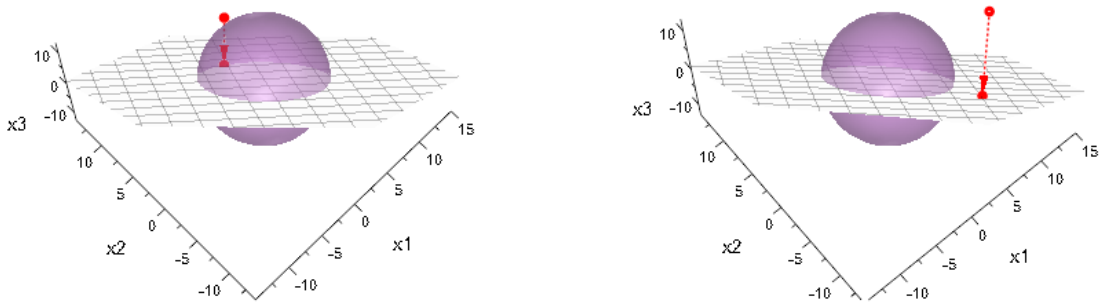
$$\mathcal{S}_I(x) \equiv (Vx)^T (V\Gamma V^T)^{-1} (Vx) = \begin{bmatrix} x_{i_1} - x_{i_s} \\ \vdots \\ x_{i_{s-1}} - x_{i_s} \end{bmatrix}^T (V\Gamma V^T)^{-1} \begin{bmatrix} x_{i_1} - x_{i_s} \\ \vdots \\ x_{i_{s-1}} - x_{i_s} \end{bmatrix} \tag{1}$$

and our continuation region is related to this quadratic form. From the definition of $\mathcal{S}_I$ it may seem that $\mathcal{S}_I$ is complicated to calculate and that it depends on the order in which its elements are listed. The following lemma is useful in deriving a simpler form of $\mathcal{S}_I(x)$ which allows us to argue that $\mathcal{S}_I(x)$ only depends on the set I, so not on the order of the elements in I. The proof is given in the appendix.

Lemma 1. *Suppose $x \in \mathbb{R}^s$ and $I \subset \{1, \ldots, k\}$ with $I = \{i_1, \ldots, i_s\}$. If $\Pi = \Gamma V^T (V\Gamma V^T)^{-1} V$, then*

$$\mathcal{S}_I(x) = \mathcal{S}_I(\Pi x).$$

The above lemma holds for Γ regardless of whether it has equal diagonal elements. The matrix Π is a (non-orthogonal) projection matrix with range $R = \{y \in \mathbb{R}^s : \sum_{i \in I} y_i/\sigma_i^2 = 0\}$ and null space $N = \{\alpha(1, \ldots, 1) : \alpha \in \mathbb{R}\}$, i.e., when Π is applied to a vector then a multiple of $(1, \ldots, 1)$ is subtracted from this vector so the result lies in R. It becomes an orthogonal projection matrix when $\sigma_i^2 = \sigma^2$ for all $i \in I$, since the null space N is orthogonal to the range R in that case. This lemma implies that the value of our statistic at any x equals the value of our statistic at the projected point on the plane determined by $\sum_{i \in I} y_i/\sigma_i^2 = 0$ (or $\sum_{i \in I} y_i = 0$ for equal variances). Since the null space and range do not change if the order of the elements in I changes, Lemma 1 shows that the quadratic form $\mathcal{S}_I$ remains the same if the indices are ordered differently, i.e., both in Γ and in x. In Section 5 this lemma is used to make the elimination decision depend on the IZ parameter δ only and not



(a) Projection inside ball: no elimination

(b) Projection outside ball: elimination

Figure 2: Projected points on plane $y_1 + y_2 + y_3 = 0$ and elimination rules. A ball (with radius 7 here) is also visible.

on the unknown mean parameter. Using the above lemma, we next derive a simpler form of $\mathcal{S}_I(x)$ when the variances are equal. As an aside, it may be tempting to think that $\mathcal{S}_I(x) = x^T V^T (V^T)^{-1} \Gamma^{-1} V^{-1} V x = x^T \Gamma^{-1} x$ in view of (1), but this is incorrect since V is not invertible.

Corollary 1. *Suppose $x \in \mathbb{R}^s$ and $I \subset \{1, \ldots, k\}$ with $I = \{i_1, \ldots, i_s\}$. If $\sigma_i^2 = \sigma^2$, then*

$$\mathcal{S}_I(x) = \frac{1}{\sigma^2} \frac{1}{|I|} \sum_{\substack{i<\ell \\ i,\ell \in I}} (x_i - x_\ell)^2 = \frac{1}{\sigma^2} \sum_{i \in I} (x_i - \bar{x})^2$$

where $\bar{x} = \frac{1}{s} \sum_{i \in I} x_i$.

Our elimination decision rule takes the form $\mathcal{S}_I(x) \geq r^2$ for $r \in \mathbb{R}^+$. Let x' denote the orthogonal projection of x on the plane with $\sum_{i=1}^s y_i = 0$. From Lemma 1, we know that $\mathcal{S}_I(x)$ is equal to $\mathcal{S}_I(x')$. As x' lies on the hyperplane $\{y : \sum_{i=1}^s y_i = 0\}$, we know that $\bar{x'} = 0$. From the second equality of $\mathcal{S}_I(x)$ in Corollary 1, it is easy to see that $\mathcal{S}_I(x')$ becomes simply the squared distance between x' and the origin. Therefore our elimination decision rule implies that no elimination occurs and sampling continues when the projected point x' is inside a sphere as in Figure 2(a); but one system with the smallest value is eliminated when the projected point x' is outside the sphere as in Figure 2(b).

One may wonder why our elimination rule only considers the largest set I but not any subset J, thinking that the screening statistics $\mathcal{S}_J(x_J)$ for $J \subseteq I$ may be larger than $\mathcal{S}_I(x_I)$. This would mean that even if an elimination does not occur with set I, an elimination might be possible for a subset J. However, the

following lemma implies that our elimination rule for the largest set I actually verifies elimination for all $2^{|I|}-1$ nonempty subsets $J \subseteq I$ by showing that we always get the largest screening statistics with set I.

Lemma 2. *Suppose $J \subseteq I \subseteq \{1,\dots,k\}$. Then $\mathcal{S}_J(x_J) \leq \mathcal{S}_I(x_I)$ for all $x \in \mathbb{R}^k$.*

5. Proofs and Approximations

This section presents an approximation for the probability of incorrect selection under $\mathcal{DK}_1$, which assumes known and equal variances σ^2. We use these approximations in lieu of possibly conservative bounds in order to choose the parameters $\eta_2,\dots,\eta_k$ of $\mathcal{DK}_1$, thus bypassing a main source of inefficiencies. In the course of the presentation, we explain how we choose the parameters $\eta_2,\dots,\eta_k$ of our procedure.

The event of incorrect selection can be partitioned according to when the best system is eliminated. If the best system is eliminated first, then we say that the level of elimination is 1. Similarly, if the second system to be eliminated is the best system, then we say that the level of elimination is 2. Thus, the possible levels of incorrect elimination are $1,\dots,k-1$. The key building block for our approximation scheme is an approximation for the probability of incorrect selection at the first elimination level, which we discuss in Section 5.1. Other levels of incorrect elimination are studied in Section 5.2. With this, we devise a procedure for choosing the parameter $\eta_{|I|}$ for $\mathcal{DK}_1$. We then explain how $\eta_2,\dots,\eta_k$ for $\mathcal{DK}_1$ are related to parameters for $\mathcal{DK}_2$ and $\mathcal{DK}_3$ in Section 5.3.

In the continuous analog of our problem, the discrete observation window is replaced with a continuous one. The analog of the random walk $\boldsymbol{X}_{\{1,\dots,k\}}(n)$ is $\sigma B(t)$, where $B(t)$ is a standard Brownian motion in $\mathbb{R}^k$ with drift $(\mu,\dots,\mu,\mu+\delta)\times 1/\sigma$. Throughout this section, we study this continuous problem as a proxy for the discrete problem, and the results we state for the $\mathcal{DK}_1$ algorithm are to be understood as for its continuous analog.

Lemma 3. *For fixed $\eta_k,\dots,\eta_2$ and $\ell \in \{2,\dots,k\}$, the probability of eliminination at level ℓ in $\mathcal{DK}_1$ is constant as a function of δ and σ. In particular, the probability of incorrect selection in $\mathcal{DK}_1$ does not depend on δ or σ.*

Proof. Consider $\sigma B(t)+\mu\mathbf{1}t+\delta wt$ instead of $\boldsymbol{X}_{\{1,\dots,k\}}(n)$, where $B(\cdot)$ is a standard Brownian motion in $\mathbb{R}^N$ with $B(0)=0$, $w=(0,\dots,0,1)$, and $\mathbf{1}=(1,\dots,1)$. Since the screening statistic uses the projection of this

process on the hyperplane $\{x : \sum_i x_i = 0\}$, i.e., the 'sample mean' is subtracted, we may assume without loss of generality that $\mu = 0$ and replace w by its projected version $v = (1/k, \ldots, 1/k, -(k-1)/k)$. Suppose that we are given some $x \in \mathbb{R}^N$ and an N-dimensional set S. We set

$$\tau_S = \inf\left\{t \geq 0 : \frac{\sigma^2}{\delta}x + \sigma B(t) + \delta v t \in \frac{\sigma^2}{\delta}S\right\},$$

so that

$$\begin{aligned}
\Pr(\tau_S < \infty) &= \Pr\left(\exists t' \geq 0 : \frac{\sigma^2}{\delta}x + \sigma B\left(\frac{\sigma^2}{\delta^2}t'\right) + \frac{\sigma^2}{\delta}vt' \in \frac{\sigma^2}{\delta}S\right) \\
&= \Pr\left(\exists t' \geq 0 : \frac{\sigma^2}{\delta}x + \frac{\sigma^2}{\delta}B(t') + \frac{\sigma^2}{\delta}vt' \in \frac{\sigma^2}{\delta}S\right) \\
&= \Pr\left(\exists t' \geq 0 : x + B(t') + vt' \in S\right),
\end{aligned}$$

where the first equality follows from rescaling time and the second from the Brownian scaling property. This argument extends to the hitting location, i.e., $\Pr(\tau_S < \infty, \frac{\sigma^2}{\delta}x + \sigma B(\tau_S) + \delta v \tau_S \in \frac{\sigma^2}{\delta}dy)$. In particular, the hitting location scales with σ^2/δ.

Elimination at level ℓ amounts to successively hitting appropriate regions of sets of the form

$$\begin{aligned}
\left\{x : \mathcal{S}_I(x_I) \geq \frac{k\sigma^2\eta^2}{(k-1)\delta^2}\right\} &= \left\{x : \frac{1}{\sigma^2}\sum_{i\in I}(x_i - \bar{x})^2 \geq \frac{k\sigma^2\eta^2}{(k-1)\delta^2}\right\} \\
&= \left\{x : \sum_{i\in I}(x_i - \bar{x})^2 \geq \frac{k\sigma^4\eta^2}{(k-1)\delta^2}\right\} \\
&= \frac{\sigma^2}{\delta}\left\{x : \sum_{i\in I}(x_i - \bar{x})^2 \geq \frac{k\eta^2}{k-1}\right\},
\end{aligned}$$

where we used Corollary 1. Such sets are of the form $(\sigma^2/\delta)S$, and the successive hitting locations scale with σ^2/δ. By the strong Markov property and the calculation in the first part of this proof, this means that the elimination probability does not depend on δ or σ. □

5.1 Immediate (Level 1) Elimination of the Best System

Our approximation for the probability of eliminating system k first is based on an asymptotic analysis as the number of systems k goes to infinity. Our results use the commonly employed idea of (i) considering the slippage configuration (SC) where $\mu_1 = \cdots = \mu_{k-1} = \mu_k - \delta = \mu$ and (ii) replacing the (discrete) Gaussian observation sequence with a (continuous) Brownian motion.

Throughout this section we use the following notation. For a given a vector $x \in \mathbb{R}^k$, we define

$$\mathrm{E}_k(x) = \frac{1}{k}\sum_{i=1}^{k} x_i, \quad \mathrm{Var}_k(x) = \frac{1}{k}\sum_{i=1}^{k} x_i^2 - \mathrm{E}_k(x)^2.$$

The $\mathcal{DK}$ algorithms require evaluating $\mathcal{S}_{\{1,\dots,k\}}$ at $\boldsymbol{X}_{\{1,\dots,k\}}(n)$, and by Lemma 1 this equals (up to $1/\sigma^2$) the squared norm of $\boldsymbol{X}_{\{1,\dots,k\}}(n) - \overline{\boldsymbol{X}_{\{1,\dots,k\}}(n)}$, which corresponds to $\sigma B(t) - \sigma \mathrm{E}_k(B(t))$. (We abuse notation and interpret subtraction of a constant as elementwise subtraction.) The following lemma specifies the probabilistic behavior of this process, and it is important to note that it is free of the unknown mean parameter μ.

Lemma 4. *$B(t) - \mathrm{E}_k(B(t))$ has drift $(-1/k, \dots, -1/k, (1-1/k)) \times \delta/\sigma$ and is a standard Brownian motion in the $(k-1)$-dimensional hyperplane*

$$H = \left\{ x \in \mathbb{R}^k : \sum_{i=1}^{k} x_i = 0 \right\}.$$

Proof. The claim that $B(t) - \mathrm{E}_k(B(t))$ takes values in H is evident. We can write $B(1) - \mathrm{E}_k(B(1)) = (\mathrm{id}_k - \mathbf{1}_k \mathbf{1}_k^T/k)B(1)$, where id_k is the $k \times k$ identity matrix and $\mathbf{1}_k$ is the $k \times 1$ vector of ones. Therefore, the covariance matrix of $B(1) - \mathrm{E}_k(B(1))$ is

$$(\mathrm{id}_k - \mathbf{1}_k \mathbf{1}_k^T/k) \times (\mathrm{id}_k - \mathbf{1}_k \mathbf{1}_k^T/k) = (\mathrm{id}_k - \mathbf{1}_k \mathbf{1}_k^T/k).$$

This matrix has one eigenvalues 0 (with corresponding eigenvector $\mathbf{1}_k$) and 1 (with corresponding eigenspace H). Therefore it acts as the identity on H and it is degenerate on the complement. □

Setting $r = \frac{\sigma \eta_k}{\delta_k}$, we define a $(k-1)$-dimensional sphere in H by

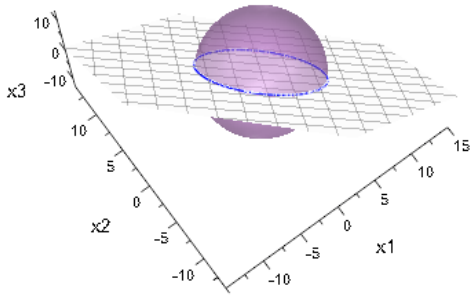
$$C = \left\{ x \in \mathbb{R}^k : \sum_{i=1}^{k} x_i = 0, \|x\| = r \right\}.$$

Elimination of the best system can be formulated as $B(t) - \mathrm{E}_k(B(t))$ hitting C in the region

$$E_k = \{x \in C : x_k = \min(x_1, \dots, x_k)\}.$$

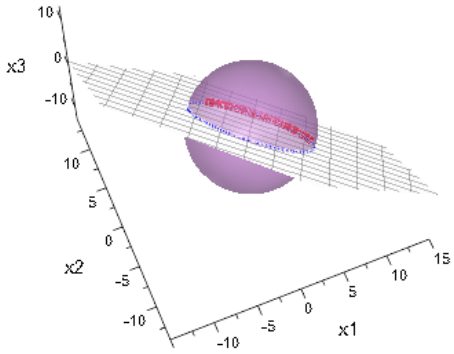
Plane H is shown in Figure 3 when $k = 3$. The blue curve in Figure 3(a) shows C when $k = 3$ and the red curve in Figure 3(b) shows E_k, which is a part of C divided by planes $x_1 = x_3$ and $x_2 = x_3$ as shown in Figure 3(c).

We now state the main result of this section.

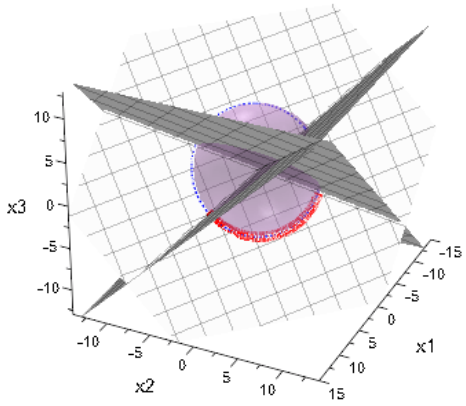


(a) Sphere C (circle here) on hyperplane H

$$C = \left\{x \in \mathbb{R}^k : \sum_{i=1}^k x_i = 0, \|x\| = r\right\}$$
$$E_k = \{x \in C : x_k = \min(x_1, \ldots, x_k)\}$$



(b) Region E_k (red) on hyperplane H



(c) Region E_k (red) with planes $x_1 = x_3$ and $x_2 = x_3$ on hyperplane H

Figure 3: Graphical depiction of C and E_k for $k = 3$.

Lemma 5. *Let $k \geq 3$. Suppose that $Z_1, \ldots, Z_k$ are iid standard normal. The probability that the process $B(t) - \mathrm{E}_k(B(t))$ first hits C in the part E_k where the best system k gets eliminated equals*

$$\frac{\int_{-r}^{r} e^{\frac{\delta_k}{\sigma} y} d_y \Pr(Z_k = \min(Z_1, \ldots, Z_k), r(Z_k - \mathrm{E}_k(Z)) \leq y \sqrt{(k-1)\mathrm{Var}_k(Z)})}{\left(\frac{\eta_k}{2}\right)^{-\nu} \Gamma(\nu+1) I_\nu(\eta_k)}, \tag{2}$$

where $\nu = (k-3)/2$, Γ stands for the Gamma function, and I_ν for the modified Bessel function of the first kind.

Proof. Writing ζ for the drift of $B(t) - \mathrm{E}_k(B(t))$, then the hitting place of $B(t) - \mathrm{E}_k(B(t))$ on C has density f with respect to the uniform distribution u_C on C. Here u_C should be interpreted as a volume element on C in the terminology of differential geometry, and by rotational invariance it has a 'simulation interpretation'

as the distribution of

$$X = \frac{r(Z_1 - \mathrm{E}_k(Z), \ldots, Z_k - \mathrm{E}_k(Z))}{\sqrt{k\mathrm{Var}_k(Z)}},$$

where Z is a standard normal vector in $\mathbb{R}^k$. The density f with respect to u_C is given by (e.g., Rogers and Pitman (1981))

$$f(x) = \frac{e^{\langle \zeta, x \rangle}}{\int_C e^{\langle \zeta, w \rangle} u_C(dw)}, \quad x \in C.$$

This distribution is known as the von Mises distribution.

According to Rogers and Pitman (1981), for any $\mu \in \mathbb{R}^k$ with $\sum_i \mu_i = 0$,

$$\int_C e^{\langle \mu, w \rangle} u_C(dw) = (\|\mu\| r/2)^{-\nu} \Gamma(\nu + 1) I_\nu(\|\mu\| r),$$

where $\nu = (k-3)/2$. Therefore, the denominator can be written as

$$\int_C e^{\langle \zeta, w \rangle} u_C(dw) = \left(\frac{\frac{\delta_k}{\sigma} r}{2} \right)^{-\nu} \Gamma(\nu + 1) I_\nu\left(\frac{\delta_k}{\sigma} r \right) = \left(\frac{\eta_k}{2} \right)^{-\nu} \Gamma(\nu + 1) I_\nu(\eta_k)$$

because $\|\zeta\| = \delta \sqrt{(k-1)/k}/\sigma = \delta_k/\sigma$ and $(\delta_k/\sigma) r = \eta_k$. Note that larger values of $B_k(t) - \mathrm{E}_k(B(t))$ are more likely than smaller values when the process hits C, which should be expected because system k is the best one.

The probability of eliminating the best system in level 1 equals

$$\begin{aligned}
\int_{E_k} f(x) u_C(dx) &= \mathrm{E}[\mathbb{1}(X \in E_k) f(X)] \\
&= \mathrm{E}[\mathbb{1}(X_k = \min(X_1, \ldots, X_k)) f(X)] \\
&= \frac{\mathrm{E}[\mathbb{1}(X_k = \min(X_1, \ldots, X_k)) e^{\langle \zeta, X \rangle}]}{\int_C e^{\langle \zeta, w \rangle} u_C(dw)} \\
&= \frac{\int_{-r}^{r} e^{\frac{\delta_k}{\sigma} y} d_y \Pr(X_k = \min(X_1, \ldots, X_k), \langle \zeta, X \rangle \le \frac{\delta_k}{\sigma} y)}{\int_C e^{\langle \zeta, w \rangle} u_C(dw)},
\end{aligned}$$

where X has a uniform distribution on C (see the beginning of this proof) and $\mathbb{1}$ stands for the indicator function. Since $\langle \zeta, x \rangle = \frac{\delta}{\sigma} x_k$ for $x \in H$, the sought probability equals

$$\frac{\int_{-r}^{r} e^{\frac{\delta_k}{\sigma} y} d_y \Pr(Z_k = \min(Z_1, \ldots, Z_k), \frac{\delta}{\sigma} r (Z_k - \mathrm{E}_k(Z)) / \sqrt{k\mathrm{Var}_k(Z)} \le \frac{\delta_k}{\sigma} y)}{\int_C e^{\langle \zeta, w \rangle} u_C(dw)},$$

as claimed. □

The preceding lemma yields a Monte Carlo method for calculating the probability of immediate elimination of the best system. Indeed, it states that this probability equals

$$\frac{\mathrm{E}\left[\exp\left(\eta_k \frac{Z_k - \mathrm{E}_k(Z)}{\sqrt{(k-1)\mathrm{Var}_k(Z)}}\right); Z_k = \min(Z_1, \ldots, Z_k)\right]}{\left(\frac{\eta_k}{2}\right)^{-\nu} \Gamma(\nu+1) I_\nu(\eta_k)}, \tag{3}$$

for iid standard normal $Z_1, \ldots, Z_k$. However, for large k, such a Monte Carlo method is not efficient and we instead approximate the level 1 probability (2) by replacing several of its components by asymptotic approximations. For instance, as $k \to \infty$, the random variables $\mathrm{E}_k(Z)$ and $\mathrm{Var}_k(Z)$ converge in distribution to 0 and 1, respectively, by the strong law of large numbers. The rate of convergence is relatively fast (order $1/\sqrt{k}$ by the central limit theorem). We, therefore, approximate those variables by their deterministic asymptotic approximations. The term with the minimum is slightly more complicated. Writing

$$c_k = \sqrt{2\log k} - \frac{\log\log k + \log(4\pi)}{2\sqrt{2\log k}},$$

$\min(Z_1, \ldots, Z_{k-1}) + c_{k-1}$ converges in distribution to 0. For example, see Example 3.3.29 in Embrechts, Kluppelberg and Mikosch (1997). The rate of convergence is relatively slow (order $1/\sqrt{2\log k}$), so we use an approximation based on the fact that

$$\sqrt{2\log k}(\min(Z_1, \ldots, Z_{k-1}) + c_{k-1})$$

converges in distribution to $-G$ where G is a Gumbel distributed random variable which is equal in distribution to $-\log(-\log(U))$ where U is standard uniformly distributed. Even when the central limit theorem is used for the sum instead of the law of large numbers, the minimum and sum are asymptotically independent (e.g., Chow and Teugels 1978). This motivates the approximation, for $y \in (-r, r)$,

$$\begin{aligned}
&\Pr(Z_k = \min(Z_1, \ldots, Z_k), r(Z_k - \mathrm{E}_k(Z)) \le y\sqrt{(k-1)\mathrm{Var}_k(Z)}) \\
&\quad\approx \Pr(Z_k \le \min(Z_1, \ldots, Z_{k-1}), rZ_k \le y\sqrt{(k-1)}) \\
&\quad\approx \Pr(Z_k \le -G/\sqrt{2\log k} - c_{k-1}, rZ_k \le y\sqrt{k-1}),
\end{aligned}$$

where Z_k and G are independent.

We are now ready to formulate our approximation for (2), and we first assume G is a given constant.

Lemma 6. *For fixed $a \in \mathbb{R}$, we have*

$$\begin{aligned}
&\int_{-r}^{r} e^{\frac{\delta_k}{\sigma} y} d_y \Pr(Z_k \le -a/\sqrt{2\log k} - c_{k-1}, rZ_k/\sqrt{k-1} \le y) \\
&= \exp\left(\frac{\eta_k^2}{2(k-1)}\right)\left[\Phi\left(\min\left(\max\left(-\sqrt{k-1}, \frac{-a}{\sqrt{2\log k}} - c_{k-1}\right), \sqrt{k-1}\right) - \frac{\eta_k}{\sqrt{k-1}}\right) - \Phi\left(-\sqrt{k-1} - \frac{\eta_k}{\sqrt{k-1}}\right)\right].
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal random variable.

Proof. Letting Y be a centered Gaussian variable with variance $r^2/(k-1)$. For any $\kappa \in \mathbb{R}$, we then have

$$\begin{aligned}
&\int_{-r}^{r} e^{(\delta_k/\sigma)y} d_y \Pr(Z_k \le \kappa, rZ_k/\sqrt{k-1} \le y) \\
&= \int_{-r}^{r\min(\max(-1,\kappa/\sqrt{k-1}),1)} e^{(\delta_k/\sigma)y} d\Pr(Y \le y) \\
&= \int_{-r}^{r\min(\max(-1,\kappa/\sqrt{k-1}),1)} e^{(\delta_k/\sigma)y} \frac{\sqrt{k-1}}{r\sqrt{2\pi}} \exp\left(-\frac{(k-1)y^2}{2r^2}\right) dy \\
&= e^{\frac{(\delta_k/\sigma)^2 r^2}{2(k-1)}} \frac{\sqrt{k-1}}{\sqrt{2\pi}r} \int_{-r}^{r\min(\max(-1,\kappa/\sqrt{k-1}),1)} \exp\left(-\frac{\left(y - \frac{(\delta_k/\sigma)r^2}{(k-1)}\right)^2}{2r^2/(k-1)}\right) dy \\
&= e^{\frac{\eta_k^2}{2(k-1)}} \frac{\sqrt{k-1}}{\sqrt{2\pi}r} \int_{-r}^{r\min(\max(-1,\kappa/\sqrt{k-1}),1)} \exp\left(-\frac{\left(y - \frac{(\delta_k/\sigma)r^2}{(k-1)}\right)^2}{2r^2/(k-1)}\right) dy \\
&= e^{\frac{\eta_k^2}{2(k-1)}} \left[\Phi\left(\min(\max(-\sqrt{k-1}, \kappa), \sqrt{k-1}) - \frac{(\delta_k/\sigma)r}{\sqrt{k-1}}\right) - \Phi\left(-\sqrt{k-1} - \frac{(\delta_k/\sigma)r}{\sqrt{k-1}}\right)\right] \\
&= e^{\frac{\eta_k^2}{2(k-1)}} \left[\Phi\left(\min(\max(-\sqrt{k-1}, \kappa), \sqrt{k-1}) - \frac{\eta_k}{\sqrt{k-1}}\right) - \Phi\left(-\sqrt{k-1} - \frac{\eta_k}{\sqrt{k-1}}\right)\right],
\end{aligned}$$

as claimed. □

In summary, we approximate the probability of first eliminating the best system by

$$\frac{\exp\left(\frac{\eta_k^2}{2(k-1)}\right)\left[\mathrm{E}\Phi\left(\min\left(\max\left(-\sqrt{k-1}, \frac{-G}{\sqrt{2\log k}} - c_{k-1}\right), \sqrt{k-1}\right) - \frac{\eta_k}{\sqrt{k-1}}\right) - \Phi\left(-\sqrt{k-1} - \frac{\eta_k}{\sqrt{k-1}}\right)\right]}{(\eta_k/2)^{-\nu}\Gamma(\nu+1)I_\nu(\eta_k)}. \tag{4}$$

The expectation in (4) can be estimated through either Monte Carlo by generating Gumbel random variates or numerical integration from 0 to 1 which is the range of a random number U. Both can be done fast but we use the latter method because it is faster and free of sampling error. In we explain in more detail how the numerical integration is performed.

Remark 1. *Shane Henderson communicated to us that the numerator in (3) can be written as*

$$\mathrm{E}\left[\exp\left(\eta_k \frac{\min(Z_1,\ldots,Z_k)-\mathrm{E}_k(Z)}{\sqrt{(k-1)\mathrm{Var}_k(Z)}}\right); Z_k = \min(Z_1,\ldots,Z_k)\right],$$

and since $\min(Z_1,\ldots,Z_k)$*,* $\mathrm{E}_k(Z)$*, and* $\mathrm{Var}_k(Z)$ *do not change when the elements of* Z *are permuted, this equals*

$$\frac{1}{k}\sum_{i=1}^{k}\mathrm{E}\left[\exp\left(\eta_k \frac{\min(Z_1,\ldots,Z_k)-\mathrm{E}_k(Z)}{\sqrt{(k-1)\mathrm{Var}_k(Z)}}\right); Z_i = \min(Z_1,\ldots,Z_k)\right] = \frac{1}{k}\mathrm{E}\left[\exp\left(\eta_k \frac{\min(Z_1,\ldots,Z_k)-\mathrm{E}_k(Z)}{\sqrt{(k-1)\mathrm{Var}_k(Z)}}\right)\right].$$

Using the approximations $\mathrm{E}_k(Z)\approx 0$*,* $\mathrm{Var}_k(Z)\approx 1$*,* $\min(Z_1,\ldots,Z_k)\approx -G/\sqrt{2\log(k)}-c_k$ *as before, the numerator in (3) can be approximated by*

$$\frac{1}{k}\mathrm{E}\left[\exp\left(-\eta_k\frac{G}{\sqrt{2(k-1)\log(k)}}-\frac{\eta_k c_k}{\sqrt{k-1}}\right)\right] = \frac{1}{k}e^{-\eta_k c_k/\sqrt{k-1}}\Gamma\left(1+\frac{\eta_k}{\sqrt{2(k-1)log(k)}}\right),$$

since $\mathrm{E}[e^{-\gamma G}] = \mathrm{E}[Y^\gamma] = \int_0^\infty x^\gamma e^{-x}dx = \Gamma(\gamma+1)$ *for a standard exponentially distributed random variable* Y*.*

We do not use this approximation in the remainder of this paper, since experiments have shown that it leads to higher PCS than (4).

5.2 Other Level Errors

For level ℓ errors for $\ell = 2,3,\ldots,k-1$, the number of survived systems $|I|$ is $|I| = k-\ell+1$ and it is natural to replace k with $|I|$ in (4) as follows:

$$\frac{\exp\left(\frac{\eta_{|I|}^2}{2(|I|-1)}\right)\left[\mathrm{E}\Phi\left(\min\left(\max\left(-\sqrt{|I|-1}, \frac{-G}{\sqrt{2\log|I|}}-c_{|I|-1}\right), \sqrt{|I|-1}\right)-\frac{\eta_{|I|}}{\sqrt{|I|-1}}\right)-\Phi\left(-\sqrt{|I|-1}-\frac{\eta_{|I|}}{\sqrt{|I|-1}}\right)\right]}{(\eta_{|I|}/2)^{-\nu}\Gamma(\nu+1)I_\nu(\eta_{|I|})} \tag{5}$$

where $\nu = (|I|-3)/2$. In our procedure, $\eta_{|I|}$ is calculated as the solution to (5) $= \beta_\ell$ for $0<\beta_\ell<\alpha$. We let $P_k(\ell/k,\beta_\ell)$ represent level ℓ error, the probability of incorrectly eliminating the best system at level ℓ when $\eta_{|I|}$ is calculated with target β_ℓ. Note that it does not depend on δ or σ by Lemma 3. The probability of incorrect selection (PICS) of $\mathcal{DK}_1$ is

$$\mathrm{PICS} = \sum_{\ell=1}^{k-1} P_k(\ell/k,\beta_\ell).$$

Let $\beta_0 = \alpha/(k-1)$. If $P_k(\ell/k,\beta_0)$ for $\ell = 1,\ldots,k-1$ are all approximately equal to β_0, then the overall PICS would be approximately equal to α. For large k, the analysis in Section 5.1 ensures that η_k, the solution

to (4) = β_0, would result in the level 1 error approximately equal to β_0. For other level errors, we do not have control over the error probability but we propose an approximation.

For the derivation of (4), it is critical that the starting point of the corresponding Brownian motion is the origin. For levels $\ell > 1$, we start at a random point from the previous level and thus we do not necessarily have $P_k(\ell/k,\beta_0) \approx \beta_0$ if we let $\eta_{|I|}$ be the solution to (5) = β_0, unless we discard all observations from previous levels. This is not desirable because too many observations would be wasted. Instead, we seek for a heuristic way to determine $\eta_{|I|}$ under the following assumption:

Assumption 3. *For* $0 < \beta_\ell < \alpha$, $\ell = 1, 2, \ldots, k-1$ *and* $\beta_0 = \alpha/(k-1)$,

1. $P_k(\ell/k,\beta_\ell) \approx \beta_\ell \cdot q_k(\ell/k)$*; and*

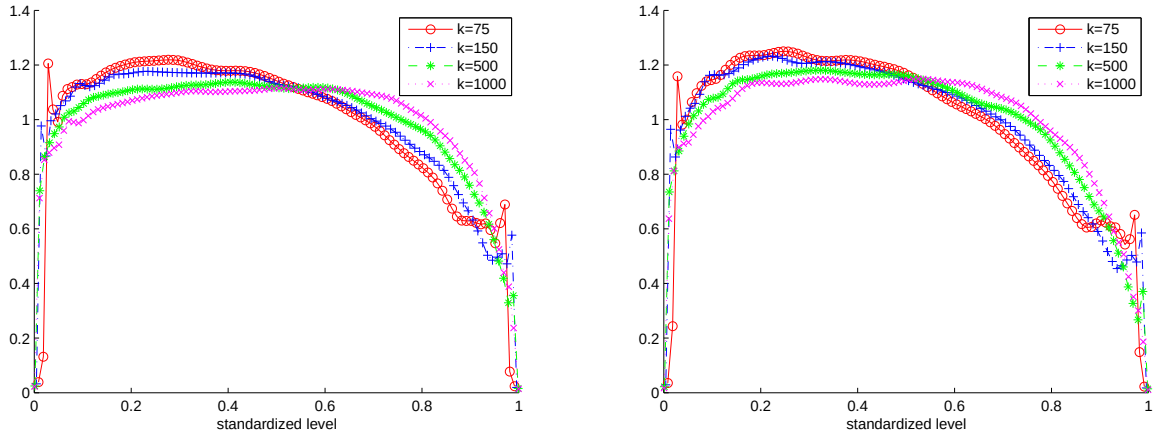
2. *If* $\beta_\ell = \beta_0$ *for all* ℓ*, then the probability that an incorrect selection (ICS) event occurs at standardized level* ℓ/k *is approximately* $\int_{\frac{\ell-1}{k-1}}^{\frac{\ell}{k-1}} g(w)dw$ *for a density function* $g(\cdot)$ *in [0,1].*

Assumption 3.1 is effectively a first-order Taylor approximation under appropriate differentiability assumptions because $\lim_{\beta \downarrow 0} P_k(\ell/k,\beta) = 0$. This assumption implies that for small β_ℓ, the level ℓ error is approximately linear in β_ℓ. For example, if β_ℓ decreases in half for level ℓ, then the level ℓ error is expected to be cut in half.

We have empirical evidence for Assumption 3.2. To test Assumption 3.2, we made one million replications and recorded standardized levels where ICS occurred for each experimental setting. Then a kernel density estimator is fitted to the data using Matlab with a normal kernel and support $[0, 1]$. A bandwidth was chosen by Matlab, which is known to be optimal for the normal kernel. Figure 4 shows kernel density estimates of standardized levels ℓ/k of having ICS for $k = 75, 150, 500$ and 1000 and $\alpha = 0.05$ and 0.10 when $\delta = 0.3$ and $\sigma^2 = 1$. Note that the specific choice of δ and σ does not matter in view of Lemma 3. From the figure, one can see that the shapes of kernel estimates for various k are similar.

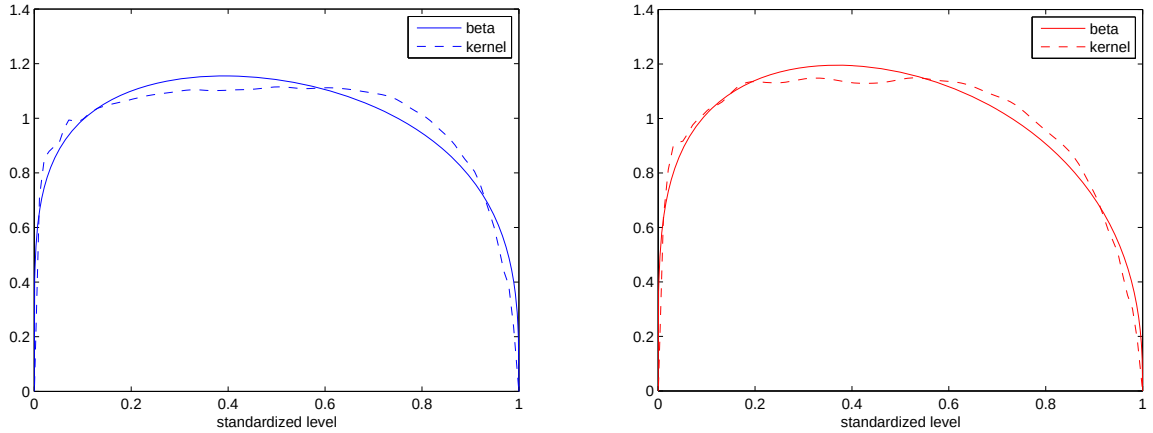
To approximate $g(w)$ for $0 < w < 1$, we use $k = 1000$ rather than $k = 75$ because $k = 75$ gives sparse points in $[0, 1]$. In addition, as kernel estimates are not stable on the boundary 0 and 1, we further fit it using a beta distribution to get a smooth function especially close to the boundary points, assuming

$$g(w) \approx \frac{1}{\text{Beta}(A, B)} \, w^{A-1} \, (1-w)^{B-1} \quad \text{for} \quad 0 < w < 1 \text{ and } A, B \in \mathbb{R}$$



(a) $\alpha = 0.10$ (b) $\alpha = 0.05$

Figure 4: Kernel density estimates on standardized levels where an incorrect selection occurs for various k when $\delta = 0.3$, and $\sigma^2 = 1$.



(a) $\alpha = 0.10$ and $g(w) \propto w^{0.19805}(1-w)^{0.30662}$ (b) $\alpha = 0.05$ and $g(w) \propto w^{0.2317}(1-w)^{0.39658}$

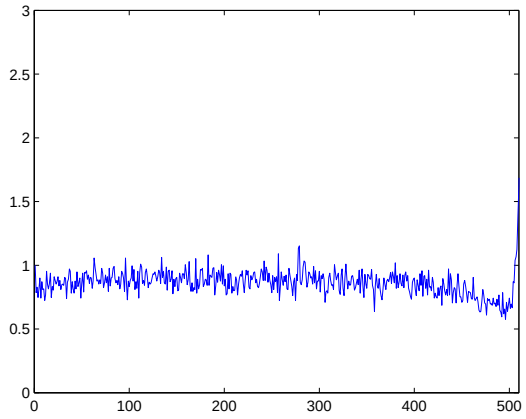
Figure 5: Kernel density estimates and beta density estimates for $g(w)$ when $k = 1000$ and $\delta = 0.3$.

where $\text{Beta}(A, B) = \int_0^1 t^{A-1}(1-t)^{B-1}dt$. Figure 5 shows the fitted beta densities for $\alpha = 0.05$ and $\alpha = 0.10$, respectively. Both beta densities are very similar.

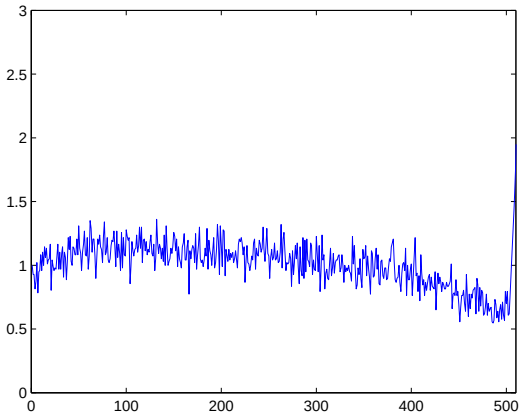
Once $g(w)$ is approximated, $\eta_{|I|}$ which ensures the probability of correct selection (PCS) of $\mathcal{DK}_1$ can be calculated as follows:

Step 1: Calculate the constant m_ℓ as follows:

$$m_\ell = \frac{G\left(\frac{\ell}{k-1}\right) - G\left(\frac{\ell-1}{k-1}\right)}{G\left(\frac{1}{k-1}\right)}$$



(a) $1-\alpha = 0.90, \text{PCS} = 0.907$



(b) $1-\alpha = 0.95, \text{PCS} = 0.950$

Figure 6: Ratios between $\hat{P}_k(\ell/k, \beta_0/m_\ell)$ and β_0 for $\alpha = 10\%$ and 5% when $k = 512$, $\delta = 0.3$ and $\sigma^2 = 1$.

where $G(w) = \int_0^w g(t)dt$, the cdf of $g(w)$.

Step 2: Set $\beta_\ell = \frac{\beta_0}{m_\ell}$ and calculate $\eta_{|I|}$ from $(5) = \beta_\ell$.

The relative magnitude between $G\left(\frac{1}{k-1}\right)$ and $G\left(\frac{\ell}{k-1}\right) - G\left(\frac{\ell-1}{k-1}\right)$ can be interpreted as the relative magnitude between level 1 error and level ℓ error when $\eta_{|I|}$ is calculated from $(5) = \beta_0$. As we know that level 1 error is β_0,

$$G\left(\frac{1}{k-1}\right) : G\left(\frac{\ell}{k-1}\right) - G\left(\frac{\ell-1}{k-1}\right) \approx \beta_0 : P_k(\ell/k, \beta_0).$$

Therefore the level ℓ error is expected to be inflated by $m_\ell = \frac{G(\frac{\ell}{k-1}) - G(\frac{\ell-1}{k-1})}{G(\frac{1}{k-1})}$ compared to target β_0. If we adjust $\beta_\ell = \beta_0/m_\ell$, then $P_k(\ell/k, \beta_\ell) \approx \beta_0$ by Assumption 3.1, which in turns implies that the overall PICS is approximately equal to α.

Figure 6 shows estimated level errors $\hat{P}_k(\ell/k, \beta_0/m_\ell)$ for $\alpha = 5\%$ and 10% when $k = 512$ with $\delta = 0.3$ and $\sigma^2 = 1$ and one million replications. One can see that the level errors do not show a beta shape as in Figure 5. Instead the ratios between level errors and β_0 fluctuate around one for the two values of α, which empirically supports Assumption 3.1. Also, it shows that the function we found $g(w)$ for $\alpha = 5\%$ and $k = 1000$ seems to work well for other popular choices of α, including $\alpha = 10\%$.

To search for $\eta_{|I|}$ for given target β_ℓ, a bisection search is used and this requires estimating the expectation in (5). Instead of using Monte Carlo by generating Gumbel random variates G, we use numerical integration

in the range of a standard uniform random variable U, $[0,1]$ using one million intervals on function $f(u)$ defined as follows:

$$f(u) \equiv \Phi\left(\min\left(\max\left(-\sqrt{|I|-1}, \frac{\log(-\log u)}{\sqrt{2\log|I|}} - c_{|I|-1}\right), \sqrt{|I|-1}\right) - \frac{\eta_{|I|}}{\sqrt{|I|-1}}\right) \quad \text{for} \quad 0 < u < 1,$$

$f(0) \equiv \lim_{u\to 0} f(u)$ and $f(1) \equiv \lim_{u\to 1} f(u)$. When $u \to 0$ or $u \to 1$, $\log(-\log U)$ converges to either ∞ or $-\infty$ but the minimum and maximum functions inside $\Phi(\cdot)$ in $f(u)$ ensure a finite number is returned. Then

$$\begin{aligned} &\mathrm{E}\Phi\left(\min\left(\max\left(-\sqrt{|I|-1}, \tfrac{\log(-\log u)}{\sqrt{2\log|I|}} - c_{|I|-1}\right), \sqrt{|I|-1}\right) - \tfrac{\eta_{|I|}}{\sqrt{|I|-1}}\right) \\ &\approx \tfrac{1}{1000000}\left[\tfrac{1}{2}f(0) + \textstyle\sum_{j=1}^{999999} f(j/1000000) + \tfrac{1}{2}f(1)\right]. \end{aligned}$$

The parameter $\eta_{|I|}$ is searched using the deterministic bisection method when the numerical integration is used. Note that the above approximation is based on the assumption that $|I|$ is large. When $|I|$ is small, say $|I| < 10$, (5) does not work well. Instead we use (3) which requires a Monte Carlo simulation with $|I|$ number of iid standard normal random variables. When a Monte Carlo simulation is used, there is a chance that a deterministic bisection method may fail due to simulation error. Therefore when $|I| < 10$, we employee a probabilistic bisection algorithm (Section 1.5 of Waeber (2013)) is used. The stochastic bisection algorithm stops when the returned median of a posterior distribution in the current search iteration is within 0.001 of the median from the previous iteration. A sequential test of power one which determines the sign of the objective function is implemented with parameters $r_0 = 50000$ and $\gamma = 0.01$. The sequential test stops either when m reaches 1000 or when its test statistics exit $(-k_m, k_m)$ where k_m is from equation (B.6) of Waeber (2013).

Since $\eta_{|I|}$ only depends on α and k, a table can be made for popular choices of α such as 5% and 10% and $k = 2, 3, \ldots, 10000$. Then the values of $\eta_{|I|}$ can be read from the table while running our procedures. Table A.1 in the appendix shows the values of $\eta_{|I|}$ for a few selected values of k when $\alpha = 10\%$.

5.3 Justification of Procedures for Unknown Variances

In this subsection, we discuss why $\mathcal{DK}_2$ and $\mathcal{DK}_3$ should be expected to work for unknown variances as well. For unknown variances, it is natural to replace variance parameters in $\mathcal{DK}_1$ to their estimated values. In general, it is not sufficient to replace the variance parameter with its estimated value to keep the statistical

validity. It is critical to account for the variability in the estimated parameter especially when variances are estimated only once based on an initial n_0 observations. Kim and Nelson (2006) and Wang and Kim (2011) show that if variance estimators are updated on the fly in a procedure as more observations are obtained, then the procedure converges to the known variance case under some appropriate asymptotic regime. In the light of these results, we employ a variance updating scheme in $\mathcal{DK}_2$ and $\mathcal{DK}_3$ to avoid the difficulty of accounting for the variability in the estimated variance parameters but without any claim for the asymptotic validity in this paper.

When the decision maker believes that the variances across systems are equal (but unknown), then the natural estimator for σ^2 is the pooled variance estimator

$$\hat{\sigma}_p^2(n) = \frac{1}{|I|}\sum_{i\in I}\hat{\sigma}_i^2(n).$$

As we update $\hat{\sigma}_p^2(n)$ as more observations become available, the estimator converges to σ^2 and thus it is expected that $\mathcal{DK}_2$ works similarly to $\mathcal{DK}_1$.

When variances are unknown and unequal, we use similar arguments as in Frazier (2014). Let $n_i = \gamma\sigma_i^2 n$ for some $\gamma > 0$ and thus the number of samples obtained by stage n for system i is proportional to its variance σ_i^2. Then

$$\frac{\sum_{j=1}^{n_i} X_{ij}}{\gamma\sigma_i^2} \sim N\left(\frac{n_i}{\gamma\sigma_i^2}\mu_i, \frac{n_i}{\gamma^2\sigma_i^2}\right) = N\left(n\mu_i, \frac{n}{\gamma}\right) \approx B_{(\mu_i,1/\gamma)}(t)$$

where $B_{(\mu_i,1/\gamma)}(t)$ is a Brownian motion with drift μ_i and variance $1/\gamma$. The $\frac{\sum_{j=1}^{n_i} X_{ij}}{\gamma\sigma_i^2}$ have equal variance as long as $n_i = \gamma\sigma_i^2 n$ and thus we can apply $\mathcal{DK}_1$ to $\frac{\sum_{j=1}^{n_i} X_{ij}}{\gamma\sigma_i^2}$. Note that when $n_i = \gamma\sigma_i^2 n$,

$$n_i\lambda^2 = n_i\frac{\sum_{i\in I}\sigma_i^2}{\sum_{i\in I} n_i} = \sigma_i^2$$

where

$$\lambda^2 = \frac{\sum_{i\in I}\sigma_i^2(n_i)}{\sum_{i\in I} n_i}.$$

Then

$$\frac{\sum_{j=1}^{n_i} X_{ij}}{\gamma\sigma_i^2} = \frac{\sum_{j=1}^{n_i} X_{ij}}{\gamma n_i\lambda^2} = \frac{W_i(n)}{\gamma\lambda^2}.$$

Finally, the screening rule in $\mathcal{DK}_1$ is

$$\frac{\sum_{i\in I}\left(\frac{W_i(n)}{\gamma\lambda^2} - \frac{1}{|I|}\sum_{i\in I}\frac{W_i(n)}{\gamma\lambda^2}\right)^2}{1/\gamma} \geq \frac{1}{\gamma}\left(\frac{\eta_{|I|}}{\delta_{|I|}}\right)^2$$

which is equivalent to

$$\frac{1}{\lambda^4}\sum_{i\in I}\left(W_i(n)-\frac{1}{|I|}\sum_{i\in I}W_i(n)\right)^2 \geq \left(\frac{\eta_{|I|}}{\delta_{|I|}}\right)^2$$

or

$$\frac{1}{\lambda^2}\sum_{i\in I}\left(W_i(n)-\frac{1}{|I|}\sum_{i\in I}W_i(n)\right)^2 \geq \left(\frac{\lambda\cdot\eta_{|I|}}{\delta_{|I|}}\right)^2 \tag{6}$$

When λ^2 is replaced with its estimator $\hat{\lambda}^2$ in (6), we get the same elimination rule in the $\mathcal{DK}_3$ procedure, which is

$$\mathcal{S}_I''(\boldsymbol{W}_I(n)) \geq \left(\frac{\hat{\lambda}\cdot\eta_{|I|}}{\delta_{|I|}}\right)^2.$$

6. Experiments

In this section, we compare the performance of $\mathcal{DK}$ procedures with KN and BIZ. For unknown variances, we use the KN procedure as originally described in Kim and Nelson (2001) with $c = 1$ and $n_0 = 30$ and Algorithm 2 of Frazier (2014) with $B_z = 1$ and $n_0 = 30$. For known variances, we use KN with $h^2 = 2\eta$ where $\eta = -\ln\left(2\frac{\alpha}{k-1}\right)$ and $n_0 = 1$, which is same as the $\mathcal{P}$ procedure in Wang and Kim (2011), and Algorithm 1 of Frazier (2014). Throughout this section, KN and BIZ refer procedures for known variances while KN-UNK and BIZ-UNK refer procedures for unknown variances.

The number of systems k varies over

$$k \in \{2,3,4,5,6,7,8,16,32,64,128,256,512,1024,2048,4096,8192\}.$$

For the mean, we consider two mean configurations, namely slippage configuration (SC) and monotonic decreasing mean configuration (MDM); and for variances, we consider three variance configurations called Equal, INC, and DEC. Thus we have total six configurations: SC-Equal, MDM-Equal, SC-INC, SC-DEC, MDM-INC and MDM-DEC. We use same parameter settings for mean, variances, δ and α as in Frazier (2014). Table 1 gives all six mean-variance configurations and other parameter settings.

When calculating $\eta_{|I|}$ for $\mathcal{DK}$ procedures, we take logs to avoid numerical overflows and underflows in the denominator, since the Gamma term can be very large and the Bessel term can be very small. When $\ell = k-1$ or only two systems are survived, we use $\eta_2 = -\ln(2\beta_\ell)$.

Table 1: Mean and variance configurations

Configuration	Means	Variances	δ	α
SC-Equal	$\mu = [\delta, 0, \ldots, 0]$	$\sigma^2 = 100$	1	0.1
MDM-Equal	$\mu_i = -\delta i$	$\sigma^2 = 100$	1	0.1
SC-INC	$\mu = [\delta, 0, \ldots, 0]$	$\sigma_i^2 = 25\left(1 + 3\frac{i-1}{k-1}\right)^2$	1	0.1
SC-DEC	$\mu = [\delta, 0, \ldots, 0]$	$\sigma_i^2 = 25\left(1 + 3\frac{k-i}{k-1}\right)^2$	1	0.1
MDM-INC	$\mu_i = -\delta i$	$\sigma_i^2 = 25\left(1 + 3\frac{i-1}{k-1}\right)^2$	1	0.1
MDM-DEC	$\mu_i = -\delta i$	$\sigma_i^2 = 25\left(1 + 3\frac{k-i}{k-1}\right)^2$	1	0.1

The nominal confidence level is set to $1 - \alpha = 0.9$. Estimated probability of correct selection (PCS) and an average number of observations per system until a decision is made (REP/k) are reported based on 10,000 macro replications. Standard errors for estimated PCS are approximately 0.003.

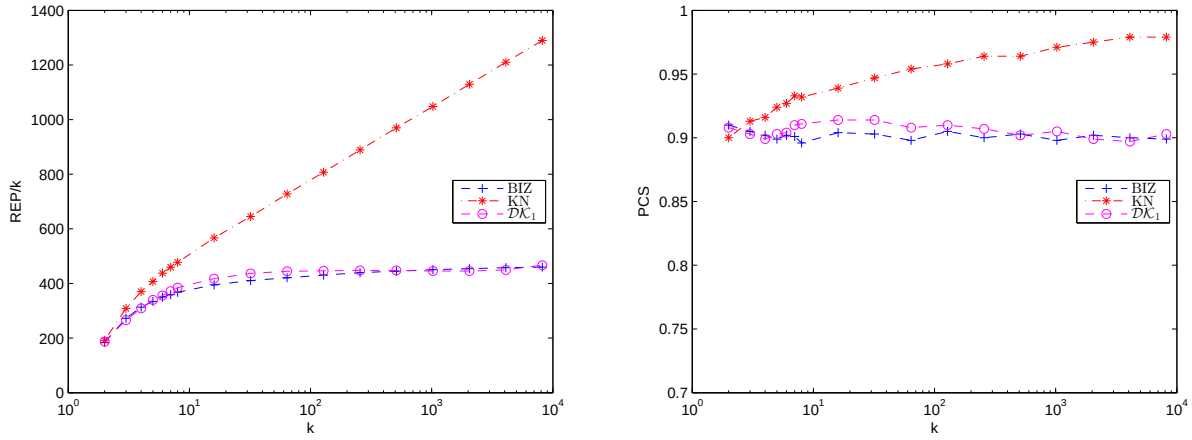
6.1 $\mathcal{DK}_1$ with Known and Equal Variances

When variances are known and equal, we compare $\mathcal{DK}_1$ with KN and BIZ. Figure 7 shows REP/k and PCS under SC and MDM configurations. Procedure $\mathcal{DK}_1$ significantly outperforms KN under both SC and MDM. When k is large, $\mathcal{DK}_1$ is more than three times better than KN in terms of REP/k. On the other hand, the performances of BIZ and $\mathcal{DK}_1$ are very similar under the slippage configuration in terms of both REP/k and PCS. When k is small, $\mathcal{DK}_1$ spends a slightly more number of observations than BIZ but its probability of correct selection is slightly higher than BIZ. For large k, their performances are very close in both measures. Under the monotonic decreasing mean configuration, $\mathcal{DK}_1$ achieves PCS greater than the nominal value 90% and clearly outperforms KN. However, BIZ achieves PCS close to the nominal value 90% than $\mathcal{DK}_1$ and spends slightly fewer but very similar number of observations than $\mathcal{DK}_1$.

6.2 $\mathcal{DK}_2$ and $\mathcal{DK}_3$ with Unknown but Equal Variances

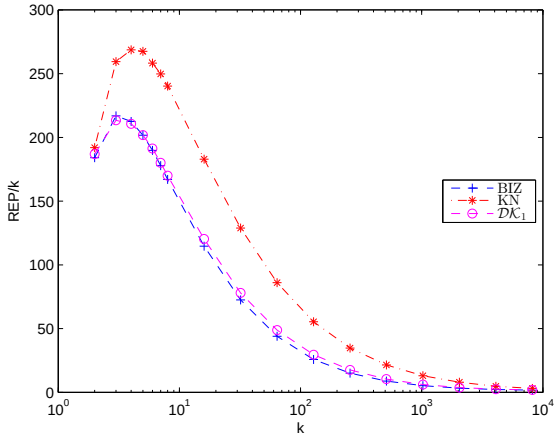
When variances are unknown but a decision maker knows that variances across systems are equal, $\mathcal{DK}_2$ or $\mathcal{DK}_3$ can be used.

Figure 8 compares performances of $\mathcal{DK}_2$ with those of KN-UNK and BIZ-UNK. As in the case of known and equal variances, $\mathcal{DK}_2$ significantly outperforms KN-UNK. Compared to BIZ-UNK, $\mathcal{DK}_2$ achieves slightly higher PCS and spends fewer number of observations for large k under the slippage configurations.

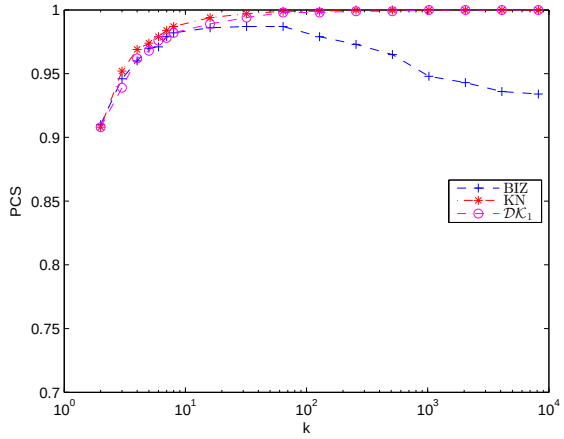


(a) SC-REP

(b) SC-PCS



(c) MDM-REP



(d) MDM-PCS

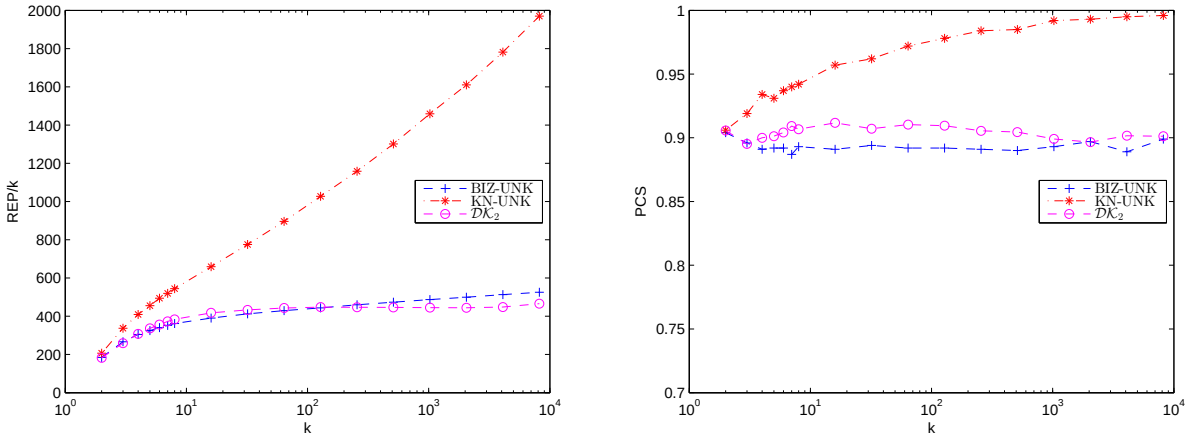
Figure 7: REP/k and PCS for $\mathcal{DK}_1$ when variances are known and equal with $1-\alpha = 0.9$

Then under the monotonic decreasing configuration, PCS is higher in $\mathcal{DK}_2$ and uses slightly more observations for small k and then similar number of observations for large k.

In reality, it is impossible to know in advance whether variances across systems are equal. In fact, equal variances across systems rarely hold. Thus we also consider $\mathcal{DK}_3$. Our experiments show that $\mathcal{DK}_3$ actually spends slightly fewer observations than $\mathcal{DK}_2$ while achieving similar PCS. Figure 9 compares $\mathcal{DK}_3$ with KN-UNK and BIZ-UNK. Graphs in Figure 9 show similar tendency as those in Figure 8.

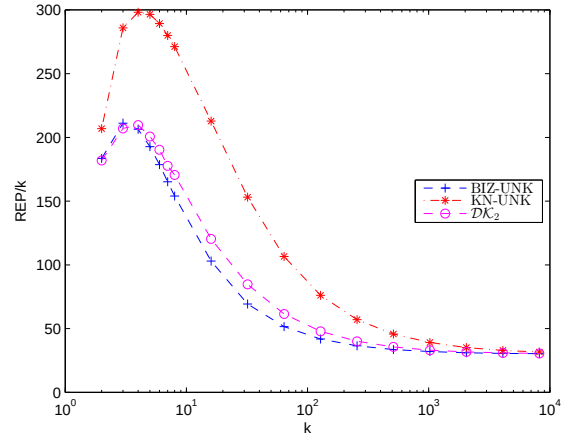
6.3 $\mathcal{DK}_3$ with Unknown and Unequal Variances

Finally, we consider unknown and unequal variances. Figure 10 compares the three procedures under the slippage configuration with increasing and decreasing variances while Figure 11 compares them under the

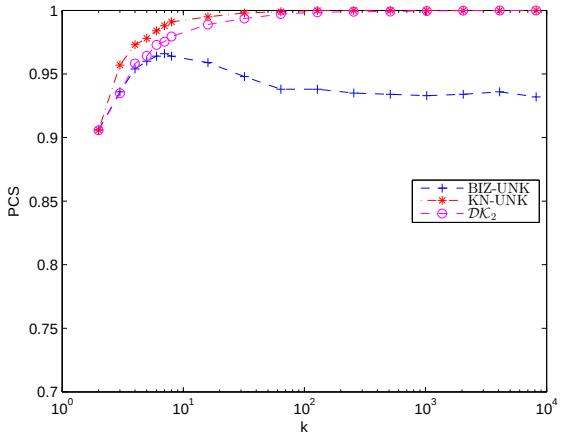


(a) SC-REP

(b) SC-PCS



(c) MDM-REP



(d) MDM-PCS

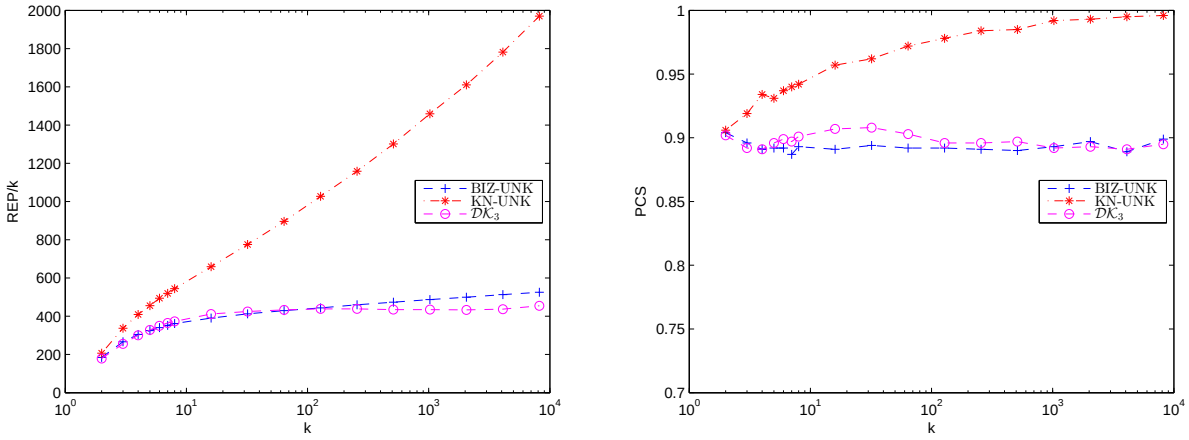
Figure 8: REP/k and PCS for $\mathcal{DK}_2$ when variances are unknown but equal with $1-\alpha = 0.9$

MDM configuration with increasing and decreasing variances.

The efficiency of $\mathcal{DK}_3$ compared to KN-UNK is more obvious. When $k = 8192$, $\mathcal{DK}_3$ is more than four times better than KN-UNK under SC-INC and six times better under SC-DEC in terms of REP/k while achieving PCS close to 90%. $\mathcal{DK}_3$ spends up to 30% fewer observations than BIZ-UNK under the slippage configuration.

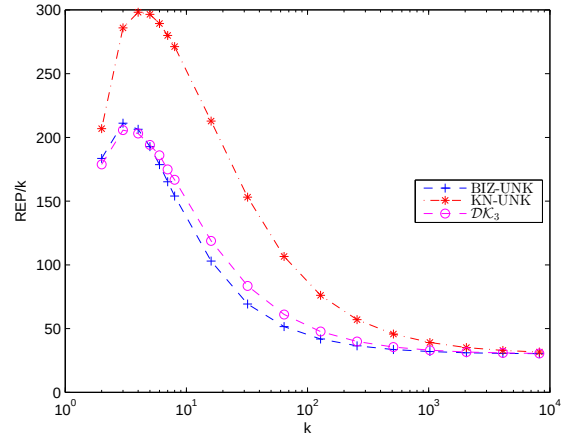
Interestingly, under the MDM configuration with increasing variances, $\mathcal{DK}_3$ significantly outperforms both KN-UNK and BIZ-UNK, showing up to 63% savings in the number of observations compared to BIZ-UNK. But $\mathcal{DK}_3$ uses slightly more observations than BIZ-UNK under decreasing variances.

Overall, $\mathcal{DK}$ procedures achieve PCS close to the nominal value for all settings we tested and they

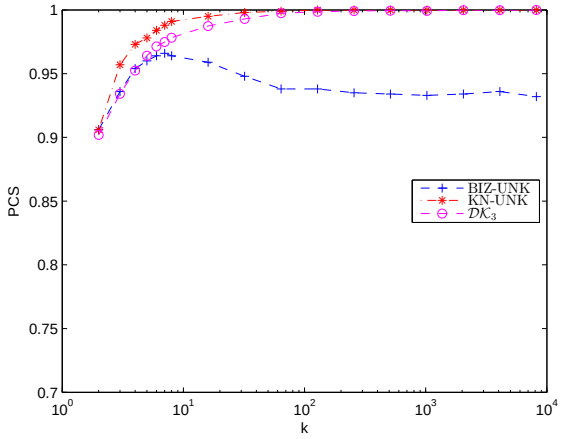


(a) SC-REP

(b) SC-PCS



(c) MDM-REP



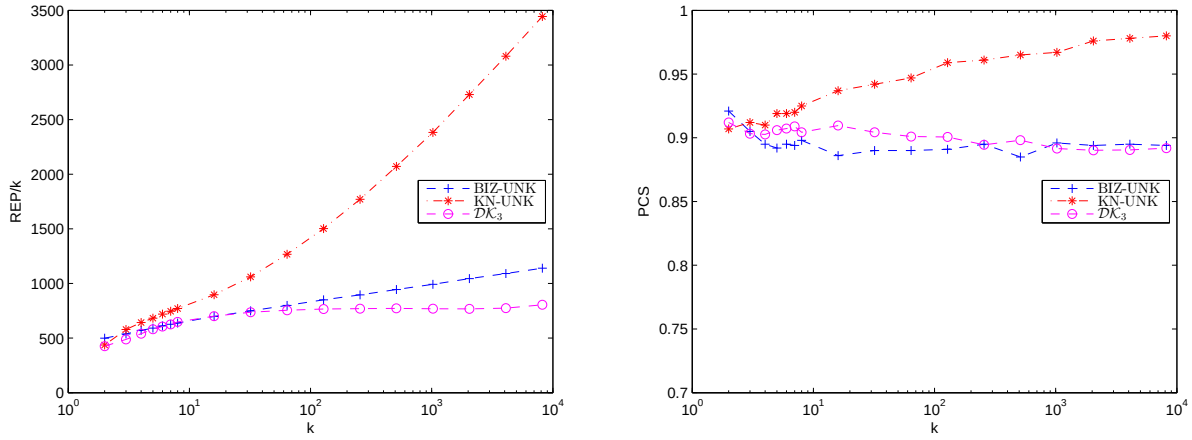
(d) MDM-PCS

Figure 9: REP/k and PCS for $\mathcal{DK}_3$ when variances are unknown but equal with $1-\alpha = 0.9$

outperform KN significantly while performing similarly to BIZ under easy mean configurations but outperforming it under difficult mean configurations especially with unknown and unequal variances.

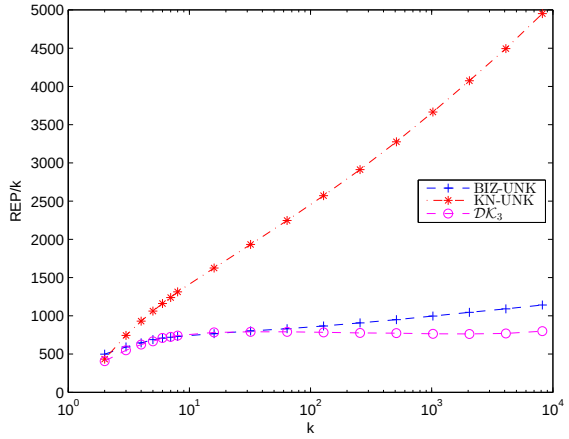
7. Conclusions

We present new fully-sequential procedures whose continuation regions are derived exploiting the properties of multidimensional Brownian motions, which is the first work in the literature. Our procedures deliver a probability of correct selection close to the nominal level. Compared to the existing state-of-art fully-sequential IZ procedure KN, the proposed procedures show a tight worst-case probability of incorrect selection under the slippage configuration and significant savings in the number of observations needed until

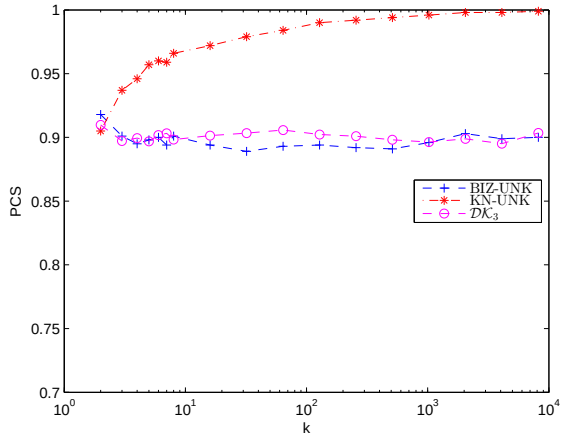


(a) SC-INC-REP

(b) SC-INC-PCS



(c) SC-DEC-REP



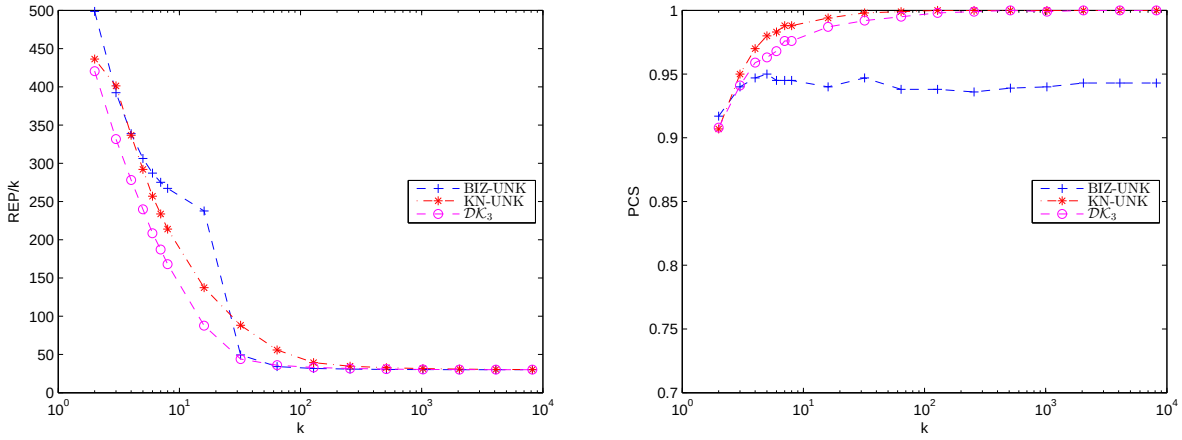
(d) SC-DEC-PCS

Figure 10: REP/k and PCS for $\mathcal{DK}_3$ when variances are unknown and unequal with $1-\alpha = 0.9$

a decision is made. Compared to BIZ, our procedures perform better for a large number of systems under difficult mean configurations, but tend to spend slightly more observations for small k but similar number of observations for large k under easier mean configurations except the increasing-variances case.

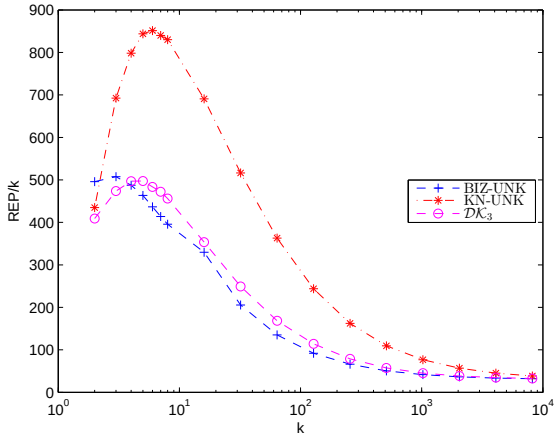
Acknowledgements

This work is supported by the National Science Foundation under grant CMMI-1131047. The authors would like to thank Seunghan Lee for his insight for Lemma 1. The authors appreciate Peter Frazier for his code and helpful comments and Barry Nelson for his helpful comments.

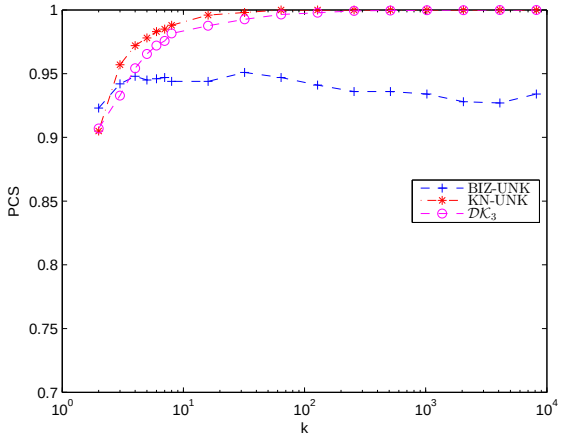


(a) MDM-INC-REP

(b) MDM-INC-PCS



(c) MDM-DEC-REP



(d) MDM-DEC-PCS

Figure 11: REP/k and PCS for $\mathcal{DK}_3$ when variances are unknown and unequal with $1-\alpha = 0.9$

References

Branke, J., Chick, S. E., and Schmidt, C. 2007. "Selecting a selection procedure". *Management Science* 53(12):1916-1932.

Chick, S. E. 2006. Subjective Probability and Bayesian Methodology. In *Handbooks in Operations Research and Management Science: Simulation*, edited by S. G. Henderson and B. L. Nelson. Oxford: Elsevier Science.

Chen, C.-H., S. E. Chick, L. H. Lee, N. A. Pujowidianto. 2014. "Ranking and Selection: Efficient Simulation Budget Allocation". In *Handbook of Simulation Optimization*, edited by M. C. Fu. Springer:NY.

Chen, C.-H., and L. H. Lee. 2011. *Stochastic Simulation Optimization: An Optimal Computing Budget Al-*

location (System Engineering and Operations Research), Vol 1. Singapore: World Scientific Publishing Company.

Chow, T. L., and J. L.Teugels. 1978. The Sum and the Maximum of I.I.D. Random Variables. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics*, edited by P. Mandl and M. Huskova, 81-92. New York: North-Holland.

Dieker, A. B., and S.-H. Kim. 2012. Selecting the Best by Comparing Simulated Systems in a Group of Three When Variances are Known and Unequal. In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 1-7. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Dieker, A. B., and S.-H. Kim 2014. "A Fully Sequential Procedure for Known and Equal Variances Based on Multivariate Brownian Motion". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 3749-3760. Piscataway, New Jersey: IEEE.

Embrechts, P., C. Kluppelberg, and T. Mikosch. 1997. *Modelling Extremal Events for Insurance and Finance*. New York: Springer.

Frazier, P. 2014. A Fully Sequential Elimination Procedure for Indifference-Zone Ranking and Selection with Tight Bounds on Probability of Correct Selection. *Operations Research* 62(4):926-942.

Hong, L. J., B. L. Nelson, J. Xu. 2014. "Discrete Optimization via Simulation". In *Handbook of Simulation Optimization*, edited by M. C. Fu. Springer:NY.

Kim, S.-H., and A. B. Dieker. 2011. Selecting the Best by Comparing Simulated Systems in a Group of Three. In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu. 4217-4226. Piscataway, New Jersey: IEEE.

Kim, S.-H., and B. L. Nelson. 2001. A Fully Sequential Procedure for Indifference-Zone Selection in Simulation. *ACM Transactions on Modeling and Computer Simulation* 11(3):251-273.

Kim, S.-H., and B. L. Nelson. 2006. "On the Asymptotic Validity of Fully Sequential Selection Procedures for Steady-State Simulation". *Operations Research* 54:475-488.

Nelson, B. L., J. Swann, D. Goldsman, and W. Song. 2001. "Simple Procedures for Selecting the Best

Simulated System when the Number of Alternatives is Large". *Operations Research* 49(6):950-963.

Powell, W. B. and Ryzhov, I. O. 2012. "Ranking and selection". In Chapter 4 in *Optimal Learning*, pages 71-88. John Wiley and Sons.

Rinott, Y. 1978. "On two-stage selection procedures and related probability inequalities". *Comm. Statist.-Theory and Methods* 7(8):799-811.

Rogers, L., and J. W. Pitman. 1981. Markov Functions. *The Annals of Probability* 9:573-582.

Waeber, R., P. I. Frazier, and S. G. Henderson. 2011. "A Bayesian Approach to Stochastic Root Finding". In *Proceedings of the 2011 Winter Simulation Conference*, edited by S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, and M. Fu. 4038-4050. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Waeber, R. 2013. *Probabilistic Bisection Search for Stochastic Root-Finding*. PhD Dissertation. Cornell University, Ithaca, NY.

Wang, H., and S.-H. Kim. 2011. Reducing the Conservativeness of Fully Sequential Indifference-Zone Procedures. *IEEE Transactions on Automatic Control* 58(6):1613-1619

Appendix

Proof of Lemma 1.

$$\begin{aligned}\mathcal{S}_I(\Pi x) &= (V\Pi x)^T(V\Gamma V^T)^{-1}(V\Pi x)\\ &= (V\Gamma V^T(V\Gamma V^T)^{-1}Vx)^T(V\Gamma V^T)^{-1}(V\Gamma V^T(V\Gamma V^T)^{-1}Vx)\\ &= (Vx)^T(V\Gamma V^T)^{-1}(Vx) = \mathcal{S}_I(x).\end{aligned}$$

□

Proof of Corollary 1. We first derive an explicit expression for $(V\Gamma V^T)^{-1}$. Without loss of generality, assume that $I = \{1,\dots,s\}$. Then by noting that $V\Gamma V^T$ is the covariance matrix of Vx, we get

$$Vx = \begin{bmatrix} x_1 - x_s \\ \vdots \\ x_{s-1} - x_s \end{bmatrix} \qquad \text{and} \qquad V\Gamma V^T = \begin{bmatrix} \sigma_1^2+\sigma_s^2 & \sigma_s^2 & \cdots & \cdots & \sigma_s^2 \\ \sigma_s^2 & \sigma_2^2+\sigma_s^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \sigma_s^2 \\ \sigma_s^2 & \cdots & \cdots & \sigma_s^2 & \sigma_{s-1}^2+\sigma_s^2 \end{bmatrix}.$$

For equal variances,

$$V\Gamma V^T = \sigma^2 \begin{bmatrix} 2 & 1 & \cdots & \cdots & 1 \\ 1 & 2 & 1 & \cdots & 1 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & 1 \\ 1 & \cdots & \cdots & 1 & 2 \end{bmatrix} = \sigma^2\left(\mathrm{id}_{s-1} + \mathbf{1}_{s-1}\mathbf{1}_{s-1}^T\right)$$

where id_s is the $s\times s$ identity matrix and $\mathbf{1}_s$ is the $s\times 1$ vector of ones.

By the Sherman-Morrison formula,

$$\begin{aligned}(V\Gamma V^T)^{-1} &= \frac{1}{\sigma^2}\left(\mathrm{id}_{s-1}^{-1} - \frac{\mathrm{id}_{s-1}^{-1}\mathbf{1}_{s-1}\mathbf{1}_{s-1}^T\mathrm{id}_{s-1}^{-1}}{1+\mathbf{1}^T\mathrm{id}_{s-1}^{-1}\mathbf{1}}\right)\\ &= \frac{1}{\sigma^2}\left(\mathrm{id}_{s-1} - \frac{\mathbf{1}_{s-1}\mathbf{1}_{s-1}^T}{1+(s-1)}\right)\\ &= \frac{1}{\sigma^2}\frac{1}{s}\left(s\cdot\mathrm{id}_{s-1} - \mathbf{1}_{s-1}\mathbf{1}_{s-1}^T\right). \qquad (7)\end{aligned}$$

Then we have

$$\begin{aligned}
\mathcal{S}_I(x) &= \frac{1}{\sigma^2}\frac{1}{s}\begin{bmatrix} x_1 - x_s \\ \vdots \\ x_{s-1} - x_s \end{bmatrix}^T \left(s \cdot \mathrm{id}_{s-1} - \mathbf{1}_{s-1}\mathbf{1}_{s-1}^T\right)\begin{bmatrix} x_1 - x_s \\ \vdots \\ x_{s-1} - x_s \end{bmatrix} \\
&= \frac{1}{\sigma^2}\frac{1}{s}\left\{(s-1)\sum_{i=1}^{s-1}(x_i - x_s)^2 - 2\sum_{1\le i<\ell<s}(x_i - x_s)(x_\ell - x_s)\right\} \\
&= \frac{1}{\sigma^2}\frac{1}{s}\sum_{\substack{i<\ell \\ i,\ell\in I}}(x_i - x_\ell)^2,
\end{aligned}$$

which shows the first equality in the corollary because $|I| = s$.

Now we show the second equality of the corollary. From (7),

$$V^T(V\Gamma V^T)^{-1}V = \frac{1}{\sigma^2}\frac{1}{s}(s\cdot\mathrm{id}_s - \mathbf{1}_s\mathbf{1}_s^T) \qquad \text{and} \qquad \Pi = \Gamma V^T(V\Gamma V^T)^{-1}V = \frac{1}{s}(s\cdot\mathrm{id}_s - \mathbf{1}_s\mathbf{1}_s^T).$$

Then

$$\Pi x = \frac{1}{s}(s\cdot\mathrm{id}_s - \mathbf{1}_s\mathbf{1}_s^T)x = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_s - \bar{x} \end{bmatrix}.$$

Finally,

$$\begin{aligned}
\mathcal{S}_I(\Pi x) &= (V\Pi x)^T(V\Gamma V^T)^{-1}(V\Pi x) \\
&= (\Pi x)^T[V^T(V\Gamma V^T)^{-1}V](\Pi x) \\
&= \frac{1}{\sigma^2}\frac{1}{s}\begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_s - \bar{x} \end{bmatrix}^T (s\cdot\mathrm{id}_s - \mathbf{1}_s\mathbf{1}_s^T)\begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_s - \bar{x} \end{bmatrix} \\
&= \frac{1}{\sigma^2}\begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_s - \bar{x} \end{bmatrix}^T\begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_s - \bar{x} \end{bmatrix} \\
&= \frac{1}{\sigma^2}\sum_{i=1}^{s}(x_i - \bar{x})^2.
\end{aligned}$$

□

Proof of Lemma 2. It suffices to prove the claim for $|I| = |J| + 1$. By relabeling systems if necessary, it suffices to prove the claim with $J = \{1,\ldots,s\}$ and $I = \{1,\ldots,s+1\}$. We set

$$H_{s+1} = \left\{(x_1, x_2, \ldots, x_{s+1})^T : \sum_{i=1}^{s+1} x_i = 0\right\}, \qquad Q_s = \left\{(x_1, x_2, \ldots, x_{s+1})^T : \sum_{i=1}^{s} x_i = 0, x_{s+1} = 0\right\}.$$

By the second equality of Corollary 1, it suffices to show that for $x \in \mathbb{R}^{s+1}$,

$$\mathcal{S}_I(x) \geq \frac{1}{\sigma^2} \sum_{i=1}^{s} (x_i - \bar{x}_s)^2, \tag{8}$$

where $\bar{x}_s = (x_1 + \cdots + x_s)/s$. To see that this holds, we define Ψ_s on H_{s+1} as the matrix that projects orthogonally on Q_s, i.e., $\Psi_s x = (x_1 - \bar{x}_s, \ldots, x_s - \bar{x}_s, 0)$. By Lemma 1 and (7), we have, for $x \in \mathbb{R}^{s+1}$,

$$\mathcal{S}_I(x) = \frac{1}{\sigma^2} \frac{1}{(s+1)} \begin{bmatrix} x_1 - x_{s+1} \\ \vdots \\ x_s - x_{s+1} \end{bmatrix}^T \left((s+1) \cdot \mathrm{id}_s - \mathbf{1}_s \mathbf{1}_s^T \right) \begin{bmatrix} x_1 - x_{s+1} \\ \vdots \\ x_s - x_{s+1} \end{bmatrix}$$

This representation immediately yields that

$$\mathcal{S}_I(\Psi_s x) = \frac{1}{\sigma^2} \sum_{i=1}^{s} (x_i - \bar{x}_s)^2.$$

Since projecting decreases any quadratic form, this establishes (8). □

Table A.1: $\eta_{|I|}$ when $\alpha = 10\%$

$\|I\|$	$k = 64$	$k = 32$	$k = 16$	$k = 8$	$k = 7$	$k = 6$	$k = 5$	$k = 4$	$k = 3$
64	6.05042								
63	6.79306								
62	7.07127								
61	7.24002								
60	7.35712								
59	7.44289								
58	7.50694								
57	7.55688								
56	7.59517								
55	7.62437								
54	7.64624								
53	7.66164								
52	7.67146								
51	7.67755								
50	7.67896								
49	7.67697								
48	7.67216								
47	7.66385								
46	7.65204								
45	7.63885								
44	7.62218								
43	7.60345								
42	7.58269								
41	7.55989								
40	7.53440								
39	7.50692								
38	7.47817								
37	7.44679								
36	7.41350								
35	7.37764								
34	7.34060								
33	7.30173								
32	7.26040	4.61250							
31	7.21664	5.16401							
30	7.17116	5.35848							
29	7.12400	5.46804							
28	7.07454	5.53579							
27	7.02218	5.57870							
26	6.96829	5.60356							
25	6.91162	5.61553							
24	6.85224	5.61724							
23	6.79024	5.61064							
22	6.72631	5.59634							
21	6.65867	5.57542							
20	6.58867	5.54845							
19	6.51518	5.51603							
18	6.43833	5.47773							
17	6.35826	5.43468							
16	6.27395	5.38576	3.55536						
15	6.18617	5.33234	3.96549						
14	6.09394	5.27360	4.09377						
13	5.99752	5.20970	4.15224						
12	5.89658	5.13991	4.17580						
11	5.79086	5.06446	4.17568						
10	5.68014	4.98455	4.15873						
9	5.46038	4.84321	4.13041						
8	5.26699	4.68543	4.02733	2.83446					
7	5.05352	4.50849	3.90131	3.05966	2.66510				
6	4.80933	4.29929	3.74380	3.04936	2.85635	2.47348			
5	4.52132	4.04717	3.54574	2.95465	2.81492	2.62431	2.25053		
4	4.16553	3.73733	3.28628	2.78163	2.67081	2.53302	2.34537	1.98200	
3	3.67921	3.30305	2.91240	2.49134	2.40324	2.29859	2.16417	1.97851	1.63182
2	2.81738	2.52053	2.21620	1.89611	1.83132	1.75468	1.66093	1.54027	1.37146