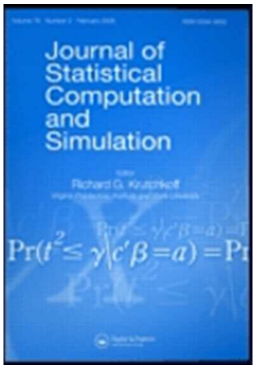

Journal of Statistical Computation and Simulation
Richard G. Krutchkoff
$\Pr(t^2 \leq \gamma | c'\beta = a) = \Pr$

Risk Inflation of Sequential Tests Controlled by Alpha Investing

Journal:	*Journal of Statistical Computation and Simulation*
Manuscript ID:	GSCS-2013-0434.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	Stine, Robert; University of Pennsylvania, Statistics Foster, Dean; University of Pennsylvania, Statistics
Areas of Interest:	LINEAR REGRESSION, MULTIPLE TESTING, MODEL SELECTION, STATISTICAL INFERENCE

To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–18

Risk Inflation of Sequential Tests Controlled by Alpha Investing

Dean P. Foster and Robert A. Stine
Department of Statistics
The Wharton School of the University of Pennsylvania
Philadelphia, PA 19104-6340

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

Streaming feature selection is a greedy approach to variable selection that evaluates potential explanatory variables sequentially. It selects significant features as soon as they are discovered rather than testing them all and picking the best one. Because it is so greedy, streaming selection can rapidly explore large collections of features. If significance is defined by an alpha investing protocol, then the rate of false discoveries will be controlled. The focus of attention in variable selection, however, should be on fit rather than hypothesis testing. Little is known, however, about the risk of estimators produced by streaming selection and how the configuration of these estimators influences the risk. To meet these needs, we provide a computational framework based on stochastic dynamic programming that allows fast calculation of the minimax risk of a sequential estimator relative to an alternative. The alternative can be data-driven or derived from an oracle. This framework allows us to compute and contrast the risk inflation of sequential estimators derived from various alpha investing rules. We find that a universal investing rule performs well over a variety of models and that estimators allowed to have larger than conventional rates of false discoveries produce generally smaller risk.

Key Phrases: stochastic dynamic programming, testimator, variable selection

Research supported by NSF grant DMS-1106743

1. Introduction

Our analysis concerns the risk of sequential estimators in which characteristics of later estimators depend on earlier estimates. Our interest in the risk of such estimators arises from their use in streaming feature selection. Streaming feature selection constructs a predictive model by greedily choosing explanatory variables from a sequence offered by an exogenous source. Think of forward stepwise regression, but without knowledge of the complete domain of explanatory variables. Rather than evaluate the collection of possible explanatory variables together, streaming selection evaluates them one-at-a-time. Searches like stepwise regression consider the full batch of, say, p potential explanatory variables together, choosing at the first step the predictor $X_{(1)}$ that obtains the best fitting model. In contrast, streaming selection is even more greedy and evaluates features sequentially as X_1, $X_2, \ldots$, judging X_j having observed $X_1, \ldots, X_{j-1}$. Hence, streaming selection does not wait to examine every explanatory variable before making a selection. It is also free to use the results of evaluating initial variables to guide the search for those to consider subsequently. For example, a streaming search might test the interaction $X_j\,X_k$ after finding significant effects for X_j and X_k. Streaming selection can thus adaptively explore collections of explanatory variables that are larger than typically considered with conventional methods. Wu et al. [10] provide several examples, including image processing, in which streaming selection is valuable. For large samples, the slowest step in forward stepwise regression is the calculation of the $X'X$ matrix. Lin et al. [8] demonstrate the speed attained by sequential selection when picking a regression from up to 100,000 explanatory variables.

Streaming selection poses a challenge, however, for variable selection. Although one gains advantages by avoiding simultaneously evaluating every predictor, the absence of a fixed set of features in streaming selection requires a different type of selection criterion from those commonly used. For example, suppose the search begins with a list of p potential features $X_1, X_2, \ldots, X_p$. As mentioned, the search could expand to include interactions in X_j once X_j joins the model. If the search were limited to second-order interactions, then the number of possible explanatory variables would become $m = p(p+1)/2$ variables. One could go further and allow the sequential search to construct and test third and fourth-order interactions derived by interacting previously selected features and pairwise interactions. The resulting search would examine only those higher-order interactions that generate nontrivial lower-order interactions. Because higher-order interactions with non-zero effects typically have nontrivial lower-order projections [6], this “bottom-up” search is likely to detect these interactions. Even so, fourth-order interactions are likely to be very sparse, and the streaming search would likely examine only a few of these. It would thus be conservative to combine this search with a criterion such as AIC or BIC that adds a penalty proportional to the size of the search space that would grow to $m = O(p^4)$. The penalties used in these criteria presume one examines all m features rather than a dynamically chosen subset. Similarly, selection using FDR requires the complete set of m possible p-values at the start of the search.

Alpha investing [5] is a sequential testing procedure that is well-suited to streaming feature selection. Because alpha investing can test an infinite sequence of hypotheses, it is well-matched to a search of a possibly unbounded collection of features. Rather than test multiple hypotheses at once, alpha investing tests hypotheses one-at-a-time in a specified order. Alpha investing begins with an initial allowance for Type I error that is called its alpha wealth. Each test consumes some of the available alpha wealth, as in alpha spending rules used in clinical trials. Alpha investing overcomes the conservatism of alpha spending rules, which include the Bonferroni method, by earning a contribution to the alpha wealth for each rejected null hypothesis. Thus rejections beget more rejections. Alpha investing further allows one to test an infinite stream of hypotheses, accommodate dependent tests, and incorporate domain knowledge.

Like other procedures for multiple testing, alpha investing controls the expected number of false rejections. Controlling the false discovery rate protects against overfitting in variable selection. One can guarantee that on average not more than, say, 5% of the rejected hypotheses spuriously add a predictor to the model. When building a predictive model, however, we find that controlling the false discovery rate is secondary to obtaining a more predictive model. Control of the false discovery rate does not imply that one finds the most predictive model. It only guarantees that a high percentage of chosen features are in fact useful. The quadratic risk of the implied estimator is more relevant. Furthermore, alpha investing is not one method, but rather a general approach to testing a sequence of hypotheses. It offers a modeler a variety of choices that control how the procedure sets the level for the next test. The impact of these choices on the risk of the resulting model is by-and-large unknown.

To study the risk of alpha investing estimators and the effects its tuning parameters, we convert the choice of features in regression into the classical normal means problem. Consider fitting the regression of Y, a vector of n observations, on the linearly independent columns $X_1, \ldots, X_p$ that make up the $n \times p$ matrix X. The Gram-Schmidt algorithm sequentially converts these explanatory variables into orthonormal variables $\tilde{X}_1, \ldots, \tilde{X}_p$. In matrix form, the regression $Y = X\beta + \epsilon$ with $\epsilon_i \sim N(0, \sigma^2)$ becomes $Y = (XT)(T^{-1}\beta) + \epsilon = \tilde{X}\mu + \epsilon$, where T is the upper triangular matrix defined by the Gram-Schmidt algorithm. Multiplying both sides of this equation on the left by $\tilde{X}'$ produces

$$\tilde{X}'Y = \tilde{Y} = \mu + \tilde{\epsilon} \,,$$

where $\tilde{\epsilon} = X'\epsilon \sim N(0, \sigma^2)$. Evaluating which of $X_1, \ldots, X_p$ to add to the regression thus becomes a test of which of the first p elements of μ differ from zero. Lehmann and Romano [7] (Chapter 7) describes this conversion in more detail. In practical implementations of sequential variable selection (such as [8]), one only sweeps selected features from those that remain, allowing $p > n$.

The methods developed here find the cumulative risk of a sequence of testimators in the context of testing a sequence of means. Rather than observe a sequence of slope estimates, we assume that the observed statistics are a sequence of p random variables $Y_j \sim N(\mu_j, 1)$. A testimator is also known as a keep-or-kill estimator or a hard thresholding estimator. The estimator of the parameter μ is zero unless $H_0 : \mu = 0$ is rejected. Within this context, our algorithm reveals the attainable risks of a testimator for *any* choice of parameters. We accomplish this task by adopting the perspective of risk inflation and comparing the competitive performance of two sequential estimators. For any pair, we find the set of attainable risks. Given the sequential nature of the problem, it should not be surprising that we rely on stochastic dynamic programming. The risks so obtained are exact (up to computational accuracy) rather than asymptotic. We further show a probabilistic model for the underlying parameters that approximates those risks. Although we consider sequential estimators, we find that bounds for the risk inflation of conventional testimators also characterize the risk of sequential testimators.

The following section provides further introduction to alpha investing. We derive two alpha investing strategies from continuous probability distributions. Our construction of these strategies in Section 2 is novel and allows us to minimize the state space required in the dynamic program. Section 3 describes the risk of a sequence of testimators. Section 4 defines the feasible set of possible risks and uses these to compare testimators to an oracle and to each other. Section 5 describes the computations in more detail. We conclude in Section 6 with a brief discussion of the results and pose conjectures motivated by our computations.

2. Alpha-investing

An alpha-investing rule [5] determines the levels for testing a sequence of hypotheses $H_1, H_2, \ldots,$. The procedure is most easily described by showing a few steps. The process begins with an initial allocation $W_0 > 0$ of alpha wealth. An alpha-investing rule can test H_1 at any level α_1 up to the initial alpha wealth, $0 \leq \alpha_1 \leq W_0$. The level α_1 is 'invested' and cannot be used for subsequent tests. We say that α_1 is invested rather than spent because rejecting H_1 produces an increment in the alpha wealth, a return on the investment. Let p_1 denote the p-value of the test of H_1. If $p_1 \leq \alpha_1$, the test rejects H_1, and the alpha investing rule earns a contribution $\omega \geq 0$ to its alpha wealth. Otherwise, the alpha wealth available to test H_2 falls to $W_1 = W_0 - \alpha_1$. In general, the alpha wealth available for testing H_{j+1} is given by the stochastic process

$$W_j = W_{j-1} - \alpha_j + \omega \, I_{\{p_j \leq \alpha_j\}}, \quad j = 1, 2, \ldots, \tag{1}$$

with the initial condition that specifies the initial wealth W_0. Alpha spending rules are alpha investing rules that constrain $\omega = 0$.

Because rejecting a null hypothesis makes it easier to reject other null hypotheses, it is essential for alpha investing to control the rate of false rejections. To this end, Foster and Stine [5] show that alpha investing controls a sequential version of the expected false discovery rate, mFDR. Let $T(j)$ count the number of hypothesis rejected among the first j tests, and let $V(j) \leq T(j)$ denote the total number of *false* rejections among the first j tests. The sequential mFDR is

$$\text{mFDR}_\eta(j) = \frac{\mathrm{E}\,V(j)}{\eta + \mathrm{E}\,T(j)}\ , \quad \eta > 0. \tag{2}$$

In contrast, FDR is the expected value of the ratio $V(j)/T(j)$ conditional on $T(j) > 0$ rather than the ratio of expected values. The constant η in the denominator of mFDR avoids dividing by zero under the complete null hypothesis in which all $\mu_j = 0$. If $W_0 \leq \eta\,\omega$, then alpha-investing rules control $\text{mFDR}_\eta(p) \leq \omega$, and this result implies weak control of the family wide error rate. Because these properties of alpha investing originate in a martingale, the index j in (2) is allowed to be an arbitrary stopping time, such as the occurrence of the kth rejection.

Alpha-investing rules are quite general. The underlying theory requires only that the level of the test of H_j is bounded by the available wealth W_{j-1} and that the test indeed have level α_j, conditional on the outcomes of prior tests. Otherwise, an alpha investing rule can use the pattern of prior rejections. We represent this dependence by writing an alpha investing rule $\mathcal{A}$ as a function of the sequence of prior wealths $W_{0:j-1} = \{W_0, W_1, \ldots, W_{j-1}\}$. The rule $\mathcal{A}$ maps this history to the interval from zero to the current wealth:

$$\mathcal{A} : W_{0:j-1} \mapsto [0, \min(W_{j-1}, 1)] \tag{3}$$

For example, $\mathcal{A}$ can set the level of the next test higher or lower depending upon the number of tests since the last rejection or the accumulated number of rejections. Although alpha investing allows this generality, we focus on a simpler class of investing rules that have a path independent, Markovian structure. The amount invested by these rules depends only on the current wealth rather than the full path, $\alpha_j = \mathcal{A}(W_{0:j-1}) = \alpha(W_{j-1})$. The only requirement is that $0 \le \alpha(w) \le w$; a rule cannot invest more wealth than the amount possessed. It does seem natural, however, for $\alpha(w)$ to be monotone increasing in w.

Remark A. To avoid adding notation, we overload the symbol α. Throughout these uses, the symbol α consistently gives the level of a test; only the context of the test changes. By itself, α represents the generic level of a test. When given an integer subscript, α_j is the level of the jth test in sequence of tests, as in (1). Finally, denoting a function, $\alpha(w)$ is the level invested in a test by an alpha investing rule that has available wealth w, as in (4) that follows.

The simplest representatives of this class of alpha investing rules are geometric rules. These invest a fixed percentage of the available wealth on each test:

$$\alpha_g(w, \psi) = \psi\, w, \quad 0 < \psi < 1. \tag{4}$$

Since the alpha wealth increases after a rejection (provided $\alpha_j < \omega$), a geometric rule α_g invests more following a rejection and then gradually – at the rate determined by ψ – reduces the level of subsequent tests.

Alternatively an alpha investing rule can vary the share of the current wealth to invest in the next test. Rather than invest a fixed share of the available wealth, the following rule invests progressively less as its wealth drops. The rule is defined by

$$\alpha_u(w) = w - \frac{\log 2}{\log(1 + 2^{1/w})} \; . \tag{5}$$

Using the procedure described in the appendix, we derived α_u from the density function

$$A_u(x) = \frac{\log 2}{x(\log x)^2} \; , \quad 2 \le x \; . \tag{6}$$

A_u is motivated by the penultimate binary code for representing positive integers defined by Elias [3]. This code is universal in the sense of producing compact representations for data sampled from a variety of probability distributions, and so we refer to α_u as a universal investing rule.

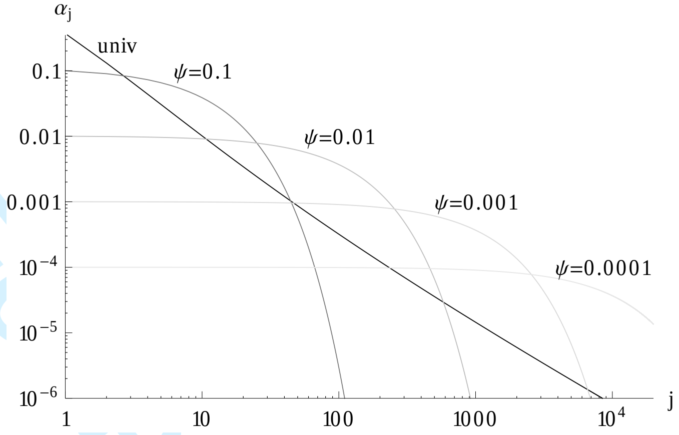
Remark B. The structure of the penultimate code and its relationship to A_u may be of independent interest. Suppose that the random variable X has a discrete distribution on the positive integers with $\mathrm{P}\,(X = j) = p_j$ for $j = 1, 2, \ldots$. A code that represents X using $-\log_2 p_X$ bits has the minimal expected length, which is given by the entropy of the distribution, $-\sum_{j=1}^n (\log_2 p_j) p_j = H_X$. Hence, the shortest binary prefix code for an *iid* sequence $\{X_1, \ldots, X_n\}$ from this distribution has expected length $n\ H_X$. Within this context, the penultimate code of Elias is asymptotically optimal and universal in the following sense. Let $L(j)$ denote the length of the penultimate code for the integer j. So long as the probabilities p_j are monotone decreasing, $p_j \geq p_k$ if $j \leq k$, then $\lim_n \sum_{j=1}^n E\, L(X_j)/(nH_X) = 1$. That is, regardless of the choice of monotone probabilities, the expected length of the penultimate code lies within a constant factor of the length of the best possible code. The structure of the penultimate code is quite simple and leads to the expression (6) for A_u. The penultimate code for an integer x contains the binary representation of x which requires $b(x) \approx \log_2 x$ bits. To form a prefix code, this code includes a binary representation of $b(x)$ that requires about $2 \log_2 b(x)$ additional bits. Combining these, $L(x) \approx \log_2 x + 2 \log_2 \log_2 x$. This length suggests a probability distribution of the form $q(x) = c_q 2^{-L(x)} = c_q/(x\ \log(x)^2)$ with c_q chosen so that $\sum_x q(x) = 1$. $A_u(x)$ is a continuous version of $q(x)$. Rissanen [9] refined the penultimate code by incorporating a more efficient representation for $b(x)$ and labeled the result the universal code for integers. For an extensive introduction to coding and its connections to statistics, see Cover and Thomas [1].

Figure 1 contrasts the investments of the universal and several geometric alpha investing rules. The figure conveys the sense in which α_u is universal in the way that it mimics a collection of geometric rules. For convenience, the initial wealth for all rules shown in Figure 1 is $W_0 = 1$. On a log-log scale, the amounts invested by the universal rule fall off approximately linearly. These amounts are initially larger than those of any of the geometric rules. Starting from $W_0 = 1$, the universal rule invests about 0.369, 0.131, 0.0693, 0.0438, and 0.0306 in the first five tests before its spending gradually slows. After this initial period, there is a range of tests over which each geometric rule invests the largest alpha level. For tests in this range, that rule is the most able to find signal. The universal rule invests almost as much as each geometric rule when that rule invests the most, and ultimately, the universal rule invests more. For example, the geometric rule that invests 1% of its wealth at each test ($\psi = 0.01$) invests more than the universal rule when testing H_{11} through H_{581}; otherwise it invests less. The graph shows that the universal rule saves enough so that it can spend close to the rate of the maximal geometric.

3. Risk Analysis

Before turning to sequential estimators, we briefly review the quadratic risk of testimators. We begin with the scalar case. Let $Y \sim N(\mu, 1)$. The scalar testimator defined by the two-sided test of $H_0 : \mu = 0$

Figure 1. *Each geometric alpha investing rule has a range of hypotheses where it invests the most and hence is the most sensitive to detecting signal. The universal rule spends almost as much as each geometric when that geometric is the highest spender.* The graph shows the alpha levels assuming no intervening test rejects and the initial wealth W_0=1.



at level α is

$$\hat{\mu}_\alpha(Y) = \begin{cases} Y & \text{if } z_\alpha^2 \le Y^2, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where z_α denotes the two-sided critical value, $z_\alpha = \Phi^{-1}(1-\alpha/2)$. The risk of $\hat{\mu}_\alpha$ is

$$\begin{aligned} R(\hat{\mu}_\alpha(Y), \mu) &= \mathrm{E}\,(\hat{\mu}_\alpha(Y) - \mu)^2 \\ &= \mu^2 \mathrm{P}\,(Y^2 \le z_\alpha^2) + \int_{z_\alpha^2 < y^2} (y-\mu)^2 \phi(y)dy \\ &= B_\alpha(\mu) + V_\alpha(\mu) \end{aligned} \tag{8}$$

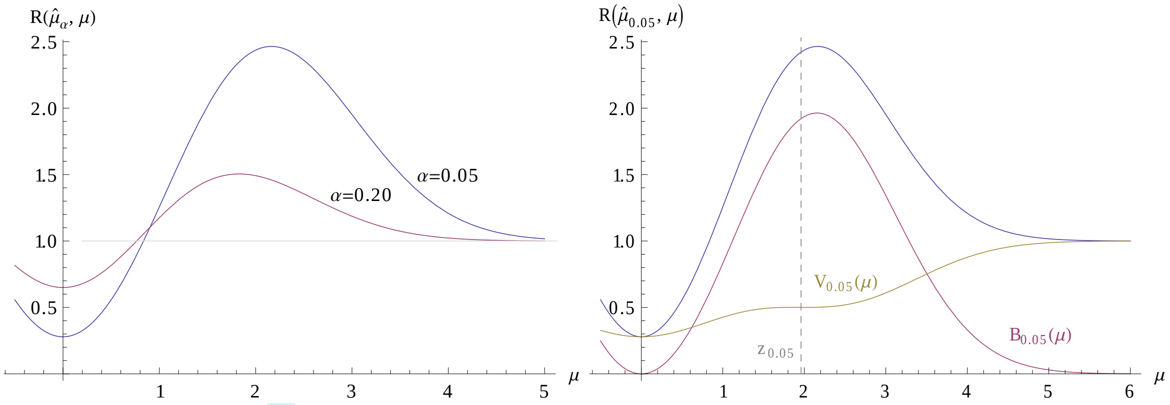
The first summand $B_\alpha(\mu)$ is the squared bias that arises if the test of $H_0 : \mu = 0$ does not reject when $\mu \neq 0$; the second summand is the variance of the estimator. Figure 1(a) shows a graph of the risk of testimators with $\alpha = 0.05$ and $\alpha = 0.20$. The maximum risk occurs near z_α and grows as α falls. Figure 1(b) shows the decomposition of the risk of $\hat{\mu}_\alpha$ into $B_\alpha(\mu)$ and $V_\alpha(\mu)$ for $\alpha = 0.05$. Because the variance component $V_\alpha(\mu)$ increases smoothly to its maximum 1 for large $|\mu|$, it is the bias that produces the noticeable peak in the risk.

The risk of testimators is typically studied in the context of estimating a vector of p means, $\boldsymbol{\mu} \equiv \mu_{1:p} = (\mu_1, \ldots, \mu_p)'$. The available data is the vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ with distribution $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, I_p)$. In this context, the estimator of $\boldsymbol{\mu}$ combines testimators with a common level, $\hat{\boldsymbol{\mu}}_\alpha = (\hat{\mu}_\alpha(Y_1), \ldots, \hat{\mu}_\alpha(Y_p))'$. The estimator $\hat{\boldsymbol{\mu}}_\alpha$ consists of zeros except for those coordinates where $z_\alpha^2 \le Y_j^2$. The risk of $\hat{\boldsymbol{\mu}}_\alpha$ is the sum of the risks of the coordinate testimators,

$$R(\hat{\boldsymbol{\mu}}_\alpha, \boldsymbol{\mu}) = \mathrm{E} \sum_{j=1}^{p} \Big(\hat{\mu}_\alpha(Y_j) - \mu_j \Big)^2 . \tag{9}$$

Minimax bounds for the risk $R(\hat{\boldsymbol{\mu}}_\alpha, \boldsymbol{\mu})$ are well-understood. We review the results of Foster and

Figure 1. *The risk of a testimator peaks near z_α due to a large contribution from bias as the level α decreases. (a) Risk of testimators with α = 0.05, 0.20 versus μ. (b) Squared bias and variance components of the risk of* $\hat\mu_{0.05}$.



George [4] who introduced the concept of the risk inflation of an estimator. Donoho and Johnstone [2] obtain similar results. The risk inflation of $\hat{\boldsymbol\mu}_\alpha$ is the supremum of the ratio of the risk of $\hat{\boldsymbol\mu}_\alpha$ to that of a testimator that obtains the optimal level from an oracle. Their results imply that the risk inflation of $\hat{\boldsymbol\mu}_\alpha$ is asymptotically about $2\log p$,

$$2\log p - o(\log p) \le \sup_\mu \frac{1 + R(\hat{\boldsymbol\mu}_\alpha, \boldsymbol\mu)}{1 + \inf_\eta R(\hat{\boldsymbol\mu}_\eta, \boldsymbol\mu)} \le 2\log p + 1 \ . \qquad (10)$$

Foster and George further show that the testimator $\hat{\boldsymbol\mu}_{1/p}$ – essentially the Bonferroni estimator – obtains the risk inflation threshold. The constant 1 added to the risks in the ratio of (10) arises in the context of regression models in which one always estimates the intercept. As a practical device, its presence avoids dividing by zero under the complete null model in which $\mu_j = 0$ for all j.

Though suggestive, these results do not reveal the risk of the testimator derived from alpha investing. The key distinction lies in the timing of the information revealed in $\boldsymbol Y$. The testimator $\hat{\boldsymbol\mu}_\alpha$ studied in risk inflation uses a fixed level α for all p coordinates, and all of the elements of $\boldsymbol Y$ are available when choosing α. In sequential testing controlled by alpha investing, the Y_j are observed sequentially. The elements of the estimator form a stochastic process, which we collect in a p-element vector as

$$\hat\mu(\alpha(\cdot), W_0, \omega) = (\hat\mu_{\alpha(W_0)}, \hat\mu_{\alpha(W_1)}, \ldots, \hat\mu_{\alpha(W_{p-1})})', \qquad (11)$$

where $\alpha(\cdot)$ denotes the defining investing rule, W_0 is the initial alpha wealth, and ω is the payout earned when rejecting a hypothesis. We omit W_0 and ω from this notation when unambiguous.

The most convenient expression for the risk of $\hat\mu(\alpha(\cdot), W_0, \omega)$ relies on a recurrence. Let

$$r_\mu(\alpha) = \Phi(\mu - z_\alpha) + \Phi(-\mu - z_\alpha)$$

denote the probability of rejecting $H_0 : \mu = 0$ using a two-sided z-test at level α (the power of the test). The risk of the testimator given by alpha investing can then be decomposed as

$$\begin{aligned} R(\hat{\mu}(\alpha(\cdot), W_0, \omega), \mu_{1:p}) &= R(\hat{\mu}_{\alpha(W_0)}, \mu_1) + \mathrm{E} \sum_{j=2}^{p} R\Big(\hat{\mu}_{\alpha(W_{j-1})}, \mu_j\Big) \\ &= R(\hat{\mu}_{\alpha_1}, \mu_1) + r_{\mu_1}(\alpha_1)\; R\Big(\hat{\mu}(\alpha(\cdot), W_0 - \alpha_1 + \omega, \omega), \mu_{2:p}\Big) \\ &\qquad +(1 - r_{\mu_1}(\alpha_1))\; R\Big(\hat{\mu}(\alpha(\cdot), W_0 - \alpha_1, \omega), \mu_{2:p}\Big)\ , \end{aligned} \tag{12}$$

where $\alpha_1 = \alpha(W_0)$ and we have suppressed the dependence of the estimator on $\boldsymbol{Y}$. The second expression for the risk emphasizes its recursive nature and motivates our method of computation. The total risk of the estimator is the risk produced by the testimator for H_1 plus the cumulative risk of testing $H_2, \ldots, H_p$. If the test of H_1 rejects, which happens with probability $r_{\mu_1}(\alpha_1)$, then testing the remaining hypotheses begins with alpha wealth $W_1 = W_0 - \alpha_1 + \omega$. Otherwise, with probability $1 - r_{\mu_1}(\alpha_1)$, testing begins with wealth $w_1 - \alpha_1$.

The calculation of the maximum risk of $\hat{\mu}(\alpha(\cdot), W_0, \omega)$ is similarly recursive, and we compute the sum by backward induction. Because the performance of subsequent tests depends on the outcome of the first, the choice of μ_1 is not so simple as setting $\mu_1 = \arg\max R(\hat{\mu}_{0.05}, \mu)$. Doing so ignores the payoff ω obtained if H_1 is rejected. By rejecting the first test, alpha investing adds ω to its alpha wealth, allowing it to increase the level – and so potentially reduce its risk – in subsequent tests. To find the maximum risk, one must choose at the first test the mean

$$\begin{aligned} \mu_1 = \arg\max_{m} \Big\{ & R(\hat{\mu}_{\alpha_1}, m) + r_m(\alpha_1) \max_{\mu_{2:p}} R(\hat{\mu}(\alpha, W_0 - \alpha_1 + \omega, \omega), \mu_{2:p}) \\ & +(1 - r_m(\alpha_1)) \max_{\mu_{2:p}}\ R(\hat{\mu}(\alpha, W_0 - \alpha_1, \omega), \mu_{2:p}) \Big\}\ . \end{aligned}$$

Notice that the sequence of means $\mu_1, \mu_2, \ldots, \mu_p$ that maximizes the risk is not deterministic because of the stochastic outcome of the tests. As a result, our calculations identify a stochastic process of means that obtains, on average, the maximum risk.

4. Feasible Risk Set

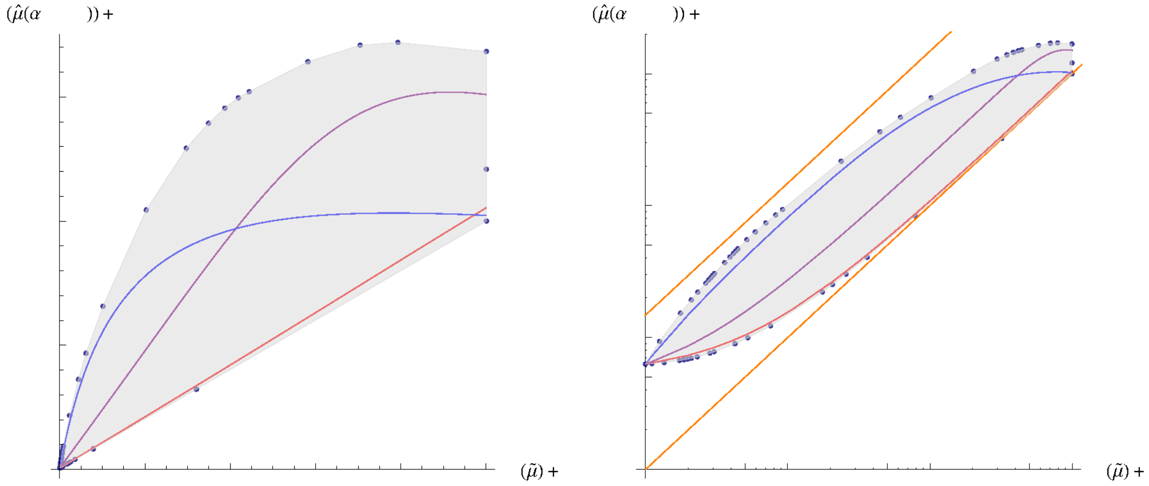
Our interest is not simply in the risk of a testimator, however, but in its risk when compared to an alternative. We want to see how various sequential testimators perform when estimating the same collection of means. In the style of risk inflation (10), we want to contrast the risk of $\hat{\mu}(\alpha(\cdot), W_0, \omega)$ to that of another realizable testimator or to a testimator that benefits from an oracle that reveals $\boldsymbol{\mu}$. The oracle-based, risk-inflation testimator $\tilde{\boldsymbol{\mu}}$ has elements

$$\tilde{\mu}_j(Y_j) = \begin{cases} 0 & \text{if } \mu_j^2 \le 1, \\ Y_j & \text{otherwise.} \end{cases} \tag{13}$$

so that its risk is

$$R(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu}) = \sum_j \min(\mu_j^2, 1)\ . \tag{14}$$

Figure 1. *The shaded feasible set identifies the possible risks of the oracle estimator* $\tilde{\boldsymbol{\mu}}$ *and universal testimator* $\hat{\mu}(\alpha_u(\cdot), W_0, \omega)$ *with* $W_0 = \omega = 0.5$*, p=1,000.* The left frame emphasizes models with substantial signal; the right frame (log scale) highlights nearly black models. Curves within the feasible set show the risks as the amount of signal varies in the model (16). Boundary points show calculation locations described in Section 5; the line parallel to the diagonal in the right frame is the risk-inflation boundary discussed in the text.



We summarize such comparisons of risks by finding the collection of all possible risks that are obtainable under any mean process. We call this collection the feasible risk set. Let $\hat{\boldsymbol{\mu}}_1$ and $\hat{\boldsymbol{\mu}}_2$ denote two sequential estimators of $\mu_{1:p}$. The feasible risk set for these two is defined as

$$\mathcal{R}_p(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) = \{(r_1, r_2) : \exists \boldsymbol{\mu} \text{ s.t. } r_1 = \mathrm{E}_{\boldsymbol{\mu}}\, R(\hat{\boldsymbol{\mu}}_1, \boldsymbol{\mu}),\ r_2 = \mathrm{E}_{\boldsymbol{\mu}}\, R(\hat{\boldsymbol{\mu}}_2, \boldsymbol{\mu})\}\ . \tag{15}$$

In words, the point $(r_1,\ r_2)$ lies in the feasible set $\mathcal{R}_p$ if there exists a stochastic process of means $\boldsymbol{\mu}$ of length p for which the risk of $\hat{\boldsymbol{\mu}}_1$ is r_1 and the risk of $\hat{\boldsymbol{\mu}}_2$ is r_2. A randomization argument proves that the feasible risk set is convex. If x and y are two points within $\mathcal{R}_p$, then there exist stochastic processes $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$, say, that produce these risks. The risk produced by the randomized process that picks $\boldsymbol{\mu}_x$ with probability $0 \le a \le 1$ and picks $\boldsymbol{\mu}_y$ with probability $1-a$ is then $a\,x + (1-a)\,y$.

Figure 1 shows two views of the feasible set that compares the risk-inflation testimator $\tilde{\boldsymbol{\mu}}$ (x-axis) to the universal testimator $\hat{\mu}(\alpha_u(\cdot), W_0, \omega)$ (y-axis). For this figure, p=1,000 tests and the initial wealth and payout $W_0 = \omega = 0.5$. The feasible risk set is the shaded region in each frame. The feasible risk set lies above the diagonal in this comparison; by construction, no realizable testimator can have smaller risk than $\tilde{\boldsymbol{\mu}}$. The frame on the left of Figure 1 shows the feasible set on the scale of risks; the frame on the right shows $\mathcal{R}_p$ on log scales. ($\mathcal{R}_p$ is not convex on a log scale but the approximation is quite close in practice.) We add 1 to the risks of both estimators, in the fashion of risk inflation, in order to be able to show the feasible risk set near 0 on a log-log scale. Points in the plot along the boundary of the feasible set identify locations at which we computed the exact risks using the method described in the following section. Consequently, because the shaded region in the graph is obtained by joining these points with lines, this region is a convex subset within the interior of $\mathcal{R}_p$. The actual risk set is slightly larger.

The two plots in Figure 1 contrast models with substantial signal (non-zero means) to those that are sparse or nearly black (most $\mu_j = 0$). The frame scaled by the risk itself emphasizes the performance in models with substantial signal. The vertical right edge of the feasible set shows the risk for saturated models in which $|\mu_j| \geq 1$; for these models, $R(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu}) = p$. The plot on the log scale emphasizes sparse models. In this frame, the line parallel to and above the diagonal is the risk-inflation boundary (10) that obtains for non-sequential estimators. These bounds suggest that the worst case risk for the testimator should be about $2 \log p$ times the risk of the oracle. The feasible set calculations show that the risk of $\hat{\mu}(\alpha_u(\cdot), W_0, \omega)$ does indeed fall below this boundary, but that is not true of all estimators.

Curves within the feasible set shown in Figure 1 identify the risks that result if $\boldsymbol{\mu}$ is determined by a two-point model. Suppose that the stochastic process that determines $\boldsymbol{\mu}$ sets μ_j independently to some $\mu^* \neq 0$ with probability π and sets $\mu_j = 0$ otherwise:

$$\mu_j = B_j\ \mu^*, \quad B_j \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi)\ . \tag{16}$$

The smooth curves within the feasible set show the risks under this model, with $\mu^* = 1.0$ (red), 1.5 (magenta), or 3 (blue), and the probability of a non-zero mean varying over the range $10^{-6} \leq \pi \leq 1 - 10^{-6}$. With $\mu^* = 1.0$, the risks nearly trace out the lower boundary of the feasible set. The crossing of the paths for $\mu^* = 1.5$ and $\mu^* = 3$ shows, however, that no one value for μ^* can reproduce the upper boundary of $\mathcal{R}_p$.

Although the simple, two-point model (16) cannot fully characterize the feasible set, the processes that define the boundary resemble its structure. Figure 2 graphs a realization of the stochastic process that generates the risks on the boundary of the feasible risk set. For this figure, we chose the process that produces the point with expected risks $(101, 657)$ which can be found along the left side of the feasible set in Figure 1. The dots in Figure 2 graph the elements μ_j versus the test index j for $j = 1, 2, \ldots, p = 1,000$. As in the two-point model, the means jump between 0 and a value that fluctuates around 2.63. The increasing trend in the figure shows the cumulative risk obtained by the universal testimator for this realization. Its risk in this instance reaches 688, whereas the cumulative risk of the oracle is 123.

Displays that combine several feasible sets allow one to compare the effects of various choices for W_0 and ω. As an example, Figure 3 considers the effect of reducing the initial wealth W_0 and payoff ω from 0.50 down to 0.25 and 0.05. As before, the left frame emphasizes estimation with greater levels of signal; the right frame on the log scale emphasizes sparse models. Within the context of hypothesis testing, $\alpha = 0.05$ is a virtual default and one may be similarly tempted to control mFDR at 0.05. Unless one believes that nature will play a nearly black strategy, however, setting $W_0 = \omega = 0.05$ generates greater risk than $W_0 = 0.25$ or 0.50. With $W_0 = 0.05$, the risks even escape the bounds suggested by risk inflation in the non-sequential setting, shown here by a portion of the feasible set above the bound provided in (10).

Remark C. Because the plots in Figure 3 show several feasible sets together, one can no longer associate a point in the graph with a single mean process $\boldsymbol{\mu}$. Points within each feasible set indicate that there exists a mean process that generates the shown risks, but at a given (x, y) location, the

Figure 2. *The mean process associated with a boundary point of the feasible set produces realizations that resemble those of the two-point model (16).* The increasing trend shows the accumulating risk of the testimator which reaches 688 in this example.

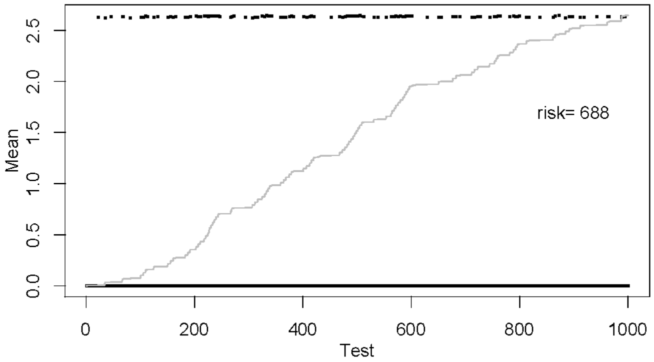
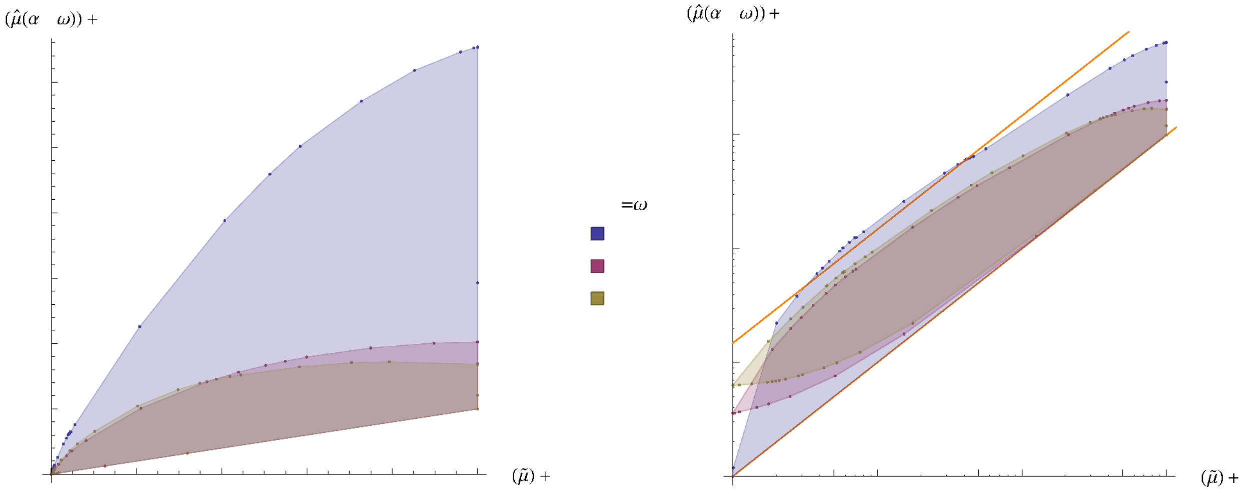


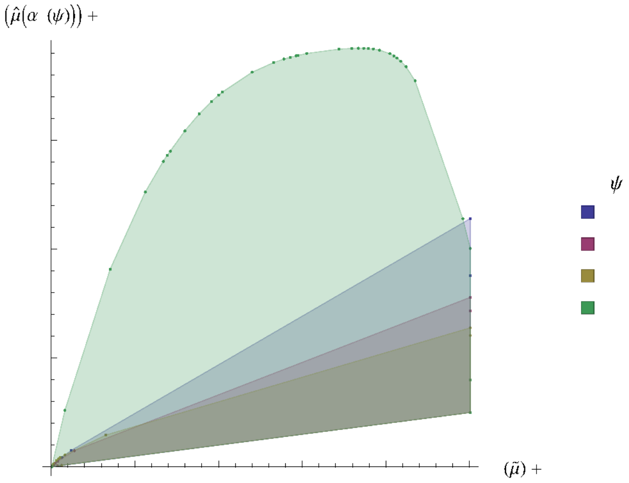
Figure 3. *The initial wealth and payout influence the feasible set of risks that contrast the risk-inflation testimator $\tilde{\boldsymbol{\mu}}$ to the universal estimator $\hat{\mu}(\alpha_u(\cdot), W_0, \omega)$.* The initial wealth varies over $W_0 = 0.05$, 0.25, and 0.50 with p=1,000 tests on the scale of risks (left) or log risks (right).



mean processes that produce the risks for the feasible sets may differ.

We have emphasized universal investing defined by α_u, and Figure 4 offers a partial explanation for our preference. Figure 4 superimposes the feasible sets obtained by geometric investing with various rates versus the risk-inflation testimator $\tilde{\boldsymbol{\mu}}$. The results are for a sequence of $p = 500$ tests. In general, increasing the spending rate ψ from 0.001 up to 0.01 reduces the risk of the geometric testimator $\alpha_g(w, \psi)$. The feasible sets for $\psi = 0.001$, 0.005, and 0.01 progressively concentrate toward the diagonal, better competing with $\tilde{\boldsymbol{\mu}}$. The move to $\psi = 0.05$, however, goes too far. The geometric estimator essentially exhausts its alpha wealth before the testing is complete, and consequently its risk soars. Because this geometric estimator essentially sets $\hat{\mu}_j \equiv 0$ as its alpha wealth approaches 0, its risk exceeds the risk inflation boundary (10).

Figure 4. *A high spending rate ψ potentially exhausts the wealth of a geometric testimator and produces excessive risks compared to $\tilde{\boldsymbol{\mu}}$.* These results are for $p = 500$ tests and rates $\psi = 0.001$, 0.005, 0.01, and 0.05.



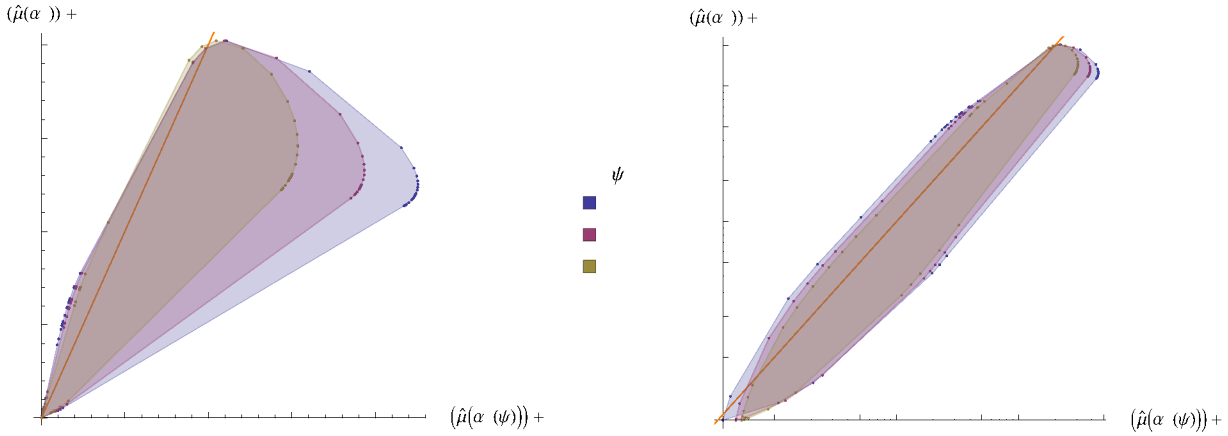
As a final example, feasible risk sets also allow us to directly compare realizable testimators produced by different methods of alpha investing. Rather than compare a realizable testimator to an oracle as in Figures 3 and 4, the feasible set $\mathcal{R}_p(\hat{\mu}(\alpha_u), \hat{\mu}(\alpha_g))$ shown in Figure 5 pits these two against each other in a head-to-head comparison. The initial value and payout for both are $W_0 = \omega = 0.25$. The rates of the geometric strategy are set to $\psi = 0.001$, 0.002, and 0.005. Small rates are necessary to avoid the surge in risk illustrated in Figure 4 when the geometric strategy runs out of wealth. There are clearly mean processes for which either choice, universal or geometric, dominate the other. That said, this figure clarifies the relative advantages of universal investing over geometric investing. A higher investing rate ψ for geometric investing reduces risk for models with more signal, but doing so necessarily leads to higher risks in nearly black models. For instance, the set in Figure 5 associated with $\psi = 0.005$ reduces the bulge toward higher risks in the left frame, but this choice is soundly dominated by slower spending rates in models with fewer non-zero parameters emphasized by the log scale in the right frame. Universal investing removes this tuning parameter.

5. Computation

We describe first the calculation of the feasible set $\mathcal{R}_p(\hat{\mu}(\alpha(\cdot), W_0, \omega), \tilde{\boldsymbol{\mu}})$ that contrasts an alpha investing testimator with the risk-inflation testimator $\tilde{\boldsymbol{\mu}}$. Because $\tilde{\boldsymbol{\mu}}$ has no wealth constraint, calculations need only track the wealth of the alpha investing estimator, which we abbreviate as $\hat{\boldsymbol{\mu}}$ with the understanding that it depends on the choice of the investing function $\alpha(\cdot)$, W_0, and ω throughout this section. Let

$$\mathcal{U}^{\theta}(\hat{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}) = \max \mathrm{E}_{\boldsymbol{\mu}} \Big(\cos(\theta) R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) + \sin(\theta) R(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu})\Big) \tag{17}$$

Figure 5. *The universal alpha investing testimator* $\hat\mu(\alpha_u(\cdot), W_0, \omega)$ *(y-axis) produces typically smaller risks than geometric testimators* $\hat\mu(\alpha_g(\cdot,\ \ ), W_0, \omega)$*(x-axis)*. The geometric rates are $\psi = 0.001, 0.002$, and 0.005 with p =1,000.



denote the maximum expected value with respect to a stochastic process $\boldsymbol{\mu}$ of the weighted sum of risks defined by the angle $0 \le \theta \le 2\pi$. Let $\boldsymbol{\mu}^\theta$ denote the mean process that maximizes $\mathcal{U}^\theta$. The point $\mathrm{E}_{\boldsymbol{\mu}^\theta}\left(R(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}), R(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu})\right)$ lies on the boundary of $\mathcal{R}_p(\hat{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}})$ where the feasible risk set is tangent to the line defined by the mixture weights in (17). Plots that show the feasible risk set, such as Figure 1, highlight the boundary points that are explicitly computed. By varying θ over the circle, we approximate the feasible risk set as the intersection of the resulting half-planes.

We compute $\mathcal{U}^\theta$ via numerical backward induction. This induction approximates the wealth of the alpha investing testimator on a grid. The wealth grid G reaches from the minimal attainable wealth $(W_0 - \sum_{j=1}^p \alpha_j)$ to a maximum allowed wealth, which we set to $W_{\max} = 5$. (Our results have not been sensitive to the choice of $W_{\max}$ so long as it is substantially larger than $W_0 + \omega$.) The wealth grid is 'logarithmically' spaced at N points, with a finer spacing 0.0001 for small wealths below 0.001 and gradually larger spacing as the wealth increases. We insure that the grid includes an element $G_{k_0} = W_0$.

The $p \times N$ matrix U^θ holds intermediate calculations of the expected value $\mathcal{U}^\theta$. The rows in this matrix identify the hypothesis H_j and the columns index the position in the wealth grid G. We fill U^θ from the 'bottom up' in a tail recursion. At the completion of the calculations,

$$\begin{aligned} U^\theta_{jk} = \max_\mu \Big\{ & \cos(\theta) R(\hat\mu_{\alpha(G_k)}, \mu) + \sin(\theta) R(\tilde\mu, \mu) \\ & + r_\mu\left(\alpha(G_k)\right) \left(c\, U^\theta_{j+1,k_c+1} + (1-c) U^\theta_{j+1,k_c}\right) \\ & + \left(1 - r_\mu(\alpha(G_k))\right) \left(d\, U^\theta_{j+1,k_d+1} + (1-d) U^\theta_{j+1,k_d}\right) \Big\} \end{aligned} \tag{18}$$

for $j = p, p-1, \dots, 1$ and $k = 1, \dots, N$ and the boundary condition $U^\theta_{p+1,k} = 0$. Note the similarity to expression (12). At the completion of the computation, $\mathcal{U}^\theta = U^\theta_{1,k_0}$. The first line of (18) adds the contribution to the weighted risk from testing H_j at the alpha level $\alpha(G_k)$. The second line adds the expected subsequent risk if the test rejects H_j, which occurs with probability $r_\mu(\alpha(G_k))$. If the test rejects, the alpha wealth increases to $G_k - \alpha(G_k) + \omega$. This wealth is unlikely to match that at any grid position, so we linearly interpolate between positions k_c and $k_c + 1$ so that $G_{k_c} \le G_k - \alpha(G_k) + \omega \le$

G_{k_c+1} and set the weight c in (18) to $c = (G_k - \alpha(G_k) + \omega)/(G_{k_c+1} - G_{k_c})$. Similarly, the third line of (18) adds the expected contribution if the testimator does not reject H_j and its wealth falls to $G_k - \alpha(G_k)$.

Remark D. One need not store the full matrix U^θ, but its use simplifies the description of the algorithm. One only needs the next row $U^\theta_{j+1,\cdot}$ to compute $U_{j,\cdot}$. Such space saving – using just two rows rather than the full matrix – becomes essential in problems that track a larger state space. Note also that one can cache the indices and weights (k_c, c and k_d, d) prior to the recursion because these can be determined from the grid positions and ω and remain fixed throughout the backward recursion.

The feasible set that compares the testimators defined by two alpha investing rules $\alpha(\cdot)$ and $\beta(\cdot)$ requires a more complex recursion that must track the wealths of both. The linear grid G remains, but the matrix U^θ defined in (18) becomes a three dimensional tensor of size $p \times N \times N$. The calculation is essentially a more messy version of (18) but for one nuance that we want to emphasize. To simplify the presentation, we suppress the linear interpolation and pretend that all of the concerned wealths are represented in the wealth grid. If the alpha investing rule $\alpha(\cdot)$ with wealth G_k rejects H_j, its wealth goes from G_k to G_{k+}; if it fails to reject, its wealth falls to G_{k-}. Similarly, we use $\ell+$ and $\ell-$ for the positions for the rule defined by $\beta(\cdot)$. If we assume $\alpha(G_k) < \beta(G_\ell)$, then the recursion can be written as

$$\begin{aligned} U^\theta_{jk\ell} = \max_\mu \Big\{ & \cos(\theta) R(\hat\mu_{\alpha(G_k)}, \mu) + \sin(\theta) R(\hat\mu_{\beta(G_\ell)}, \mu) \\ & + r_\mu\left(\alpha(G_k)\right)\, U_{j+1,k+,\ell+} + \left[r_\mu(\beta(G_\ell)) - r_\mu(\alpha(G_k))\right] U_{j+1,k-,\ell+} \\ & + \left[1 - r_\mu(\beta(G_\ell))\right] U_{j+1,k-,\ell-} \Big\}, \end{aligned} \tag{19}$$

where the boundary condition is $U^\theta_{p+1,\cdot,\cdot} = 0$. The first line in (19) is the expected risk produced by the test of H_j, and following summands denote the expected contributions to the risk if both reject, if only the rule with the larger alpha level rejects, and if neither rejects. The point of writing this out is to emphasize these testimators see the same data, not independent samples. Hence, $\alpha(G_k) < \beta(G_\ell)$ implies that if the first rule $\alpha(\cdot)$ rejects H_j, then the second rule must also reject H_j because it tests the same hypothesis using the same data, but with a larger alpha level.

6. Discussion

Feasible risk sets allow us to study the risks of testimators in sequential problems. The comparisons shown here suggest that universal alpha investing does well and can compete with the risks attained by any geometric procedure. It is also of interest to point out that a large initial alpha wealth and payout $W_0 = \omega = 0.25$ produce a noticeable reduction in the risk (Figure 3). This choice for ω implies that controlling the expected false discovery rate at 25%, quite a bit larger than would usually be chosen, produces lower risk unless the mean process is quite sparse. For example, the comparisons of streaming estimators in Wu et al. [10] set W_0 and $\omega = 0.01, 0.05$. It would be useful to investigate whether larger choices improve the performance of these estimates in their models (which are more complex than the idealized testing of independent means shown here).

A particular benefit of these computations is that they suggest conjectures about asymptotic properties of these estimators. For example, it appears that we can approximate the boundary of the feasible risk set using two-point models defined in (16). Figure 1 shows that by varying the signal probability π such a model can be found that approaches the boundary of the feasible risk set. Further, the simulation shown in Figure 2 shows that (at least for this location) the boundary mean process generates either zero or approximately a single, non-zero value. Hence, it would appear that, asymptotically in p, there exists a two-point model (π, μ^*) for which the risks lie within some epsilon ball of the boundary of the feasible risk set.

The shapes of the various feasible sets are also intriguing. For instance in Figure 1, the set has a vertical segment where the risk of the oracle attains its maximum at p. These risks occur when the mean process is saturated in the sense that every $\mu_j^2 \geq 1$ so that the risk-inflation oracle "fits everything." Although the oracle then has fixed risk p, the risk of the testimator varies with the size of μ_j. This property of the feasible risk set for saturated mean processes is rather different from the behavior for sparse processes. As the risk of the oracle estimator approaches its minimum, the feasible risk set approaches a single point. That the set comes to a point is not surprising. Unlike the saturated case, a unique process produces the minimum risk, namely the process for which every $\mu_j = 0$. What is surprising is the lack of evident curvature. Does the feasible set come to a point or instead form a very tight curve?

As a final conjecture, the performance of testimators derived from the universal rule $\alpha_u(\cdot)$ suggests that this investing rule can simplify the choice of an alpha investing method. Figure 5 shows its ability to match the risks obtained by various geometric testimators. Ideally, we would like to obtain results that show that universal alpha investing is about as good as one can do, analogous to those in Rissanen [9]. Such a proof would then simplify the use of alpha investing in practice as one would not need to struggle with the choice of an investing rule but instead could focus on generating a better stream of features.

Appendix

We obtained the universal rule α_u defined in (5) by the following construction. The idea is to define a spending rule by a probability distribution that allocates wealth over subsequent tests and then shift from discrete to continuous distributions. We illustrate the construction for the geometric rule. If the initial wealth is W_0, then the alpha wealth invested in the jth test (assuming no intervening test rejects) is

$$\alpha_j = W_0 \, \psi \, (1-\psi)^{j-1} \, , \quad j = 1, 2, \ldots \, . \tag{20}$$

Rather than define the investing rule using a discrete distribution on $j = 1, 2, \ldots$ as in (20), consider the continuous density

$$A_g(x) = c_g\ \psi\,(1-\psi)^{x-1}, \quad 1 \le x\ , \tag{21}$$

where the normalizing constant $c_g = -\log(1-\psi)/\psi$ implies $\int_1^\infty A_g(x)dx = 1$. Notice that the wealth invested in the jth test (20) matches W_0 times the integral of $A_g(x)$ from $x = j$ to $j+1$,

$$\alpha_j = W_0 \int_j^{j+1} A_g(x)dx = W_0\,\psi\,(1-\psi)^{j-1}\ . \tag{22}$$

To move away from discrete indexing, we use the following tail integral,

$$W_g(x) = W_0 \int_x^\infty A_g(t,\psi)dt = W_0(1-\psi)^{x-1}\ . \tag{23}$$

For integers j, $W_g(j)$ is the wealth available to test H_j if none of $H_1, H_2, \ldots, H_{j-1}$ are rejected. By inverting this tail integral, we can write the investing rule as a function of just the available wealth,

$$\alpha_g(w,\psi) = A_g(W_g^{-1}(w)) = \psi\,w\ , \tag{24}$$

as in (4). This construction uses the inverse of the tail wealth to determine a ‘hypothesis index’ $W_g^{-1}(w)$ that corresponds to wealth w. The universal rule $\alpha_u(w)$ follows from the same construction, but starts with the density A_u defined in equation (6).

Acknowledgement

The authors gratefully acknowledge support from NSF (award #1106743) and the helpful suggestions from an anonymous referee.

References

[1] Cover, T. M. and J. A. Thomas. 2006. *Elements of Information Theory (2nd Edition)*. Hoboken, NJ: Wiley.

[2] Donoho, D. L. and I. M. Johnstone. 1994. “Ideal spatial adaptation by wavelet shrinkage.” *Biometrika* 81:425–455.

[3] Elias, P. 1975. “Universal codeword sets and representations of the integers.” *IEEE Trans. on Info. Theory* 21:194–203.

[4] Foster, D. P. and E. I. George. 1994. “The Risk Inflation Criterion for Multiple Regression.” *Annals of Statistics* 22:1947–1975.

[5] Foster, D. P. and R. A. Stine. 2008. "α-investing: a procedure for sequential control of expected false discoveries." *Journal of the Royal Statist. Soc., Ser. B* 70:429–444.

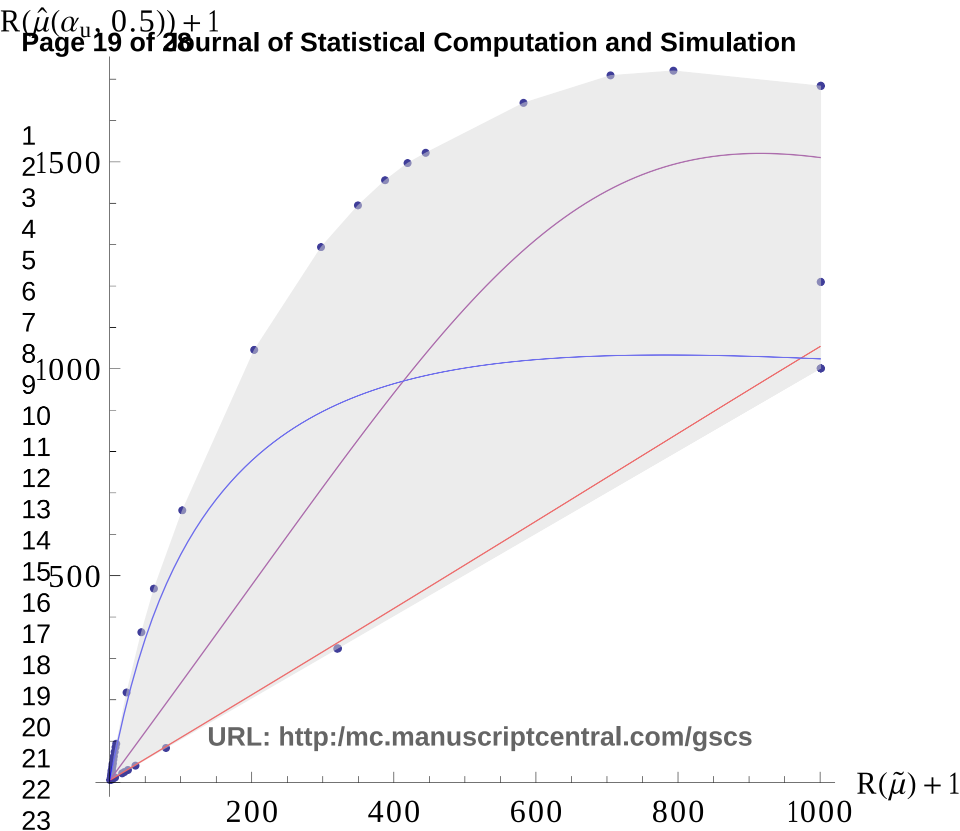
[6] Kalai, A. T., A. Samorodnitsky, and S.-H. Teng. 2009. "Learning and Smoothed Analysis." In *Proceedings of the 50th Annual Symposium on Computer Science (FOCS)*.

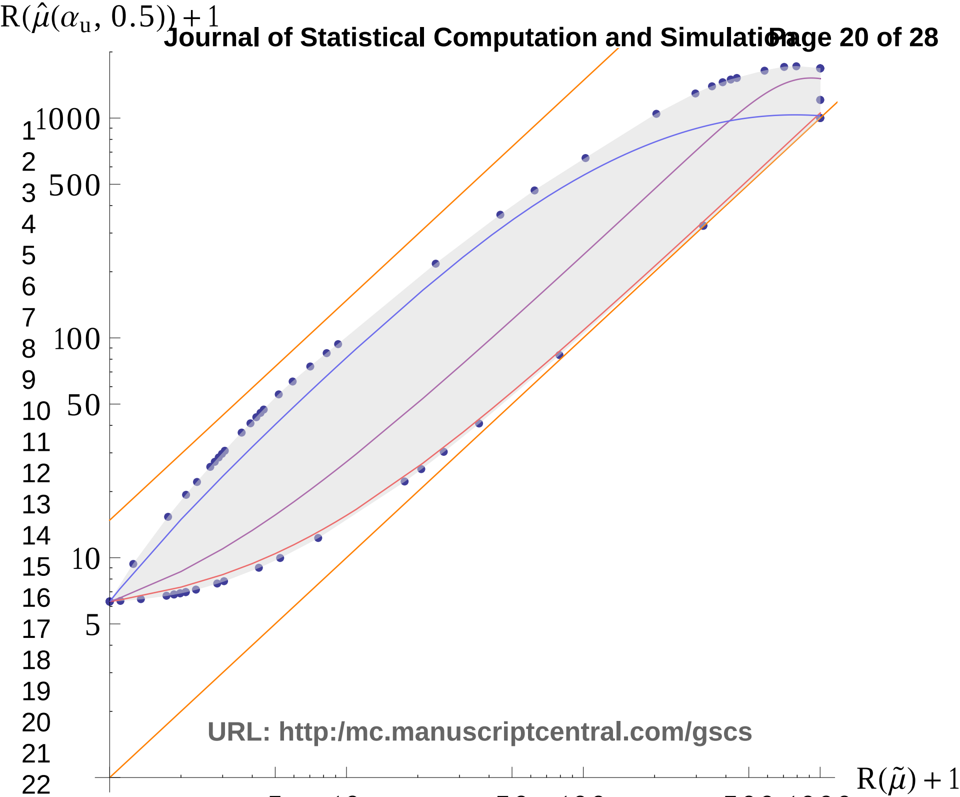
[7] Lehmann, E. L. and J. P. Romano. 2005. *Testing Statistical Hypotheses (Third Edition)*. New York: Springer.

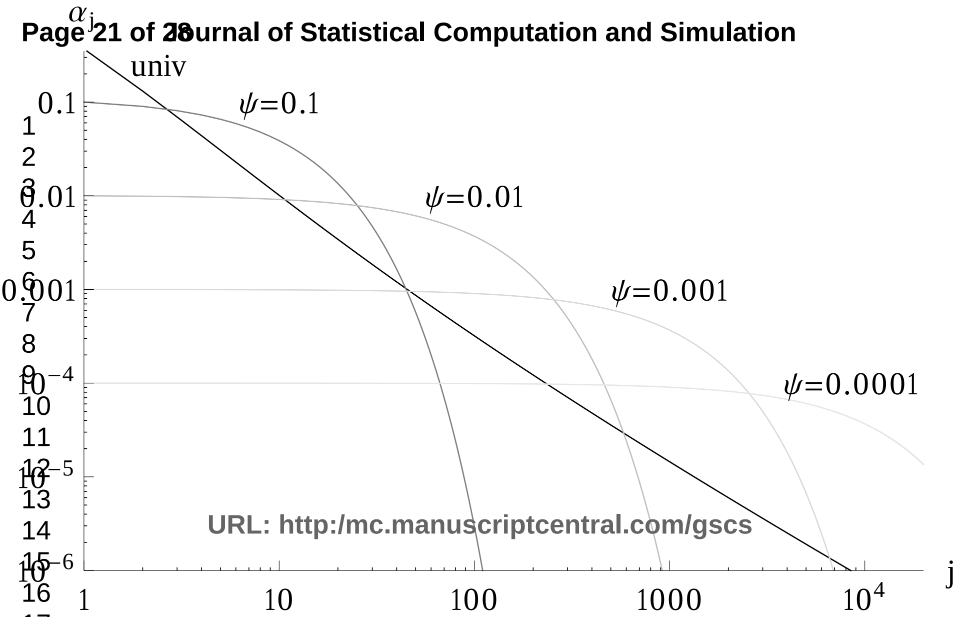
[8] Lin, D., D. Foster, and L. Ungar. 2011. "VIF regression: a fast regression algorithm for large data." *Journal of the Amer. Statist. Assoc.* 106:232–247.

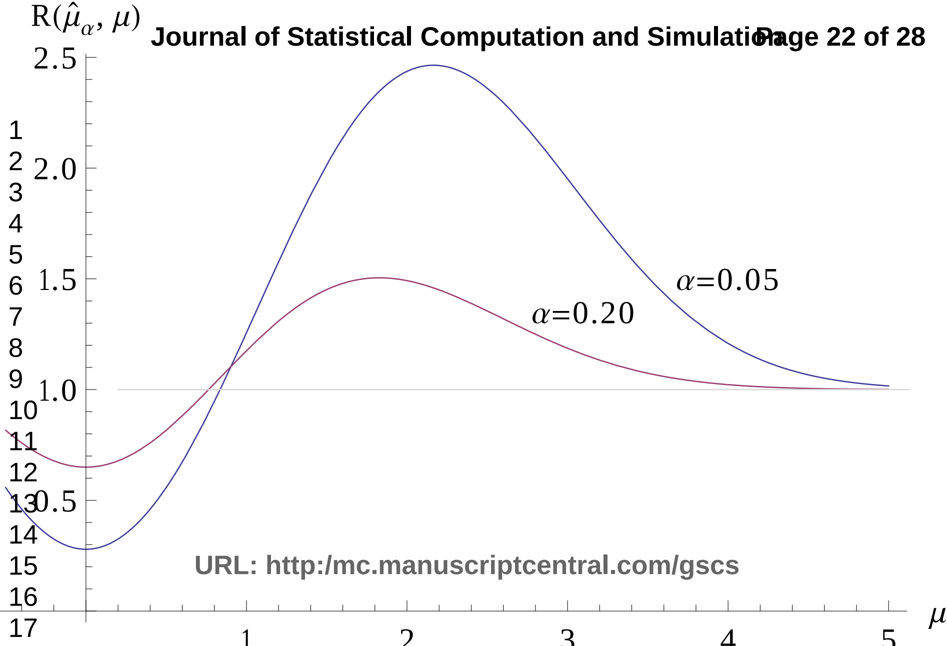
[9] Rissanen, J. 1983. "A universal prior for integers and estimation by minimum description length." *Annals of Statistics* 11:416–431.

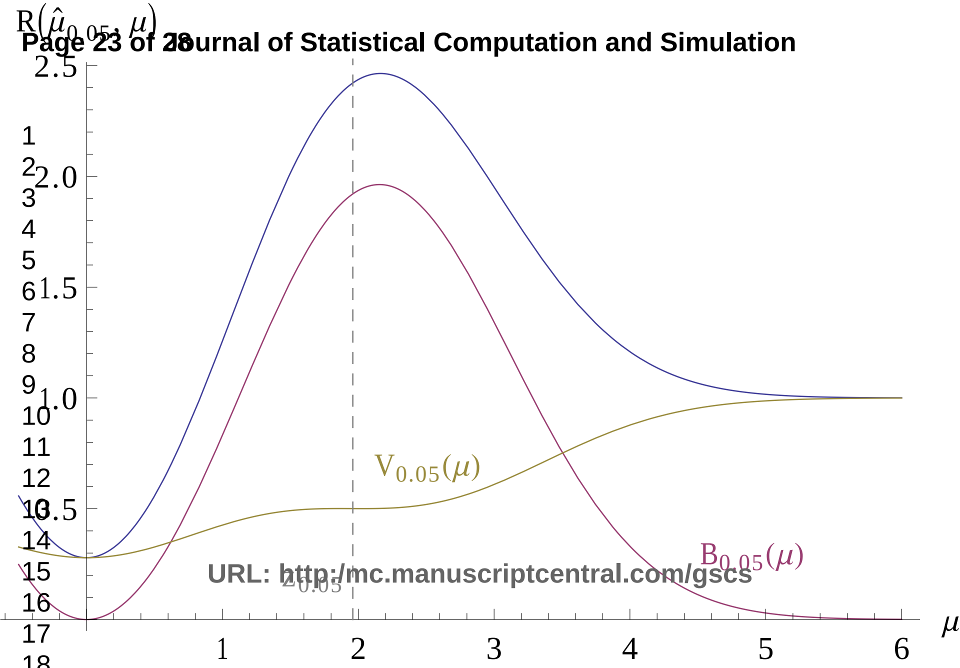
[10] Wu, X., K. Yu, W. Ding, H. Wang, and X. Zhu. 2013. "Online Feature Selection with Streaming Features." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35:1178–1191.

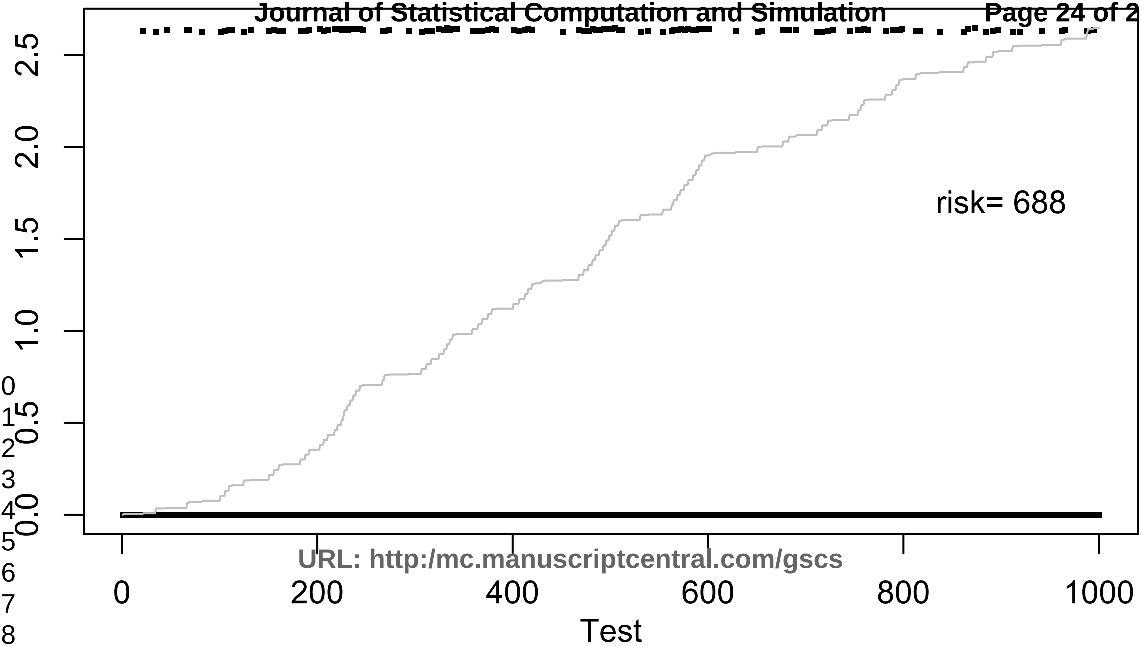

$R(\hat{\mu}(\alpha_u, 0.5)) + 1$
1500
1000
500
200
400
600
800
1000
$R(\tilde{\mu}) + 1$

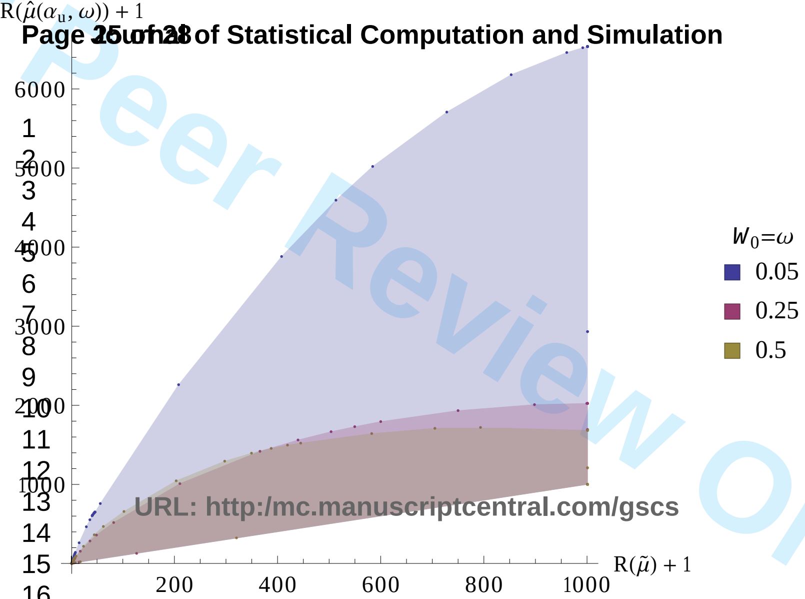

$R(\hat{\mu}(\alpha_u, 0.5)) + 1$
1000
500
100
50
10
5
5
10
50
100
500
1000
$R(\tilde{\mu}) + 1$

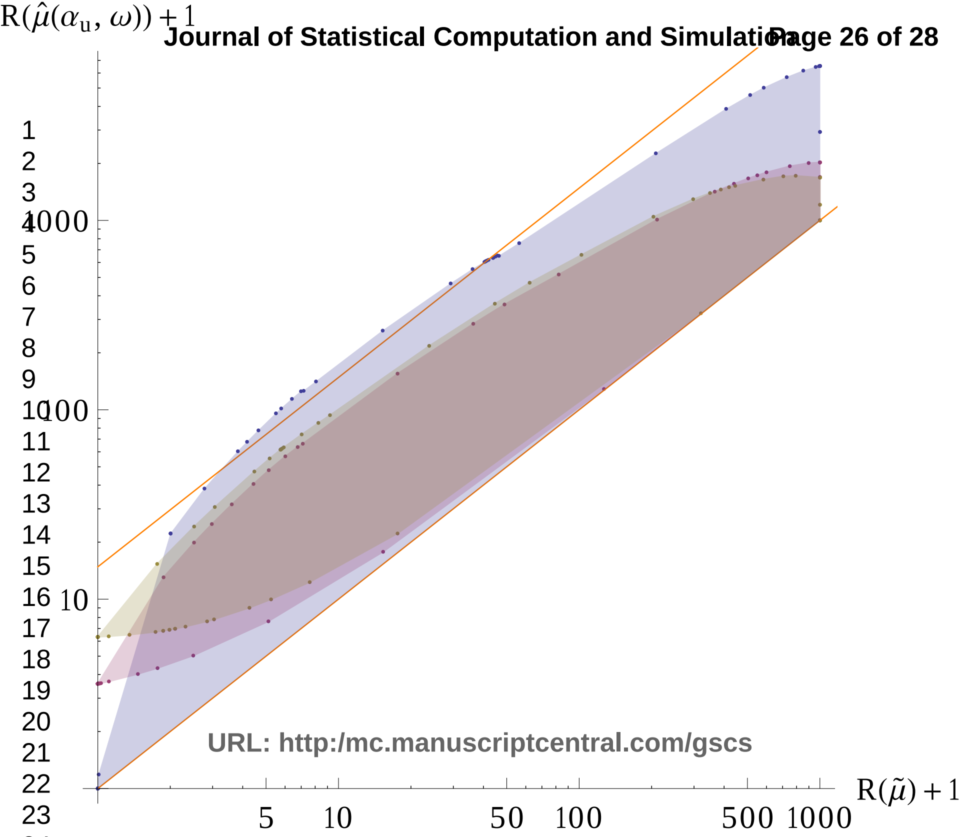

α_j
univ
ψ=0.1
ψ=0.01
ψ=0.001
ψ=0.0001
0.1
0.01
0.001
10^{-4}
10^{-5}
10^{-6}
1
10
100
1000
10^4
j

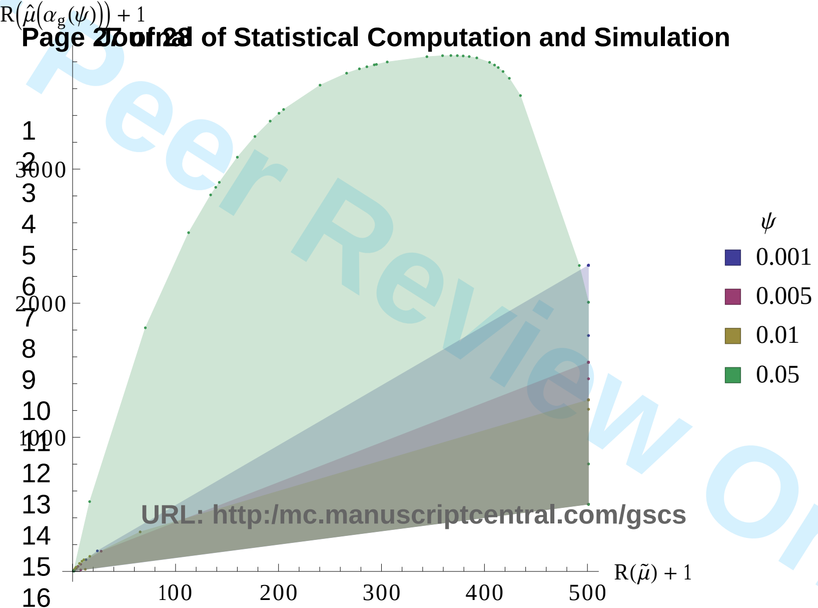

$R(\hat{\mu}_\alpha, \mu)$
2.5
2.0
1.5
1.0
0.5
α=0.05
α=0.20
1
2
3
4
5
μ

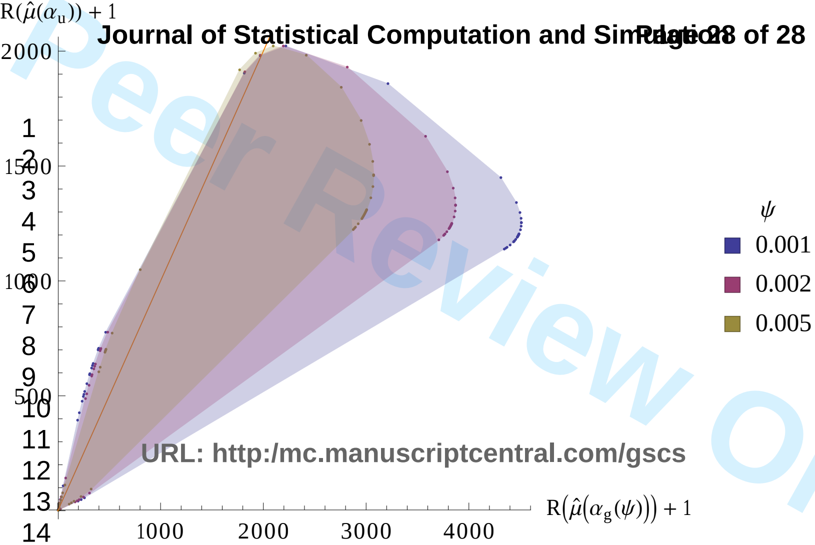

$R(\hat{\mu}_{0.05};\mu)$
2.5
2.0
1.5
1.0
0.5
$V_{0.05}(\mu)$
$B_{0.05}(\mu)$
$z_{0.05}$
1
2
3
4
5
6
μ


Mean
0.0
0.5
1.0
1.5
2.0
2.5
0
200
400
600
800
1000
Test
risk= 688


$R(\hat{\mu}(\alpha_u, \omega)) + 1$
6000
5000
4000
3000
2000
1000
200
400
600
800
1000
$R(\tilde{\mu}) + 1$
$W_0 = \omega$
0.05
0.25
0.5


$R(\hat{\mu}(\alpha_{u}, \omega))+1$
1000
100
10
5
10
50
100
500
1000
$R(\tilde{\mu})+1$


$\mathrm{R}\left(\hat{\mu}\left(\alpha_{\mathrm{g}}(\psi)\right)\right)+1$
3000
2000
1000
100
200
300
400
500
$\mathrm{R}(\tilde{\mu})+1$
ψ
0.001
0.005
0.01
0.05


$R(\hat{\mu}(\alpha_u)) + 1$
2000
1500
1000
500
1000
2000
3000
4000
$R\left(\hat{\mu}\left(\alpha_g(\psi)\right)\right) + 1$
ψ
0.001
0.002
0.005