Tight Bounds On Expected Order Statistics

Dimitris Bertsimas[*] Karthik Natarajan[†] Chung-Piaw Teo[‡§]

April 2004

Abstract

In this paper, we study the problem of finding tight bounds on the expected value of the kth order statistic $E[x_{k:n}]$ under moment information on n real-valued random variables. Given means $E[x_i] = \mu_i$ and variances $Var[x_i] = \sigma_i^2$, we show that the tight upper bound on the expected value of the highest order statistic $E[x_{n:n}]$ can be computed with a bisection search algorithm. An extremal discrete distribution is identified that attains the bound and two new closed form bounds are proposed. Under additional covariance information $Cov[x_i, x_j] = Q_{ij}$, we show that the tight upper bound on the expected value of the highest order statistic can be computed with semidefinite optimization. We generalize these results to find bounds on the expected value of the kth order statistic under mean and variance information. For $k < n$, this bound is shown to be tight under identical means and variances. All our results are distribution-free with no explicit assumption of independence made. Particularly, using optimization methods, we develop tractable approaches to compute bounds on the expected value of order statistics.

1 Introduction

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ denote $n \geq 2$ jointly distributed real-valued random variables. The order statistics of this set is a reordering of the x_i in terms of non-decreasing values, expressed as $x_{1:n} \leq$

[*] Boeing Professor of Operations Research, Sloan School of Management and Operations Research Center, Massachusetts Institute of Technology, E53-363, Cambridge, MA 02139, dbertsim@mit.edu.

[†] High Performance Computation for Engineered Systems, Singapore-MIT Alliance, Singapore 119260, karthik_natarajan@yahoo.com.

[‡] Department of Decision Sciences, NUS Business School, Singapore 117591, bizteocp@nus.edu.sg.

[§] This research was partially supported by the Singapore-MIT alliance.

$\ldots \leq x_{k:n} \leq \ldots \leq x_{n:n}$. The smallest and highest order statistics are denoted by $x_{1:n}$ and $x_{n:n}$ respectively. One of the central problems in statistics is to find, bound or approximate the expected value of order statistics under varying assumptions on the distribution of the random variables. For detailed reviews on this subject, the reader is referred to [9] and [3].

In this paper, we focus on finding bounds on the expected value of order statistics under *moment* information on the random variables. Let $\boldsymbol{x} \sim_\theta \boldsymbol{m}$ denote the set of feasible distributions θ that satisfies the given moments $\boldsymbol{m}$ for the random variables.

Definition 1 *$Z^*_{k:n}$ is a tight upper bound on the expected value of the kth order statistic if:*

$$Z^*_{k:n} = \sup_{\boldsymbol{x} \sim_\theta \boldsymbol{m}} E_\theta[x_{k:n}],$$

i.e., there exists a feasible distribution or a limit of a sequence of feasible distributions that achieves the upper bound.

No other assumptions on independence or the type of distribution are made. In this paper, we develop methods to compute $Z^*_{k:n}$ under first and second moment information on the random variables. Next, we review some of the classical bounds for order statistics.

Some Known Bounds

Given identical means and variances (μ, σ^2) for the random variables, one of the earliest known bounds for the expected highest order statistic was derived by Gumbel [10] and Hartley and David [11]. Under the assumption of independence, they obtained the upper bound $\mu + \sigma(n-1)/(2n-1)$. Moriguti [17] extended this result to the special case of symmetrically distributed random variables.

For more general distributions (not necessarily independent or identically distributed), Arnold and Groeneveld [2] obtained an upper bound on the expected value of the kth order statistic:

$$E_\theta[x_{k:n}] \leq \frac{\sum_{i=1}^n \mu_i}{n} + \sqrt{\frac{k-1}{n(n-k+1)} \sum_{i=1}^n \left[\sigma_i^2 + \left(\mu_i - \frac{\sum_{i=1}^n \mu_i}{n}\right)^2\right]}. \tag{1}$$

Under identical means and variances, this bound reduces to:

$$E_\theta[x_{k:n}] \leq \mu + \sigma\sqrt{\frac{k-1}{n-k+1}}. \tag{2}$$

For this particular case, Arnold and Groeneveld show that (2) is tight by explicitly constructing a distribution that achieves the bound. However, for general mean-variance information, (1) is not necessarily tight. Aven [4] proposed an alternative upper bound on the expected value of the highest order statistic:

$$E_\theta[x_{n:n}] \quad \leq \quad \max_{1 \leq i \leq n} \mu_i + \sqrt{\frac{n-1}{n} \sum_{i=1}^{n} \sigma_i^2}. \tag{3}$$

This bound is also not tight under general mean-variance information. In this paper, we develop an algorithmic approach to find (possibly) tight bounds on the expected value of the order statistic $Z^*_{k:n}$. We characterize cases for which the bound can be computed tractably, else we propose simple closed form bounds that seem promising.

Contributions

Our main contributions in this paper are as follows:

(a) In Section 2, we find the tight upper bound on expected value of the highest order statistic $Z^*_{n:n}$ under mean-variance information on the random variables. An efficiently solvable bisection search approach is developed to compute $Z^*_{n:n}$. A discrete extremal distribution is identified that attains the tight bound. Two simple closed form bounds for the expected highest order statistic are proposed. Under additional covariance information, we propose a semidefinite programming approach to find the tight bound on the expected highest order statistic.

(b) In Section 3, we extend the bisection search method to obtain bounds on the expected value of the general kth order statistic under mean-variance information. For $k < n$, we show that the bound is tight under identical means and variances. For general mean-variance information, the bound found with the bisection search method, while not necessarily tight, is at least as strong as (1).

(c) In Section 4, we provide computational experiments to test the performance of the different bounds. Application of the results to an option-pricing problem is considered.

2 Bounds On Expected Highest Order Statistic

We first compute the tight upper bound on the expected highest order statistic $Z^*_{n:n}$ under mean-variance information on the random variables. The mean and variance information on the random variables are denoted as $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_n^2)$. The set of feasible distributions satisfying these moment restrictions is represented by $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$. For simplicity of presentation, we will assume that all the σ_i are strictly positive. As discussed later, this condition can in fact be relaxed.

The approach to compute the tight upper bound on the expected value of the highest order statistic is based on a convex reformulation technique, initially proposed by Meilijson and Nadas [18] and developed later in Bertsimas, Natarajan and Teo [6]. The reformulation is based on the observation that the highest order statistic $x_{n:n}$ is a convex function in the x_i variables. We review the key ideas of this reformulation next.

Theorem 1 *(Bertsimas, Natarajan and Teo [6]) The tight upper bound on the expected value of the highest order statistic $Z^*_{n:n}$ given $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ is obtained by solving:*

$$Z^*_{n:n} = \min_{\boldsymbol{z}} \left(z_{n:n} + \sum_{i=1}^{n} \sup_{x_i \sim_{\theta_i} (\mu_i, \sigma_i^2)} E_{\theta_i}[x_i - z_i]^+ \right), \tag{4}$$

where $x^+ = \max(0, x)$.

Sketch of Proof. We first show that Eq. (4) provides an upper bound on $Z^*_{n:n}$. To see this, note that we have the following inequality for each variable x_i:

$$\begin{aligned} x_i &= z_i + (x_i - z_i), \\ &\leq z_{n:n} + \sum_{i=1}^{n} [x_i - z_i]^+. \end{aligned}$$

Since the right hand side of this inequality is independent of the particular i, we have:

$$x_{n:n} \leq z_{n:n} + \sum_{i=1}^{n} [x_i - z_i]^+.$$

Taking expectations and minimizing over the z_i variables, we obtain the best upper bound:

$$E_\theta[x_{n:n}] \leq \min_{\boldsymbol{z}} \left(z_{n:n} + \sum_{i=1}^{n} E_\theta[x_i - z_i]^+ \right).$$

Optimizing over distributions with given mean-variance information, we obtain an upper bound:

$$Z^*_{n:n} \leq \min_{\boldsymbol{z}} \left(z_{n:n} + \sum_{i=1}^{n} \sup_{x_i \sim_{\theta_i} (\mu_i, \sigma_i^2)} E_{\theta_i}[x_i - z_i]^+ \right).$$

Note that the inner problem is optimization over probability distributions of single random variables θ_i, since no cross moment information is specified. For a proof that the bound is tight, the reader is referred to [6]. Alternatively, we construct an extremal distribution in Theorem 3 that attains the bound. ■

The solution for the inner problem in Formulation (4) is in fact known in closed form from [13] and [22]. We outline a simple proof for this bound next.

Proposition 1 *The tight upper bound on the expected value $E_{\theta_i}[x_i - z_i]^+$ given $x_i \sim_{\theta_i} (\mu_i, \sigma_i^2)$ is:*

$$\sup_{x_i \sim_{\theta_i} (\mu_i, \sigma_i^2)} E_{\theta_i}[x_i - z_i]^+ = \frac{1}{2}\left[\mu_i - z_i + \sqrt{(\mu_i - z_i)^2 + \sigma_i^2}\right]. \qquad (5)$$

Proof. We have the basic equality:

$$[x_i - z_i]^+ \;=\; \frac{1}{2}\Big(x_i - z_i + |x_i - z_i|\Big).$$

Taking expectations, we obtain:

$$\begin{aligned} E_{\theta_i}[x_i - z_i]^+ &= \frac{1}{2}\Big(E_{\theta_i}[x_i - z_i] + E_{\theta_i}|x_i - z_i|\Big), \quad \forall x_i \sim_{\theta_i} (\mu_i, \sigma_i^2), \\ &\le \frac{1}{2}\Big(\mu_i - z_i + \sqrt{(\mu_i - z_i)^2 + \sigma_i^2}\Big), \quad \text{(From Cauchy-Schwarz inequality)}. \end{aligned}$$

Furthermore, this bound can be shown to be tight since it is attained by the distribution:

$$x_i = \begin{cases} z_i + \sqrt{(\mu_i - z_i)^2 + \sigma_i^2}, & \text{w.p. } p = \frac{1}{2}\left(1 + \frac{\mu_i - z_i}{\sqrt{(\mu_i - z_i)^2 + \sigma_i^2}}\right), \\ z_i - \sqrt{(\mu_i - z_i)^2 + \sigma_i^2}, & \text{w.p. } 1 - p = \frac{1}{2}\left(1 - \frac{\mu_i - z_i}{\sqrt{(\mu_i - z_i)^2 + \sigma_i^2}}\right). \end{cases}$$

■

Using this closed form bound, we now show that the tight upper bound on the expected highest order statistic can be found by solving a univariate convex minimization problem.

Theorem 2 *The tight upper bound on the expected value of the highest order statistic $Z^*_{n:n}$ given $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ is obtained by solving the strictly convex univariate minimization problem:*

$$Z^*_{n:n} = \min_{z \in \Re} f_{n:n}(z) = \min_{z \in \Re}\left(z + \sum_{i=1}^{n} \frac{1}{2}\left[\mu_i - z + \sqrt{(\mu_i - z)^2 + \sigma_i^2}\right]\right) \qquad (6)$$

Proof. Combining Theorem 1 and Proposition 1, the tight upper bound on the expected highest order statistic is:

$$Z^*_{n:n} = \min_{\boldsymbol{z}} \left(z_{n:n} + \sum_{i=1}^{n} \frac{1}{2} \left[\mu_i - z_i + \sqrt{(\mu_i - z_i)^2 + \sigma_i^2} \right] \right). \tag{7}$$

We next show that Formulation (7) can be simplified to a single variable optimization problem. Let $\boldsymbol{z}^*$ be an optimal solution to Problem (7) and $z^*_{n:n}$ denote the highest order statistic. Note that the second term $\sum_{i=1}^{n} \frac{1}{2} \left[\mu_i - z_i + \sqrt{(\mu_i - z_i)^2 + \sigma_i^2} \right]$ is decreasing in z_i. Hence for any $i < n$ with $z^*_{i:n} < z^*_{n:n}$, by increasing $z^*_{i:n}$ upto $z^*_{n:n}$ the first term remains unaffected while the second term decreases, thus reducing the objective. Since we are minimizing the objective, the optimal solution will set all the z^*_i values equal to $z^*_{n:n}$. ■

It can be easily checked that $f_{n:n}$ is a strictly convex function implying that the function has a unique global minimum. The optimal decision variable z^* in Formulation (6) hence satisfies the first order condition obtained by setting the derivative $\partial f_{n:n}(z^*)$ to zero:

$$\partial f_{n:n}(z^*) = \sum_{i=1}^{n} \left(\frac{z^* - \mu_i}{\sqrt{(\mu_i - z^*)^2 + \sigma_i^2}} \right) - (n-2) = 0. \tag{8}$$

Remark:

(a) Our result can be viewed as an extension of the bound from Lai and Robbins [15] and Ross [21]. In their case, under completely known marginal distributions $x_i \sim_\theta \theta_i$, they obtain the following tight bound on the highest order statistic:

$$\sup_{x_i \sim_\theta \theta_i \forall i} E_\theta[x_{n:n}] = \min_{d \in \Re} \left(d + \sum_{i=1}^{n} E_{\theta_i}[c_i - d]^+ \right) \tag{9}$$

Note that this result follows also from Meilijson and Nadas [18].

2.1 An Extremal Probability Distribution

We now construct a discrete distribution that satisfies that mean-variance requirements and attains the bound in Problem (6).

Theorem 3 *Given $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$, there is an extremal distribution for the random variables that achieves the upper bound in Problem (6).*

Proof. Let z^* denote the optimal minimizer to Problem (6). Define:

$$p_j = \frac{1}{2}\left(1 + \frac{\mu_j - z^*}{\sqrt{(\mu_j - z^*)^2 + \sigma_i^2}}\right), \quad j = 1, \ldots, n. \tag{10}$$

Clearly $p_j \geq 0$ for all j and:

$$\begin{aligned} \sum_{j=1}^n p_j &= \sum_{j=1}^n \frac{1}{2}\left(1 + \frac{\mu_j - z^*}{\sqrt{(\mu_j - z^*)^2 + \sigma_i^2}}\right), \\ &= \frac{n}{2} + \frac{2-n}{2} & \text{(From optimality condition in Eq. (8))} \\ &= 1. \end{aligned}$$

For $j = 1, \ldots, n$, we let:

$$x_i^{(j)} = \begin{cases} z^* + \sqrt{(\mu_i - z^*)^2 + \sigma_i^2}, & \text{if } i = j, \\ z^* - \sqrt{(\mu_i - z^*)^2 + \sigma_i^2}, & \text{if } i \neq j. \end{cases} \tag{11}$$

Let $\boldsymbol{x}$ take value $\boldsymbol{x}^{(j)}$ with probability p_j for $j = 1, \ldots, n$. It can be verified for this n atom distribution that:

$$\begin{aligned} E_\theta[x_i] &= \sum_{j=1}^n p_j x_i^{(j)} &= \mu_i, \quad i = 1, \ldots, n, \\ Var_\theta[x_i] &= \sum_{j=1}^n p_j (x_i^{(j)} - \mu_i)^2 &= \sigma_i^2, \quad i = 1, \ldots, n. \end{aligned}$$

Furthermore, it is easily seen from Eq. (11), that the maximum among the n random variables for the jth atom is attained by $x_j^{(j)}$. Thus:

$$E_\theta[x_{n:n}] = \sum_{j=1}^n p_j x_j^{(j)} = \left(z^* + \sum_{j=1}^n \frac{1}{2}\left[\mu_j - z^* + \sqrt{(\mu_j - z^*)^2 + \sigma_j^2}\right]\right) = f_{n:n}(z^*).$$

This n atom distribution attains the upper bound on the expected value of the highest order statistic and satisfies the mean and variance requirements. This provides an alternative proof to show that the bound in Theorem 1 is tight. ■

2.2 Solution Techniques

In general, it does not seem possible to find $Z_{n:n}^*$ in closed form. A special case under which this is possible is discussed next.

Identical mean and variance

For identical mean-variance pairs (μ, σ^2), solving Eq. (8) yields the optimal value for z^*:

$$z^* = \mu + \sigma \frac{n-2}{2\sqrt{n-1}}.$$

Substituting this into Eq. (6) yields the tight bound:

$$\sup_{x_i \sim_\theta (\mu,\sigma^2) \forall i} E_\theta[x_{n:n}] \quad = \quad \mu + \sigma\sqrt{n-1}. \tag{12}$$

Note that is exactly (2) obtained by Arnold and Groeneveld for $k = n$. A distribution that attains this bound is randomly selecting n elements without replacement from the set where one element has value $\mu + \sigma\sqrt{n-1}$ and the remaining $n-1$ elements have value $\mu - \sigma/\sqrt{n-1}$.

General mean-variance pairs

For the general case, we outline a simple bisection search algorithm to find $Z^*_{n:n}$.
Description of the algorithm:

1. Initialize z_l, z_u such that $\partial f_{n:n}(z_l) \leq 0$ and $\partial f_{n:n}(z_u) \geq 0$ and $\epsilon > 0$ to given tolerance level.

2. Let $z = \frac{z_l + z_u}{2}$.

3. While $|\partial f_{n:n}(z)| \geq \epsilon$, do

   (a) If $\partial f_{n:n}(z) >= 0$, set $z_u = z$; else set $z_l = z$.

   (b) Go back to 2.

4. Output $Z^*_{n:n} = f_{n:n}(z)$.

We propose two simple upper and lower bounds z_u and z_l on the range of the optimal z^* to initialize the algorithm. Consider the problem of finding a z_u such that $f'(z_u) \geq 0$. One such z_u is constructed such that each term on the left hand side of Eq. (8) contributes at least a fraction $(n-2)/n$:

$$\frac{z_u - \mu_i}{\sqrt{(\mu_i - z_u)^2 + \sigma_i^2}} \geq \frac{n-2}{n}, \quad i = 1, \ldots, n,$$

which reduces to:

$$z_u \geq \mu_i + \sigma_i \frac{n-2}{2\sqrt{n-1}}, \quad i = 1, \ldots, n.$$

We choose z_u as:

$$z_u = \max_{1 \leq i \leq n} \left(\mu_i + \sigma_i \frac{n-2}{2\sqrt{n-1}} \right). \tag{13}$$

Similarly, a lower bound z_l can be found such that:

$$\frac{z_l - \mu_i}{\sqrt{(\mu_i - z_l)^2 + \sigma_i^2}} \leq \frac{n-2}{n}, \quad i = 1, \ldots, n.$$

A z_l that satisfies this condition is:

$$z_l = \min_{1 \leq i \leq n} \left(\mu_i + \sigma_i \frac{n-2}{2\sqrt{n-1}} \right). \tag{14}$$

Our computational tests indicate that these values of z_u and z_l lead to the tight bound quickly.

New Closed Form Bounds

Based on the two endpoints, we now propose simple closed form bounds on the expected value of the highest order statistic.

Theorem 4 *Two closed form upper bounds on the expected value of the highest order statistic given* $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ *are:*

$$\frac{1}{2} \left(\sum_{i=1}^{n} \left[\mu_i + \sqrt{\left(\mu_i - \max_{1 \leq i \leq n} \left\{ \mu_i + \frac{n-2}{2\sqrt{n-1}} \sigma_i \right\} \right)^2 + \sigma_i^2} \right] + (2-n) \left[\max_{1 \leq i \leq n} \left\{ \mu_i + \frac{n-2}{2\sqrt{n-1}} \sigma_i \right\} \right] \right), \tag{15}$$

$$\frac{1}{2} \left(\sum_{i=1}^{n} \left[\mu_i + \sqrt{\left(\mu_i - \min_{1 \leq i \leq n} \left\{ \mu_i + \frac{n-2}{2\sqrt{n-1}} \sigma_i \right\} \right)^2 + \sigma_i^2} \right] + (2-n) \left[\min_{1 \leq i \leq n} \left\{ \mu_i + \frac{n-2}{2\sqrt{n-1}} \sigma_i \right\} \right] \right). \tag{16}$$

Proof. Substitute $z = z_l$ and $z = z_u$ in Eq. (6) respectively. ■

Note that (15) and (16) reduces to the tight upper bound (12) on the expected highest order statistic for random variables with identical mean-variance pairs.

2.3 Extensions

We now extend the results to the case where some of the $\sigma_i^2 = 0$, i.e., x_i is deterministic. Without loss of generality, we assume that exactly one variable is deterministic since the case with multiple constants can be reduced to this case by choosing the maximum of the constants. Given $n \geq 1$ random variables with strictly positive variances and a constant K, we want to find the tight upper bound on $E_\theta[\max(x_{n:n}, K)]$. By introducing an extra decision variable z_{n+1} variable for the term K, Eq. (4) reduces to:

$$\sup_{\boldsymbol{x}\sim_\theta(\boldsymbol{\mu},\boldsymbol{\sigma^2})} E_\theta[\max(x_{n:n}, K)] = \min_{\boldsymbol{z}} \left(z_{n+1:n+1} + \sum_{i=1}^{n} \frac{1}{2}\left[\mu_i - z_i + \sqrt{(\mu_i - z_i)^2 + \sigma_i^2}\right] + (K - z_{n+1})^+ \right).$$

Using an argument similar to Theorem 2, it can be checked that the optimal solution will set all the z_i values the same at a value greater than or equal to K. Hence, the tight upper bound on the expected highest order statistic is:

$$\sup_{\boldsymbol{x}\sim_\theta(\boldsymbol{\mu},\boldsymbol{\sigma^2})} E_\theta[\max(x_{n:n}, K)] = \min_{z\geq K} \left(z + \sum_{i=1}^{n} \frac{1}{2}\left[\mu_i - z + \sqrt{(\mu_i - z)^2 + \sigma_i^2}\right] \right), \tag{17}$$

which reduces to the constrained version of Formulation (6):

$$\sup_{\boldsymbol{x}\sim_\theta(\boldsymbol{\mu},\boldsymbol{\sigma^2})} E_\theta[\max(x_{n:n}, K)] \quad = \quad \min_{z\geq K} f_{n:n}(z). \tag{18}$$

The tight upper bound can be found by a modified bisection search method:

1. Solve the unconstrained version of Formulation (18) with bisection search to find z^*.

2. Output $f_{n:n}(\max(z^*, K))$.

We propose using the following two closed form bounds in this case:

$$f_{n:n}\left[\max\left(\max_{1\leq i\leq n}\left\{\mu_i + \frac{n-2}{2\sqrt{n-1}}\,\sigma_i\right\}, K\right)\right], \tag{19}$$

and:

$$f_{n:n}\left[\max\left(\min_{1\leq i\leq n}\left\{\mu_i + \frac{n-2}{2\sqrt{n-1}}\,\sigma_i\right\}, K\right)\right]. \tag{20}$$

2.4 Extensions To Additional Covariance Information

In this section, we propose an algorithmic approach to find the tight upper bound on the expected value of the highest order statistic under covariance information. Given the mean and covariance matrix for the random variables $\boldsymbol{x}\sim_\theta(\boldsymbol{\mu},\boldsymbol{Q})$, the tight upper bound is computed by finding a distribution θ that solves:

$$\begin{aligned} Z^*_{n:n} = \sup\nolimits_\theta \quad & E_\theta[x_{n:n}] \\ \text{s.t.} \quad & E_\theta[\boldsymbol{x}] = \boldsymbol{\mu}, \\ & E_\theta[\boldsymbol{x}\boldsymbol{x}'] = \boldsymbol{Q} + \boldsymbol{\mu}\boldsymbol{\mu}', \\ & E_\theta[\mathbb{I}_{\Re^n}] = 1. \end{aligned} \tag{21}$$

Here $\mathbb{I}_{\Re^n}(\boldsymbol{x}) = 1$ if $\boldsymbol{x} \in \Re^n$ and 0 otherwise represents the indicator function. This problem has been well studied under the class of *moment problems* in Isii [12] and Karlin and Studden [14]. To solve Formulation (21), we construct the dual problem by introducing variables $\boldsymbol{y}$, $\boldsymbol{Y}$ and y_0 for each of the moment constraints. The dual problem [12] is formulated as:

$$\begin{aligned} Z^* = \min \quad & \big(\boldsymbol{y}'\boldsymbol{\mu} + \boldsymbol{Y}.(\boldsymbol{Q} + \boldsymbol{\mu}\boldsymbol{\mu}') + y_0\big) \\ \text{s.t.} \quad & \boldsymbol{y}'\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{Y}\boldsymbol{x} + y_0 \geq x_{n:n}, \qquad \forall \boldsymbol{x} \in \Re^n. \end{aligned} \tag{22}$$

The constraints in Formulation (22) imply the non-negativity of a quadratic function over $\Re^n$. By taking the expectation of the dual constraints, it is easy to see that $Z^* \geq Z^*_{n:n}$. Furthermore, Isii [12] shows that if the covariance matrix $\boldsymbol{Q} \succ \boldsymbol{0}$ is strictly positive definite, then $Z^* = Z^*_{n:n}$. Under this assumption, the convexity of $x_{n:n}$ implies that the tight upper bound on the expected highest order statistic is:

$$\begin{aligned} Z^*_{n:n} = \min \quad & \big(\boldsymbol{y}'\boldsymbol{\mu} + \boldsymbol{Y}.(\boldsymbol{Q} + \boldsymbol{\mu}\boldsymbol{\mu}') + y_0\big) \\ \text{s.t.} \quad & \boldsymbol{y}'\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{Y}\boldsymbol{x} + y_0 \geq x_i, \qquad i = 1,\ldots,n,\ \forall \boldsymbol{x} \in \Re^n. \end{aligned} \tag{23}$$

Let $e^{(i)}$ denotes a unit vector with the ith component $e^{(i)}_i = 1$ and 0 otherwise. The equivalence between the global non-negativity of a quadratic polynomial and the semidefinite representation [20] implies that Formulation (23) can be rewritten as:

$$\begin{aligned} Z^*_{n:n} = \min \quad & \big(\boldsymbol{y}'\boldsymbol{\mu} + \boldsymbol{Y}.(\boldsymbol{Q} + \boldsymbol{\mu}\boldsymbol{\mu}') + y_0\big) \\ \text{s.t.} \quad & \begin{pmatrix} \boldsymbol{Y} & (\boldsymbol{y} - \boldsymbol{e_i})/2 \\ (\boldsymbol{y} - \boldsymbol{e_i})'/2 & y_0 \end{pmatrix} \succeq 0, \quad i = 1,\ldots,n. \end{aligned} \tag{24}$$

Here $A \succeq 0$ denotes the constraint that the matrix A is positive semidefinite. Formulation (24) is a semidefinite optimization problem that can be solved within $\epsilon > 0$ of the optimal solution in polynomial time in the problem data and $\log(\frac{1}{\epsilon})$ [19]. In practice, standard semidefinite optimization codes such as SeDuMi [23] can be used to find the tight upper bound on the expected highest order statistic under covariance information.

3 Bounds On Expected kth Order Statistic

In this section, we generalize our results to find bounds on the expected value of the kth order statistic for $k < n$ under mean-variance information on the random variables i.e.,

$$Z^*_{k:n} = \sup_{\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})} E_\theta[x_{k:n}].$$

Our results are based on the simple observation that:

$$x_{k:n} \leq \frac{\sum_{i=k}^{n} x_{i:n}}{n-k+1}. \tag{25}$$

We find tight bounds on the expected value of the right hand side of Eq. (25), to obtain bounds on the expected value of the kth order statistic.

Theorem 5 *The tight upper bound on the expected value of the sum of the kth to nth order statistic given $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ is obtained by solving:*

$$\sup_{\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})} E_\theta[\sum_{i=k}^{n} x_{i:n}] = \min_{z} \left((n-k+1)z + \sum_{i=1}^{n} \frac{1}{2} \left[\mu_i - z + \sqrt{(\mu_i - z)^2 + \sigma_i^2} \right] \right). \tag{26}$$

Proof. Using the result from Bertsimas, Natarajan and Teo [6], the upper bound on the sum of the expected value of the kth to nth order statistic is:

$$\sup_{\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})} E_\theta[\sum_{i=k}^{n} x_{i:n}] = \min_{\boldsymbol{z}} \left(\sum_{i=k}^{n} z_{i:n} + \sum_{i=1}^{n} \frac{1}{2} \left[\mu_i - z_i + \sqrt{(\mu_i - z_i)^2 + \sigma_i^2} \right] \right). \tag{27}$$

As before, Formulation (27) can be reduced to a single variable optimization problem. To see this, let $\boldsymbol{z}^*$ be an optimal solution to Problem (27). For any $l < k$ with $z^*_{l:n} < z^*_{k:n}$, we can increase $z^*_{l:n}$ to $z^*_{k:n}$ since the first term is unaffected ($\sum_{i=k}^{n} z^*_{i:n}$ is unaffected by change in $z^*_{l:n}$, for $l < k$, provided $z^*_{l:n} < z^*_{k:n}$) while the second term decreases in $z^*_{i:n}$. Hence, we have $z^*_{l:n} = z^*_{k:n}$ for $l < k$.

Furthermore for $l > k$ with $z^*_{l:n} > z^*_{k:n}$, by decreasing $z^*_{l:n}$ to $z^*_{k:n}$, the first term decreases at a rate of 1 while the second term increases at a rate of at most 1. Since we want to minimize our objective, we have $z^*_{l:n} = z^*_{k:n}$ for $l = 1, \dots, n$. ■

Using Eq. (25) and Theorem 5, we now obtain a bound on the expected kth order statistic.

Theorem 6 *An upper bound on the expected value of the kth order statistic $Z^*_{k:n}$ given $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ is obtained by solving:*

$$Z^*_{k:n} \leq \min_{z \in \Re} f_{k:n}(z) = \min_{z} \left(z + \sum_{i=1}^{n} \frac{1}{2(n-k+1)} \left[\mu_i - z + \sqrt{(\mu_i - z)^2 + \sigma_i^2} \right] \right). \qquad (28)$$

Note that the non-convex structure of the kth order statistic for $k < n$ implies that (28) is not necessarily tight for general mean-variance pairs. However, (28) is at least as tight as (1) proposed by Arnold and Groeneveld. This follows from observing that they obtain their bound also obtained by bounding Eq. (25), though not in the tightest manner. A special case under which (28) is tight is described next.

Identical mean and variance

For identical mean-variance pairs (μ, σ^2), Eq. (28) yields the optimal value for z^*:

$$z^* = \mu + \sigma \frac{2k - n - 2}{2\sqrt{(k-1)(n-k+1)}}.$$

Substituting this into (28) yields:

$$\sup_{x_i \sim_\theta (\mu, \sigma^2) \forall i} E_\theta[x_{k:n}] \leq \mu + \sigma \sqrt{\frac{k-1}{n-k+1}}. \qquad (29)$$

This is exactly (2) obtained by Arnold and Groeneveld. To see that (29) is tight, consider a distribution obtained by randomly selecting n elements without replacement from the set where $n - k + 1$ elements has value $\mu + \sigma\sqrt{(k-1)/(n-k+1)}$ and the remaining $k - 1$ elements have value $\mu - \sigma\sqrt{(n-k+1)/(k-1)}$. It is easy to verify that this distribution attains the bound as described above.

General mean-variance pairs

For the general case, we propose the use of the bisection search algorithm to find the bound on the expected kth order statistic by solving $\min_z f_{k:n}(z)$. The lower and upper bounds on the range of the optimal z^* to initialize the bisection search method in this case reduces to:

$$z_u = \max_{1 \le i \le n} \left(\mu_i + \sigma_i \frac{2k-n-2}{2\sqrt{(k-1)(n-k+1)}} \right), \tag{30}$$

and:

$$z_l = \min_{1 \le i \le n} \left(\mu_i + \sigma_i \frac{2k-n-2}{2\sqrt{(k-1)(n-k+1)}} \right). \tag{31}$$

Theorem 7 *Two closed form upper bounds on the expected value of the kth order statistic given $\boldsymbol{x} \sim_\theta (\boldsymbol{\mu}, \boldsymbol{\sigma^2})$ are:*

$$Z^*_{k:n} \le f_{k:n}\left(\max_{1 \le i \le n} \left(\mu_i + \sigma_i \frac{2k-n-2}{2\sqrt{(k-1)(n-k+1)}} \right) \right), \tag{32}$$

$$Z^*_{k:n} \le f_{k:n}\left(\min_{1 \le i \le n} \left(\mu_i + \sigma_i \frac{2k-n-2}{2\sqrt{(k-1)(n-k+1)}} \right) \right). \tag{33}$$

4 Computational Results

In this section, we evaluate the quality of the various bounds proposed in this paper. The first example is an application of the highest order statistic bound in a financial context. The second example is a simulation experiment to compare the performance of the bounds for the general kth order statistic. The computations were conducted on a Pentium II (550 MHz) Windows 2000 platform with the total computational time under a minute.

4.1 Application in option pricing

One of the central questions in financial economics is to find the price of a derivative security given information on the underlying assets. Under a geometric Brownian motion assumption on the prices of the underlying assets and using the no-arbitrage assumption, the Black-Scholes [7] formula provides an insightful answer to this question. Assuming no-arbitrage, but without making specific distributional assumptions, Lo [16], Bertsimas and Popescu [5] and Boyle and Lin [8] derive moment bounds on prices of options. Our particular focus is on finding bounds on the price of an option known as the *lookback option* under moment information on the asset prices.

Let $x_1, x_2, \ldots, x_n$ denote the price of an asset at n different times. A simple lookback European call option on these assets with strike price $K \geq 0$ has a payoff of $\max(x_{n:n} - K, 0)$. Let r denote the risk free interest rate and T denote the maturity date. Under the no-arbitrage assumption, the price of the lookback option is:

$$P(K) = e^{-rT} E_\theta \left[\max(x_{n:n} - K, 0)\right], \tag{34}$$

where the expectation is taken over the martingale measure. Clearly, the price of this option depends on the highest order statistic. Under mean and variance information on x_i, Boyle and Lin [8] proposed the following upper bound on the price of the lookback option:

$$P(K) \quad \leq \quad e^{-rT} \sum_{i=1}^{n} \frac{1}{2} \left[\mu_i - K + \sqrt{(\mu_i - K)^2 + \sigma_i^2}\right]. \tag{35}$$

We use the results from Section 2 to find the best bounds on $P(K)$. Note that while the asset prices are non-negative in practice, we do not model this explicitly here to compute our bounds.

The specific lookback option-pricing example is taken from Andreasen [1]. An upper bound on price of a European call lookback option over $n = 10$ time steps is calculated. The risk free interest rate (r) is 5% and the time to maturity (T) is 1 year. Table 4.1 provides the mean and variance information of the asset prices over the ten periods.

Asset	Mean μ_i	Variance σ_i^2	Asset	Mean μ_i	Variance σ_i^2
x_1	100.50	40.48	x_6	103.05	257.92
x_2	101.00	81.94	x_7	103.56	304.55
x_3	101.51	124.4	x_8	104.08	352.26
x_4	102.02	167.87	x_9	104.60	401.08
x_5	102.53	212.37	x_{10}	105.13	451.03

Table 1: Mean-variance data on asset prices from Andreasen [1].

The bounds on the option price are computed for strike prices K from 70 to 140 in steps of 10. Table 2 provides six bounds under mean-variance information and an additional bound under covariance information. For the last bound, we assumed that the asset prices were uncorrelated and solved Formulation (24) with the semidefinite optimization code SeDuMi. From Table 2, it is observed that Boyle and Lin's bound is very loose for small values of K. On average, our proposed closed form bound (19) outperforms both Arnold and Groeneveld's and Aven's bound respectively. While the closed form bound (20) is weaker for smaller K, it is in fact tight for larger K, indicating its usefulness. In Figure 1, we provide the graphical comparison of the bounds (excluding Boyle and Lin's bound which is tight only for large K).

Bound/K	70	80	90	100	110	120	130	140
Tight mean-variance bd. (18)	75.38	65.87	56.35	46.84	37.33	27.81	19.58	14.82
Our closed form bd. (19)	78.00	68.49	58.98	49.46	39.95	30.44	20.93	14.82
Our closed form bd. (20)	85.49	75.97	66.46	56.95	45.71	28.14	19.58	14.82
Boyle & Lin bd. (35)	327.97	238.52	154.36	84.85	45.71	28.14	19.58	14.82
Arnold & Groeneveld bd. (1)	81.20	68.46	57.00	47.06	38.79	32.12	26.88	22.80
Aven bd. (3)	77.79	68.28	58.77	49.25	44.38	44.38	44.38	44.38
Tight mean-var-cov bd. (24)	73.23	63.73	54.25	44.79	35.40	26.41	19.30	14.75

Table 2: Upper bound on lookback call option price from Andreasen [1].

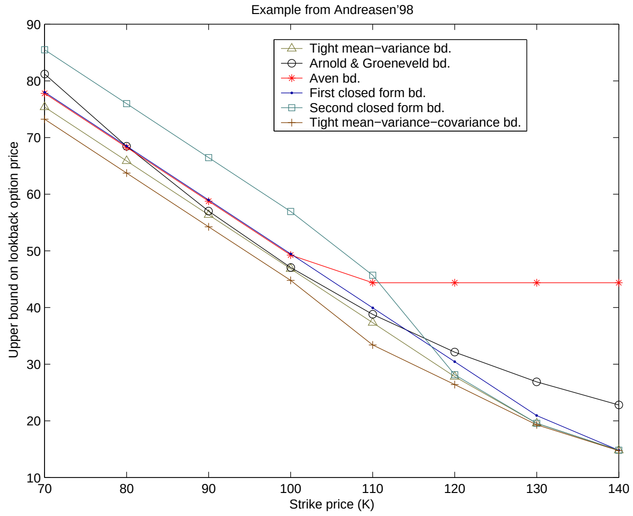

Figure 1: Upper bound on lookback call option price from Andreasen [1].

4.2 Simulation Test

The second example is a simulation test to compare the relative performance of the different bounds under randomly generated moment information. We consider $n = 30$ random variables. The mean-variance pairs for each random variable were independently chosen from a uniform distribution with $\mu_i \sim U[0, 50]$ and $\sigma_i^2 \sim U[100, 400]$. Hundred mean-variance pairs were sampled in these ranges and the bounds on the expected order statistics were computed. For each closed form bound, we evaluate the relative percentage error:

$$\text{Percentage error} = \left(\frac{\text{Closed form bound - Bisection search bound}}{\text{Bisection search bound}} \right) \times 100\%.$$

For the highest order statistic, the percentage error of the bounds are provided in Figure 2 and Table 3.

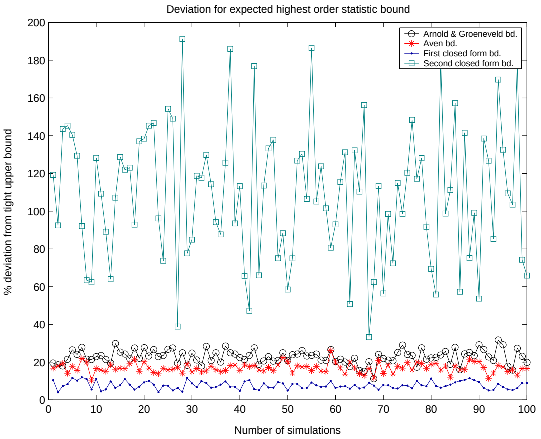

Figure 2: Deviation of closed form bounds from tight bound on expected highest order statistic.

Bound	Mean % error	Std. dev % error
Our closed form bd. (15)	7.73	1.96
Our closed form bd. (16)	108.93	35.57
Arnold & Groeneveld bd. (1)	22.86	3.64
Aven bd. (3)	16.91	2.56

Table 3: Statistics of deviation of closed form bounds for expected highest order statistic.

Note that in this case, the bisection search method finds the tight bound $Z^*_{n:n}$. In this case, our closed form bound (15) performs the best while bound (16) is relatively weaker.

We next consider the results for a smaller order statistic. Since the upper bound for the smallest order statistic $Z^*_{1:n}$ from (25) simply reduces to $\sum_{i=1}^{n} \mu_i/n$, we use the second smallest order statistic $Z^*_{2:n}$ to compare the bounds. For this case, the bisection search method does not guarantee finding the tight bound. The results obtained are presented in Figure 3 and Table 4. For this case, our closed form bound (16) is observed to be tightest among the closed form bounds with an average percentage error of about 1%.

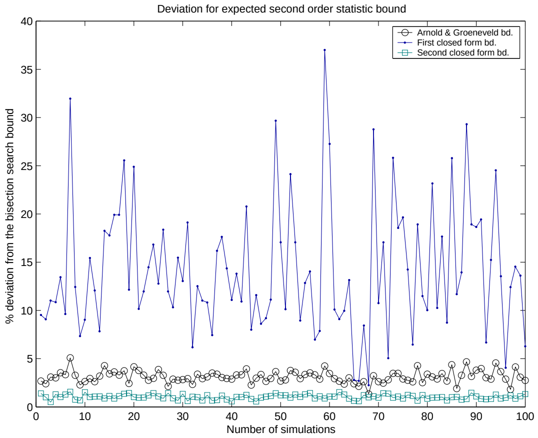

Figure 3: Deviation of closed form bounds from bisection bound on second order statistic.

Bound	Mean % error	Std. dev % error
Our closed form bd. (32)	14.21	6.74
Our closed form bd. (33)	1.04	0.25
Arnold & Groeneveld bd. (1)	3.13	0.63

Table 4: Statistics of deviation of closed form bounds for expected second order statistic.

The simulation results seem to indicate that the two closed form bounds perform well in reasonable settings. Interestingly, in each of the two simulations, the best closed form bounds were observed to be one of our bounds. While cases can be constructed for which both the bounds are weaker that either of Arnold and Groeneveld and Aven's bounds, the results suggest that the bounds are useful.

5 Summary

In this paper, we studied the problem of finding tight bounds on the expected value of order statistics under first and second moment information on the random variables. For the highest order statistic, we showed the tight upper bound could be found efficiently under mean-variance information with a bisection search method and under mean-variance-covariance information with semidefinite programming. For the general kth order statistic, we provided efficiently computable bounds (not necessarily tight) under mean-variance information. Finding tight bounds for the general kth order statistic under mean-variance and possibly covariance information is a potential research area for the future.

References

[1] Andreasen, J. 1998. The pricing of discretely sampled Asian and lookback options: a change of numeraire approach. *Journal of Computational Finance* **2**, 1, 5-30.

[2] Arnold, B. C., R. A. Groeneveld. 1979. Bounds on expectations of linear systematic statistics based on dependent samples. *Mathematics of Operations Research* **4**, No 4 441-447.

[3] Arnold, B. C., N. Balakrishnan. 1989. *Relations, Bounds and Approximations for Order Statistics.* Lecture Notes in Statistics **53**, Springer-Verlag.

[4] Aven, T. 1985. Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of Applied Probability* **22**, 723-728.

[5] Bertsimas, D., I. Popescu. 2002. On the relation between option and stock prices: A convex optimization approach. *Operations Research* **50**, 2, 358-374.

[6] Bertsimas, D., K. Natarajan and Chung Piaw Teo. 2004. Probabilistic combinatorial optimization: Moments, Semidefinite Programming and Asymptotic Bounds. To appear in *SIAM Journal of Optimization.*

[7] Black, F., M. Scholes. 1973. The pricing of options and corporate liabilities. *Journal of Political Economy.* **81**, 637-654.

[8] Boyle, P., X. S. Lin. 1997. Bounds on contingent claims based on several assets. *Journal of Financial Economics.* **46**, 383-400.

[9] David, H. A. 1981. *Order Statistics.* Second Edition, John Wiley and Sons.

[10] Gumbel , E. J. 1954. The maximum of the mean largest value and of the range. *The Annals of Mathematical Statistics* **25**, 76-84.

[11] Hartley, H. O. and H. A. David. 1954. Universal bounds for mean range and extreme observations. *The Annals of Mathematical Statistics* **25**, 85-89.

[12] Isii, K. 1963. On the sharpness of chebyshev-type inequalities. *Ann. Inst. Stat. Math* **14**, 185-197.

[13] Jagannathan, R. 1976. Minimax procedure for a class of linear programs under uncertainty. *Operations Research* **25**, No.1 173-176.

[14] Karlin, S., W. J. Studden. 1966. *Tchebycheff Systems: with Applications in Analysis and Statistics.* Pure and Applied Mathematics, A Series of Texts and Monographs. Interscience Publishers, John Wiley and Sons.

[15] Lai, T. L., H. Robbins. 1976. Maximally dependent random variables. *Proceedings of the National Academy of the Sciences* **73**, 2, 286-288.

[16] Lo, A. W. 1987. Semi-parametric upper bounds for option prices and expected payoffs. *Journal of Financial Economics.* **19**, 373-387.

[17] Moriguti, S. 1951. Extremal properties of extreme value distributions. *The Annals of Mathematical Statistics* **22**, 523-536.

[18] Meilijson, I., A. Nadas. 1979. Convex majorization with an application to the length of critical path. *Journal of Applied Probability* **16**, 671-677.

[19] Nesterov, Y., A. Nemirovkii. 1994. Interior point polynomial algorithms for convex programming. *Studies in Applied Mathematics* **13**.

[20] Parillo, P. A. 2000. Structured semidefinite programs and semi-algebraic geometry methods in robustness and optimization. PhD Thesis, California Institute of Technology.

[21] Ross, S. M. 2003. *Introduction To Probability Models.* 8th Ed. Academic Press.

[22] Scarf, H. 1958. A min-max solution of an inventory problem. K.J. Arrow, S.Karlin, H.Scarf, eds. *Studies in the mathematical theory of inventory and production.* Stanford University Press, Stanford, CA, 201-209.

[23] Sturm, J. F. SeDuMi version 1.03, Available from http://fewcal.kub.nl/sturm/software/sedumi.html.