Journal of

Applied Mathematics and Decision Sciences

Special Issue
10th Anniversary Special Issue

Hindawi Publishing Corporation
http://www.hindawi.com

Volume 2006
Number 4

Journal of Applied Mathematics and Decision Sciences

James Moffat
Defence Science and Technology Laboratory, UK
jmoffat@dstl.gov.uk

Daoji Shi
Tianjin University, China
daoji_shi@china.com

Graeme Charles Wake
Massey University, New Zealand
g.c.wake@massey.ac.nz

Journal of Applied Mathematics and Decision Sciences
Volume 2006, Number 4

Special Issue

10th Anniversary Special Issue
Guest Editors: Mahyar A. Amouzegar, Khosrow Moshirvaziri, and Roger Ríos

Contents

The 10th anniversary special issue (Editorial), *Mahyar A. Amouzegar, Khosrow Moshirvaziri, and Roger Z. Ríos-Mercado*
Volume 2006, Article ID 43435, 3 pages

Nonparametric analysis of blocked ordered categories data: some examples revisited, *D. J. Best, J. C. W. Rayner, and O. Thas*
Volume 2006, Article ID 31089, 9 pages

Stochastic dominance theory for location-scale family, *Wing-Keung Wong*
Volume 2006, Article ID 82049, 10 pages

Comparison of two common estimators of the ratio of the means of independent normal variables in agricultural research, *C. G. Qiao, G. R. Wood, C. D. Lai, and D. W. Luo*
Volume 2006, Article ID 78375, 14 pages

Effectiveness of high interest rate policy on exchange rates: a reexamination of the Asian financial crisis, *Tim Brailsford, Jack H. W. Penm, and Chin Diew Lai*
Volume 2006, Article ID 35752, 9 pages

Loss protection in pairs trading through minimum profit bounds: a cointegration approach, *Yan-Xia Lin, Michael McCrae, and Chandra Gulati*
Volume 2006, Article ID 73803, 14 pages

Mapping the convergence of genetic algorithms, *Zvi Drezner and George A. Marcoulides*
Volume 2006, Article ID 70240, 16 pages

A measure of the variability of revenue in auctions: a look at the revenue equivalence theorem, *Fernando Beltrán and Natalia Santamaría*
Volume 2006, Article ID 27417, 14 pages

A simulation framework for networked queue models: analysis of queue bounds in a G/G/c supply chain, *Mahyar Amouzegar and Khosrow Moshirvaziri*
Volume 2006, Article ID 87514, 13 pages

An analytical characterization for an optimal change of Gaussian measures, *Henry Schellhorn*
Volume 2006, Article ID 95912, 9 pages

Editorial

THE 10TH ANNIVERSARY SPECIAL ISSUE

MAHYAR A. AMOUZEGAR, KHOSROW MOSHIRVAZIRI,
AND ROGER Z. RÍOS-MERCADO

Received 12 September 2006; Accepted 12 September 2006

Ten years later

More than ten years ago, two colleagues and I developed an idea for a new journal that promised, as we stated in the preface of the first issue, "to bring together three main areas of applied mathematics, namely, classical applied mathematics, applied statistics, and operations research." We believe that we have succeeded in our goal as is evident in over 170 articles spanning such areas as biofuel production, dynamics growth games, convex and nonconvex optimization, robustness of sample correlation, global optimization, dynamical system modeling, decision support system, and much more.

Although the initial focus of this journal was to fill the gap in applied mathematics and decision sciences research that existed in New Zealand, the journal quickly grew to become an international journal with over thirty editors across the globe.

In our first issue, we also promised our readers and potential contributors that the *Journal of Applied Mathematics and Decision Sciences* (JAMDS) would "appeal to practitioners as well as theoreticians with carefully reviewed articles, be inexpensively priced, and be published rapidly." Today, we continue striving to satisfy all these goals. Every article that appeared in our journal has been diligently reviewed by two or more referees, followed by a careful examination by the editor in charge, and depending on the nature of the article, a final review by an area editor or the editor-in-chief. We continue to be a leader in our response time to authors and though our journal has been reasonably priced throughout its life, as of this year, we have made the journal available at no cost to the readers through a novel business model called *Open Access Program*.

To celebrate a decade of successful publications and our association with a new publisher, Hindawi Publishing Corporation, we have devoted this special issue to the current

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 43435, Pages 1–3
DOI 10.1155/JAMDS/2006/43435

and former editors of JAMDS. The articles that appear in this issue illustrate the quality and the diversity of our editorial board.

D. J. Best et al. present a paper on "Nonparametric analysis of blocked ordered categories data: some examples revisited." This article demonstrates the use of Cochran-Mantel-Haenszel (CMH) statistics in nonparametric analysis of general block design. Several important examples for randomized block designs with or without missing values, for balanced incomplete block designs, and for supplemented balanced designs are given and investigated. By implementing the idea on four known examples in the literature, the authors show how CMH statistics can also be applied in less standard situations. Additionally, several well-known nonparametric statistics are shown to be special cases of CMH statistics.

In "Stochastic dominance for location-scale family," Wing-Keung Wong makes an interesting contribution to the theory of mean-variance criterion by extending some results previously developed independently by Meyer, Tobin, and Levy. His results include the development of some properties for first- and second-degree stochastic-dominance efficient sets and the mean-variance efficient set.

In "Comparison of two common estimators of the ratio of the means of independent normal variables in agricultural research," Chin-Diew Lai et al. address the problems of estimating the ratio of the means of independent normal variables in agriculture research. Their results, tested in data from rice breeding multienvironment trials in Jilin, China, demonstrate the validity of this proposed approach.

In "Effectiviness of high-interest rate policy on exchange rates: a re-examination of the Asian financial crisis," Jack Penm et al. examine the effects of higher-interest rates during the Asian financial crisis. Their results indicate that sharply higher-interest rates helped support the exchange rates in various Asian countries.

Y. X. Lin et al. in the "Loss protection in pairs trading through minimum profit bounds: a cointegration approach" use cointegration principles to develop a procedure that embeds a minimum profit condition within a pairs trading strategy. Necessary conditions for such a procedure are derived and incorporated in the implementation of a five-step procedure for identifying eligible trades. Using this technique, in which its statistical validity is verified through simulation data, the author provides exploitable information on long-run time series behavior of share pairs that is not currently available in statistical methods.

In "Mapping the convergence of genetic algorithms," Zvi Drezner and George A. Marcoulides apply "MD cluster analytic" procedure, which was devised by the authors, to fully investigate the structure of the population and convergence of genetic algorithms. This is illustrated using a hybrid genetic algorithm and applying it to the well-known quadratic assignment problem (QAP). The use of the tools provided here is highly recommended and is shown to be effective in the construction of better and more efficient genetic algorithms.

F. Beltrán and N. Santamaría use simulation in "A measure of the variability of revenue in auctions: a look at the revenue equivalence theorem" to verify certain known results in auction theory, such as revenue equivalence theorem. They also attempt to develop a criterion to guide the auctioneer in deciding about the type of auction to be used. The

paper presents an interesting statistical analysis in its verification process. The variability of the results obtained about the average is measured for each type of auction, for increasing number of auctions, and for increasing number of bidders. These results are further illustrated in several companion plots.

In "A simulation framework for networked queue models: analysis of queue bounds in a G/G/c supply chain," M. Amouzegar and K. Moshirvaziri present a closed stochastic simulation network model and several approximation and bounding schemes for G/G/c systems. The analysis was, originally, conducted to verify the integrity of simulation models used to develop alternative policy options for the United States Air Force. The authors showed that the theoretical bounds could be used to derive superior approximation for mean capacities at various queues. In "An analytical characterization for an optimal change of Gaussian measures," H. Schellhorn presents an alternate characterization of the solution of an optimal control problem by considering two Gaussian measures. The author is also interested in the optimal speed of mean reversion that is shown to follow a Riccati equation. This equation is solved analytically when the volatility curve takes specific shapes. An application of the result to simulation is further discussed.

Mahyar A. Amouzegar
Khosrow Moshirvaziri
Roger Z. Ríos-Mercado

NONPARAMETRIC ANALYSIS OF BLOCKED ORDERED CATEGORIES DATA: SOME EXAMPLES REVISITED

D. J. BEST, J. C. W. RAYNER, AND O. THAS

Received 5 October 2005; Revised 12 May 2006; Accepted 15 May 2006

Nonparametric analysis for general block designs can be given by using the Cochran-Mantel-Haenszel (CMH) statistics. We demonstrate this with four examples and note that several well-known nonparametric statistics are special cases of CMH statistics.

1. Introduction

In this paper, we will use Cochran-Mantel-Haenszel (CMH) statistics to analyse four data sets which have appeared in the literature. It is well known that tests based on the CMH statistics are equivalent to certain standard rank tests but here we show how CMH statistics also apply in less standard situations. In particular, examples are given for randomized block designs both with and without missing values, for balanced incomplete block designs, and for supplemented balanced designs.

Recent descriptions of CMH statistics have been in Davis [6, Chapter 8] and Agresti [1, Chapter 7, Section 5]. We now give a very brief outline of CMH statistics, mainly from Davis [6, Chapter 8].

2. Cochran-Mantel-Haenszel statistics

The CMH statistics apply to counts N_{ijh} in which $i = 1,\dots,r$, $j = 1,\dots,c$, and $h = 1,\dots,s$. Typically, the layer index h reflects the subjects or experimental units, usually referred to as the strata; the row index i reflects the levels of the factor of interest, and the column index j reflects the values of the response variable. The marginal totals $\{n_{\cdot jh}\}$ and $\{n_{i\cdot h}\}$ for each of the s strata are taken to be fixed. For each stratum, the vector of counts $N_h = (n_{11h},\dots,n_{1ch},\dots,n_{r1h},\dots,n_{rch})^T$ has probability function

$$\left\{\prod_{i=1}^{r} n_{i\cdot h}!\right\}\left\{\prod_{j=1}^{c} n_{\cdot jh}!\right\}\Bigg/\left\{n_{\cdot\cdot h}!\prod_{i=1}^{r}\prod_{j=1}^{c} n_{ijh}!\right\}. \tag{2.1}$$

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 31089, Pages 1–9
DOI 10.1155/JAMDS/2006/31089

Initially no assumption is made about the ordering of the row and column variables: both are taken to be nominal. The null hypothesis of interest, that there is no association between row and column variables in any of the s tables, is first tested against its negation.

Davis [6, Section 8.2.2] shows that for a table consisting of only a single stratum, the CMH statistic to test for randomness in a 2×2 table is $\{(n-1)/n\}X^2$, where X^2 is the familiar Pearson test statistic $\sum(\text{observed} - \text{expected})^2/\text{expected}$. A test statistic for testing no association between row and column variables across s 2×2 tables is due to Cochran [5] and Mantel and Haenszel [11]. For an arbitrary single stratum $r \times c$ table, a test for randomness may be based on $\{(n-1)/n\}X^2$. The test for no association between s 2×2 tables can be generalized to s $r \times c$ tables. The details follow.

2.1. CMH general association statistic. Suppose now we have counts in s independent $r \times c$ tables. The test statistic may be derived by considering the vector of counts for the hth stratum, N_h, modified by removing the redundant counts for the final row and column; these are known if the row and column totals and the other row or column entries are known. We also need the expected value under the null hypothesis of no association, $E[N_h]$, and the difference, $G_h = N_h - E[N_h]$. Now $G = \sum_h G_h$ is the aggregation over all strata of $(r-1)$ $(c-1)$ differences between observation and expectation, and G has expectation zero and covariance matrix V_G, say under the null hypothesis, so that $Q_G = G^T V_G^{-1} G$ has asymptotic distribution $\chi^2_{(r-1)(c-1)}$ as the total sample size $n\ldots = \sum_h n_{\cdot\cdot h}$ approaches infinity. This is known as the CMH *general association statistic*. The Anderson [3] and McNemar [12] statistics are particular cases of the CMH general association statistic.

2.2. CMH mean score statistic. Assume now that the column variable is ordinal or interval, and that every observation in the jth column of the hth stratum is scored as b_{hj}, $j = 1,\ldots,c$. The null hypothesis, that there is no association between row and column variables in any of the s tables, is now tested against the alternative that the r row mean scores differ, on average, across strata. First, define N_{jh} as the $r-1$ vector of counts N_{ijh}, $i = 1,\ldots,r-1$, and then define $M_h = (\sum_{j=1}^{c} b_{hj}(N_{jh} - E[N_{jh}]))$ as the vector containing the first $r-1$ row sums for the hth stratum. It is routine to show that under the null hypothesis of no association $M = \sum_h M_h$ has expectation zero and covariance matrix V_M say, so that $Q_M = M^T V_M^{-1} M$ has asymptotic distribution χ^2_{r-1} as the total sample size $n\ldots = \sum_h n_{\cdot\cdot h}$ approaches infinity. The statistic Q_M is known as the CMH *mean score statistic*. If mid-rank scores are used, then if $s = 1, Q_M$ is the Wilcoxon-Mann-Whitney statistic for $r = 2$ and the Kruskal-Wallis [10] statistic for $r > 2$, while if $s > 1$ and all row totals for all strata are unity, Q_M is the Friedman [8] statistic. If the "natural" scores, $b_{hj} = j$, $j = 1,\ldots,c$, are used when $s = 1$ and $r > 2$, then a statistic due to Yates [19] is obtained.

2.3. CMH correlation statistic. Assume now that both the row and column variables are ordinal or interval, and that every observation in the ith row of the hth stratum is scored as a_{hi}, $i = 1,\ldots,r$, and that every observation in the jth column of the hth stratum is scored as b_{hj}, $j = 1,\ldots,c$. The null hypothesis, that there is no association between row

and column variables in any of the s tables, is now tested against the alternative that across strata there is a consistent association, positive or negative, between the row scores and column scores. Let C_h be a scalar given by $C_h = \sum_i \sum_j a_{hi} b_{hj} \{N_{ijh} - E[N_{ijh}]\}$. It is routine to show that under the null hypothesis of no association, $C = \sum_h C_h$ has expectation zero and variance V_C say, so that $Q_C = C^T V_C^{-1} C = C^2/V_C$ has asymptotic distribution χ_1^2 as the total sample size $n\ldots = \sum_h n_{..h}$ approaches infinity. The statistic Q_C is known as the CMH *correlation statistic*. If $s = 1$, then Q_C is $(n\ldots - 1)$ times the square of the Pearson correlation between the row and column variables; if $s = 1$ and natural scores $a_{hi} = i$ and $b_{hj} = j$ are used, then Q_C is $(n\ldots - 1)$ times the square of the Spearman correlation. The CMH correlation test is a detector of linear-linear association.

2.4. Generalized CMH statistics. Suppose that the row variable is not ordered (nominal) while the column variable is ordinal or interval, with scores $\{b_{hj}\}$. Suppose that the scores satisfy $b_{hj} = b_v(j)$ for all h with $\sum_j b_r(j) b_s(j) N_{\cdot j \cdot}/n\ldots = \delta_{rs}$. Then M is an $(r-1)$ vector with typical element $\sum_j b_v(j)\{N_{ij\cdot} - E[N_{ij\cdot}]\}$. It follows from Rayner and Best [16, Section 4.4] that M standardised is the vth component of Pearson's X^2 in the sense that the sum of the squares of the $(c-1)$ components is X^2. This order v component detects departures of the data from the model of homogeneity of row means. As before, if natural linear scores are used, the resulting test is related to that of Yates [19]. However, if the scores are quadratic, the resulting test detects dispersion differences between rows. The set of p-values resulting from applying all $(c-1)$ component tests gives a detailed and informative scrutiny of the data, albeit an informal one.

Suppose that both row and column variables are ordinal or interval, $a_{hi} = a_u(i)$ for all h with $\sum_i a_r(i) a_s(i) N_{i\cdot\cdot}/n\ldots = \delta_{rs}$, and $b_{hj} = b_v(j)$ for all h with $\sum_j b_r(j) b_s(j) N_{\cdot j \cdot}/n\ldots = \delta_{rs}$. Then $C = \sum_i \sum_j a_u(i) b_v(j)\{N_{ij\cdot} - E[N_{ij\cdot}]\}$. It follows from Rayner and Best [16, Section 8.2] that C standardised is the uvth component of Pearson's X^2, detecting departures of the data from the model of independence in the uvth bivariate moment. As previously noted, if natural linear scores are used for both row and column variables, then C is Spearman's ρ. However, if one set of scores is linear while the other is quadratic, this leads to interesting tests of bivariate skewness.

3. Randomized blocks

Possibly, the most commonly used experimental design is the randomized block design. We begin this section by illustrating how the three CMH statistics Q_G, Q_M, and Q_C introduced in the previous section are equivalent to three nonparametric rank statistics for randomized blocks.

Suppose, as in Bradley [4, page 127], that we consider measures of visual acuity for five subjects which have been given drugs designated as A, B, C, and D. The data are presented in Table 3.1. Suppose further that we wish to carry out nonparametric tests based on the within blocks (subjects) rankings for this data set. These rankings are given in parentheses in Table 3.1. We wish to use these ranks to test for equality of median drug effects, that is, to test $H_0 : \tau_A = \tau_B = \tau_C = \tau_D$ against K: not H_0, that at least two medians differ. Friedman's [8] test statistic T takes the value 8.28 with corresponding p-value 0.04 based on an χ_3^2 approximation.

Table 3.1. Visual acuity data from Bradley [4].

Drug\subject	1	2	3	4	5
A	0.39 (3)	0.21 (2)	0.73 (1)	0.41 (2)	0.65 (1)
B	0.55 (1)	0.28 (1)	0.69 (2)	0.57 (1)	0.57 (3)
C	0.33 (4)	0.19 (3)	0.64 (3)	0.28 (4)	0.53 (4)
D	0.41 (2)	0.16 (4)	0.62 (4)	0.35 (3)	0.60 (2)

Table 3.2. Stratum 1 contingency table for visual acuity data.

Drug\rank	1	2	3	4
A	0	0	1	0
B	1	0	0	0
C	0	0	0	1
D	0	1	0	0

Table 3.3. Partition of A for visual acuity data.

Source	df	SS	p-value
Friedman	3	8.28	0.04
Dispersion	3	0.60	0.90
Residual	3	1.32	0.72
Anderson	9	10.20	0.33

If we wish to test $H_0 : \tau_A = \tau_B = \tau_C = \tau_D$ against $K : \tau_A > \tau_B > \tau_C > \tau_D$, then Page's [14] test is appropriate. We find the Page test statistic L takes the value 4.7 with corresponding p-value 0.03 based on an χ^2_1 approximation. To test for the equality of the distributions of the ranks for the four drugs, we use Anderson's [3] test based on A, which here takes the value 10.20 with p-value 0.33 based on an χ^2_9 approximation.

The T statistic is simply Q_M, the L statistic is Q_C, and the A statistic is Q_G. To calculate the three CMH statistics, we need to form five 4×4 tables of counts. For subject or block 1, the 4×4 stratum table is shown as Table 3.2. Notice that each row and each column sum is one.

Software for calculating Q_G, Q_M, andQ_C is available in the IMSL, SAS, and StatXact (version 6) computer packages. To calculate Q_M and Q_C, scores are needed. To obtain T and L, the scores 1, 2, 3, and 4 are required. The usual parametric F test for mean drug differences gives $F_{3,12}$ with p-value 0.014.

Before proceeding to use the CMH approach to obtain analogues of T, L, and A for more complicated designs, we note as an aside that using the orthogonal polynomial methods of Rayner and Best [16] and Rayner et al. [17], we can partition the statistic A for randomized block designs. Results for Table 3.1 data are given in Table 3.3. The CMH approach can be used to obtain the dispersion statistic in Table 3.3 by using Q_M with the quadratic scores 9, 1, 1, 9.

4. Balanced incomplete blocks

We now illustrate the CMH approach for data from a balanced incomplete block design.

Off-flavour in six ice cream samples was rated by 15 subjects tasting four samples each. A seven-point scale was used, with "1" meaning little off-flavour and "7" meaning considerable off-flavour. The data were given in Meilgaard et al. [13, Table 7.11] and are shown here in Table 4.1. Notice that the original data in Meilgaard et al. [13, Table 7.11] is in error for subject 14, in that, a rating of "1" should be given to ice cream *F*, not to ice cream *E*. All six ice cream samples were not given to each subject as it was thought six samples were too many to evaluate at once. Sensory fatigue is well documented and often only three or four samples are judged at one sitting.

To apply CMH statistics, we form an $r \times c$ contingency table for each of the s subjects. Here r is the number of ice creams and c is the number of categories, so that $s = 15$, $r = 6$, and $c = 7$. The rows relate to ice creams and the columns to categories. Thus for each subject, a 6×7 contingency table of 0 s and 1s is formed. For example, for subject 1 the contingency table is given by Table 4.2. Of course, for a complete block design, rows *E* and *F* would have a "1" in one of the columns. Summing the 6×7 contingency tables for all 15 subjects, we obtain Table 4.3. Notice that not all rows and columns sum to one as they did for randomized blocks. Tied data would give us some column sums greater than one.

Are the six histograms whose counts are given in Table 4.3 significantly different? To answer this, we calculate Q_G, the generalized association CMH statistic, or the mean scores CMH statistic, Q_M. This can easily be done by using as data the 15 (0, 1) subject tables for IMSL [9] routine CTRAN. The SAS and StatXact routines for generalized CMH statistics will not now do all the analysis needed.

For the Table 4.3 data, we find $Q_G = 32.86$ with an approximate p-value, based on the χ^2_{29} distribution, of 0.28. Note that because here the covariance matrix is a generalized inverse of rank 29, the degrees of freedom are 29, not 30. It appears that Q_G is not too sensitive for these data. Perhaps this is because Q_G does not take into account that the data are ordered. If we use the category identifiers as scores, then we find $Q_M = 19.8$ with a p-value of 0.001 based on the χ^2_5 approximation. An F test using the same scores gives a p-value less than 0.001 according to Meilgaard et al. [13]. The F test relies on more assumptions than the test based on Q_M.

Rayner et al. [17] give an alternative analysis of the Table 3.1 data using ranks. Also note that for $r = 2$, Q_G is the Stuart [18] test of marginal homogeneity.

5. Missing values

Alvo and Cabilio [2] derive a nonparametric ranks-based test for an ordered alternative $\tau_1 \geq \tau_2 \geq \tau_3 \geq \cdots$ when the data are from a randomized block design with missing values. We now illustrate how to apply Q_C to obtain an alternative test statistic. We consider the same lymph heart pressure (in mm of Hg) as did Alvo and Cabilio [2]. These data are reproduced in Table 5.1 and concern measurements on eight toads which were dehydrated for 6-, 12-, 18-, and 24-hour periods. Biologists expect that on average, a toad's lymph heart pressure will decrease with increasing dehydration.

Table 4.1. Off-flavour ratings for six ice creams.

Subject\ice cream	A	B	C	D	E	F
1	6	1	1	2	—	—
2	6	—	—	1	3	3
3	—	4	2	—	5	2
4	7	2	3	—	2	—
5	3	5	—	1	—	1
6	—	—	1	1	3	2
7	7	4	4	—	—	3
8	2	—	1	1	1	—
9	—	2	—	2	2	3
10	4	2	—	2	5	—
11	5	—	3	—	1	1
12	—	3	2	1	—	2
13	4	2	—	—	1	1
14	5	—	2	2	—	1
15	—	2	4	5	3	—

Table 4.2. Off-flavour ratings of six ice creams for subject 1.

Ice cream\category	1	2	3	4	5	6	7
A	0	0	0	0	0	1	0
B	1	0	0	0	0	0	0
C	1	0	0	0	0	0	0
D	0	1	0	0	0	0	0
E	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0

Table 4.3. Off-flavour ratings combined for subjects.

Ice cream\category	1	2	3	4	5	6	7
A	0	1	1	2	2	2	2
B	1	5	1	2	1	0	0
C	3	3	2	2	0	0	0
D	5	4	0	0	1	0	0
E	3	2	3	0	2	0	0
F	4	3	3	0	0	0	0

To find Q_C, we need to rank the data within toads and we use eight indicator matrices or contingency tables which are similar in form to Tables 3.2 and 4.2. For toad 24, this indicator table is given as Table 5.2.

Table 5.1. Lymph heart pressure (in mm of Hg) data of Alvo and Cabilio [2].

Toad\dehydration time	6 hours	12 hours	18 hours	24 hours
21	11.9	9.8	7.6	10.2
22	5.6	4.9	4.0	3.1
23	—	14.4	14.2	7.8
24	13.3	—	—	10.0
25	8.0	7.9	—	7.6
27	17.7	16.6	15.3	11.6
28	9.0	8.0	11.9	6.8
29	9.8	8.0	7.7	7.8

Table 5.2. Rankings for toad 24 in a 4×4 table.

Hours\rank	1	2	3	4
6	1	0	0	0
12	0	0	0	0
18	0	0	0	0
24	0	1	0	0

Table 6.1. Growth of strawberry plants after applying pesticides.

Block I	Block II	Block III	Block IV
C, 107	A, 136	B, 118	O, 173
A, 166	O, 146	A, 117	C, 95
D, 133	C, 104	O, 176	C, 109
B, 166	B, 152	D, 132	A, 130
O, 177	D, 119	B, 139	D, 103
A, 163	O, 164	O, 186	O, 185
O, 190	D, 132	C, 103	B, 147

Using routine CTRAN from IMSL [9], we find $Q_C = 11.9$ with p-value 0.0006 based on an χ^2_1 approximation. Alvo and Cabilio [2] found that for these data, their recommended test statistic took the value 226.75 and quoted exactly the same p-value as do we, namely 0.0006.

6. Supplemented balance designs

Pearce [15] suggested the use of supplemented balanced designs and used these designs to analyse data when pesticides designated as A, B, C, D, and O are applied to strawberry plants. The pesticides were intended to control weeds and allow the strawberry plants to grow bigger and presumably produce more strawberries. However, while eradicating the weeds, do the pesticides inhibit strawberry growth? Pearce [15] gave the results and the design that we reproduce in Table 6.1. The figures quoted represent the spread of the strawberry plants. Pesticide "O" is a control.

Table 6.2. Rankings for block I in a 5×8 table.

Drug\rank	1	2	3	3.5	4	5	6	7
A	0	0	0	1	0	1	0	0
B	0	0	0	1	0	0	0	0
C	0	0	0	0	0	0	0	1
D	0	0	0	0	0	0	1	0
O	1	1	0	0	0	0	0	0

To use CMH to obtain an analogue of Friedman's T for this more complex design, we proceed as before, ranking within blocks and forming four 5×8 indicator matrices. Notice in block I, there are two tied observations. Table 6.2 shows the indicator matrix for this block.

We find $Q_M = 20.1$ with a p-value of 0.0005 based on an χ^2_4 approximation. An F test based on a regression routine gives, for these data, $F_{4,20} = 24.6$ with p-value less than 0.0001. Desu and Raghavarao [7] give an analogue of Friedman's T for general block designs that have the same asymptotic chi-squared distribution as T. For Table 6.1 data, their statistic has the value 20.0, almost identical to Q_M. Perhaps the difference is in the treatment of the tied observations.

References

[1] A. Agresti, *Categorical Data Analysis*, 2nd ed., Wiley Series in Probability and Statistics, John Wiley & Sons, New York, 2002.

[2] M. Alvo and P. Cabilio, *Testing ordered alternatives in the presence of incomplete data*, Journal of the American Statistical Association **90** (1995), no. 431, 1015–1024.

[3] R. L. Anderson, *Use of contingency tables in the analysis of consumer preference studies*, Biometrics **15** (1959), 582–590.

[4] J. V. Bradley, *Distribution-Free Statistical Tests*, Prentice-Hall, New Jersey, 1968.

[5] W. G. Cochran, *Some methods for strengthening the common χ^2 tests*, Biometrics **10** (1954), 417–451.

[6] C. S. Davis, *Statistical Methods for the Analysis of Repeated Measurements*, Springer Texts in Statistics, Springer, New York, 2002.

[7] M. M. Desu and D. Raghavarao, *Nonparametric Statistical Methods for Complete and Censored Data*, Chapman & Hall/CRC, Florida, 2004.

[8] M. Friedman, *The use of ranks to avoid the assumption of normality implicit in the analysis of variance*, Journal of the American Statistical Association **32** (1937), no. 200, 675–701.

[9] IMSL, *IMSL User's Guide-Mathematical & Statistical Functions*, Houston: Visual Numerics. 1995.

[10] W. H. Kruskal and W. A. Wallis, *Use of ranks in one-criterion analysis of variance*, Journal of the American Statistical Association **47** (1952), no. 260, 583–621.

[11] N. Mantel and W. Haenszel, *Statistical aspects of the analysis of data from retrospective studies of disease*, Journal of the National Cancer Institute **22** (1959), 719–748.

[12] Q. McNemar, *Note on the sampling error of the difference between correlated proportions or percentages*, Psychometrika **12** (1947), no. 2, 153–157.

[13] M. Meilgaard, G. V. Civille, and B. T. Carr, *Sensory Evaluation Techniques*, 3rd ed., CRC Press, Florida, 1999.

[14] E. B. Page, *Ordered hypotheses for multiple treatments: a significance test for linear ranks*, Journal of the American Statistical Association **58** (1963), no. 301, 216–230.
[15] S. C. Pearce, *Supplemented balance*, Biometrika **47** (1960), no. 3-4, 263–271.
[16] J. C. W. Rayner and D. J. Best, *A Contingency Table Approach to Nonparametric Testing*, Chapman & Hall/CRC, Florida, 2001.
[17] J. C. W. Rayner, D. J. Best, P. B. Brockhoff, and G. D. Rayner, *Nonparametrics for Sensory Science: A More Informative Approach*, Blackwell, Iowa, 2005.
[18] A. Stuart, *A test for homogeneity of the marginal distributions in a two-way classification*, Biometrika **42** (1955), no. 3-4, 412–416.
[19] F. Yates, *The analysis of contingency tables with groupings based on quantitative characters*, Biometrika **35** (1948), no. 1-2, 176–181.

D. J. Best: School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia
E-mail address: donald.j.best@newcastle.edu.au

J. C. W. Rayner: School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia
E-mail address: john.rayner@newcastle.edu.au

O. Thas: Department of Applied Mathematics, Biometrics and Process Control, Ghent University, 9000 Gent, Belgium
E-mail address: olivier.thas@ugent.be

STOCHASTIC DOMINANCE THEORY FOR LOCATION-SCALE FAMILY

WING-KEUNG WONG

Received 17 January 2006; Revised 1 August 2006; Accepted 2 August 2006

Meyer (1987) extended the theory of mean-variance criterion to include the comparison among distributions that differ only by location and scale parameters and to include general utility functions with only convexity or concavity restrictions. In this paper, we make some comments on Meyer's paper and extend the results from Tobin (1958) that the indifference curve is convex upwards for risk averters, concave downwards for risk lovers, and horizontal for risk neutral investors to include the general conditions stated by Meyer (1987). We also provide an alternative proof for the theorem. Levy (1989) extended Meyer's results by introducing some inequality relationships between the stochastic-dominance and the mean-variance efficient sets. In this paper, we comment on Levy's findings and show that these relationships do not hold in certain situations. We further develop some properties among the first- and second-degree stochastic dominance efficient sets and the mean-variance efficient set.

1. Introduction

Mean-variance (MV) efficient sets have been widely used in both economics and finance to analyze how people make their choices among risky assets. Markowitz [21] demonstrated that if the ordering of alternatives is to satisfy the von Neumann-Morgenstern [39] (NM) axioms of rational behavior, only a quadratic (NM) utility function is consistent with an ordinal expected utility function that depends solely on the mean and variance of the return. Thereafter, Feldstein [7], Hanoch and Levy [12], Rothschild and Stiglitz [31, 32], and others commented that the MV criterion is applicable only when the decision maker's utility function is quadratic and the probability distribution of return is normal. Moreover, Baron [2] pointed out that even if the return for each alternative has a normal distribution, the MV framework cannot be used to rank alternatives consistently

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 82049, Pages 1–10
DOI 10.1155/JAMDS/2006/82049

with the NM axioms unless a quadratic NM utility function is specified. Meyer [25] extended the MV theory to include general utility functions and comparison between distributions that differ only by location and scale parameters.

Meyer's extensions are important as it is well known that the distribution of investment returns is usually nonnormal and the restriction of the utility function to the quadratic form is too limited in scope. These restrictions were popular in the literature not only before Meyer's findings but remained common after Meyer's findings. For example, Zhao and Ziemba [45] restricted the use of mean-variance criterion to normal or log-normal distributions and the quadratic utility function. Chow [4] pointed out that the mean-variance portfolio theory assumes that investor utility functions are quadratic and/or the return distributions of assets are multivariate normal. In this paper, we make some comments on Meyer's paper and extend the results from Tobin [37], who postulated that the indifference curve is convex upwards for risk averters and is concave downwards for risk lovers, to include a wide family of distributions for the returns as well as to include general utility functions as stated in Meyer [25]. We also provide an alternative proof for the theorem.

Levy [16] extended Meyer's results to prove that the first- (FSD) and second-degree (SSD) stochastic dominance efficient sets are equal to the mean-variance (MV) efficient set under certain conditions and established some inequality relationships between the variables in the same location-scale family. In this paper, we comment on Levy's findings and show that the inequality relationships developed by Levy do not hold in certain situations. We further explore the relationships among the FSD, SSD, and MV efficient sets, which culminate in three important findings: (1) the SSD efficient set is a proper subset of the FSD efficient set, (2) the SSD efficient set is a proper subset of the MV efficient set, and (3) the FSD efficient set is not equal to the MV efficient set in a way that neither is a proper subset of each other.

Being of both theoretical and practical interests, the main challenge of the MV and SD analyses is to identify the assets that constitute attainable efficient portfolios. Unfortunately, the relationships between the MV efficient sets and the SD efficient sets have not been well established. With this in mind, we seek to develop the relationships between the MV and SD efficient sets to capture the essence of portfolio selection here. In addition, we explore the shapes of indifference curves for risk averters, risk lovers, and risk-neutral investors. Our findings could be useful in facilitating the MV and SD procedures and enabling investors to make wiser decisions in their investments.

We begin by introducing a brief literature in this section. In Section 2, we first review, discuss, and give comments on some properties stated in Meyer [25], Levy [16], and Sinn [34]. We then proceed to develop some properties on the expected utility maximization and the stochastic dominance theory for the location-scale family. The concluding remarks are in Section 3.

2. Theory

In this section, we first review and discuss some properties stated in Meyer [25], Levy [16], and Sinn [34], and further extend their work by developing some additional properties.

In order to avoid confusion, we use "proposition" to state our results and "property" to state the results produced by Meyer [25] and Levy [16].

Let the return X be the random variable with zero mean and variance one, with the location-scale family $\mathcal{D}$ generated by X such that

$$\mathcal{D} = \{Y \mid Y = \mu + \sigma X,\ -\infty < \mu < \infty,\ \sigma > 0\}. \tag{2.1}$$

The expected utility $V(\sigma,\mu)$, see Meyer [25], for the utility U on the random variable Y can then be expressed as

$$V(\sigma,\mu) = E[U(Y)] = \int_a^b u(\mu + \sigma x)dF(x), \tag{2.2}$$

where $[a,b]$ is the support of X, F is the distribution function of X, and the mean and variance of Y are μ and σ^2, respectively. We note that the requirement of the zero mean and unit variance for X is not necessary. However, without loss of generality, we can make these assumptions as we will always be able to find such a seed random variable in the location-scale family.

For any constant α, the indifference curve drawn on the (σ,μ) plane such that $V(\sigma,\mu)$ is a constant can be expressed as

$$C_\alpha = \{(\sigma,\mu) \mid V(\sigma,\mu) \equiv \alpha\}. \tag{2.3}$$

In the indifference curve, we follow Meyer [25] to have

$$V_\mu(\sigma,\mu)d\mu + V_\sigma(\sigma,\mu)d\sigma = 0 \tag{2.4}$$

or

$$V_\mu(\sigma,\mu)\frac{d\mu}{d\sigma} + V_\sigma(\sigma,\mu) = 0, \tag{2.5}$$

where

$$\begin{aligned} V_\mu(\sigma,\mu) &= \frac{\partial V(\sigma,\mu)}{\partial \mu} = \int_a^b u'(\mu + \sigma x)dF(x), \\ V_\sigma(\sigma,\mu) &= \frac{\partial V(\sigma,\mu)}{\partial \sigma} = \int_a^b u'(\mu + \sigma x)x\,dF(x). \end{aligned} \tag{2.6}$$

The following proposition is then obtained by applying Meyer [25, Properties 1 and 2] and the implicit function theorem.

Proposition 2.1. *If the distribution function of the return with mean μ and variance σ^2 belongs to a location-scale family and for any utility function u, if $u' > 0$, then the indifference curve C_α can be parameterized as $\mu = \mu(\sigma)$ with slope*

$$S(\sigma,\mu) = -\frac{V_\sigma(\sigma,\mu)}{V_\mu(\sigma,\mu)}. \tag{2.7}$$

In addition,

(1) *if* $u'' \le 0$, *then the indifference curve* $\mu = \mu(\sigma)$ *is an increasing function of* σ; *and*

(2) *if* $u'' \ge 0$, *then the indifference curve* $\mu = \mu(\sigma)$ *is a decreasing function of* σ.

Proof. From (2.6), we have

$$S(\sigma,\mu) = -\frac{\int_a^b u'(\mu+\sigma x)x\,dF(x)}{\int_a^b u'(\mu+\sigma x)dF(x)} \tag{2.8}$$

in which $\int_a^b u'(\mu+\sigma x)dF(x) > 0$ because $u' > 0$. For the numerator, as $E(X) = 0$, we have $\int_a^0 x\,dF(x) = -\int_0^b x\,dF(x)$. If $u'' < 0$, we have

$$\begin{aligned}\int_0^b u'(\mu+\sigma x)x\,dF(x) &< \int_0^b u'(\mu)x\,dF(x) = -\int_a^0 u'(\mu)x\,dF(x)\\ &< -\int_a^0 u'(\mu+\sigma x)x\,dF(x).\end{aligned} \tag{2.9}$$

Hence, $S(\sigma,\mu) > 0$. Similarly, if $u'' > 0$, we have $S(\sigma,\mu) < 0$. □

Meyer [25] continued to investigate the properties of $\partial S(\sigma,\mu)/\partial\mu$ without the restriction of $V(\sigma,\mu) \equiv \alpha$ and obtained the following property (we refer to Property 5 in Meyer's paper).

Property 2.2. $\partial S(\sigma,\mu)/\partial\mu \le (=\ge)0$ for all μ and for all $\sigma \ge 0$ if and only if $u(\mu+\sigma x)$ displays decreasing (constant, increasing) absolute risk aversion.

We note that Sinn [34] obtained similar results as the above property in Meyer's paper. But similar to Meyer's approach, the proof of the results in Sinn [34] was also done without the restriction of $V(\sigma,\mu) \equiv \alpha$. It should be equally important to study the convexity of the indifference curve C_α with the restriction of $V(\sigma,\mu) \equiv \alpha$. Under the constraint of $(\sigma,\mu) \in C_\alpha$, we have the following proposition for $\partial S(\sigma,\mu)/\partial\sigma$ as a complement of Meyer's Property 5 and Sinn's work.

Proposition 2.3. *The distribution function of the return with mean* μ *and variance* σ^2 *belongs to a location-scale family. For any utility function* u *with* $u' > 0$,

(1) *if* $u'' \le 0$, *then* $\mu = \mu(\sigma)$ *is a convex function of* σ, *and*

(2) *if* $u'' \ge 0$, *then* $\mu = \mu(\sigma)$ *is a concave function of* σ.

Proof. As

$$\frac{d\mu}{d\sigma} = -\frac{\int_a^b u'(\mu+\sigma x)x\,dF(x)}{\int_a^b u'(\mu+\sigma x)dF(x)} = -\frac{I_1}{I_2}, \tag{2.10}$$

we have

$$\begin{aligned}
\frac{d^2\mu}{d\sigma^2} &= \frac{1}{I_2^2}\left(I_1\frac{\partial I_2}{\partial\sigma} - I_2\frac{\partial I_1}{\partial\sigma}\right)\\
&= \frac{I_1}{I_2^2}\int_a^b u''(\mu+\sigma x)\left(\frac{d\mu}{d\sigma}+x\right)dF - \frac{1}{I_2}\int_a^b u''(\mu+\sigma x)\left(\frac{d\mu}{d\sigma}+x\right)x\,dF\\
&= -\frac{1}{I_2}\frac{d\mu}{d\sigma}\int_a^b u''(\mu+\sigma x)\left(\frac{d\mu}{d\sigma}+x\right)dF - \frac{1}{I_2}\int_a^b u''(\mu+\sigma x)\left(\frac{d\mu}{d\sigma}+x\right)x\,dF\\
&= -\frac{1}{I_2}\int_a^b u''(\mu+\sigma x)\left(\frac{d\mu}{d\sigma}+x\right)^2 dF \qquad (2.11)\\
&= -\frac{\int_a^b u''(\mu+\sigma x)(d\mu/d\sigma+x)^2 dF}{\int_a^b u'(\mu+\sigma x)dF}\\
&\geq (>)0 \quad \text{as } u'>0,\ u''\leq(<)0\\
&\leq (<)0 \quad \text{as } u'>0,\ u''\geq(>)0.
\end{aligned}$$

□

The above proposition can be easily extended to include the situation in which $u' \geq 0$ and $u'' \leq 0$ and the situation in which $u' \geq 0$ and $u'' \geq 0$ with the condition Prob$(u' > 0) > 0$. It may be rewritten as the indifference curve C_α is convex upwards for risk averters, concave downwards for risk lovers, and horizontal for risk neutral investors.

In addition, we note that Tobin [37] had proven the above proposition on the quadratic utility functions with the normality assumption for the distributions of the return. Our proposition is then an extension of Tobin [37] results to include the general utility functions, as well as the distributions in the location and scale family as in Meyer's paper. Furthermore, since Sinn [34] also obtained similar results for risk averters, our proof is an alternative to the results reported by Tobin and Sinn.

Levy [16] stated the first-degree stochastic dominance (FSD), the second-degree stochastic dominance (SSD), and the mean-variance (MV) rules (Levy called it mean-standard deviation rule); and defined the FSD, SSD, and MV efficient sets (see Levy for the detailed definitions). He also extended Meyer's results to prove that the first- and second-degree stochastic dominance efficient sets are equal to mean-variance efficient set under certain conditions and showed the relationships between the support of the seed random variable X and the parameters in the two linear functions Y_i and Y_j of X in the following property (Levy termed it as "proposition" in his paper).

Property 2.4. Let X be a random variable with a finite mean and variance, but with no further restriction on its distribution, and let Y_i and Y_j differ from X by location and scale parameters, such that $Y_i = \alpha_i + \beta_i X$, $Y_j = \alpha_j + \beta_j X$. The support of X is $[a,b]$. Then

(1) Y_i and Y_j are in the MV-efficient set for all nondecreasing preferences if and only if

$$a < \frac{\alpha_j - \alpha_i}{\beta_i - \beta_j}. \qquad (2.12)$$

(2) (a) If Y_i dominates Y_j in MV, then such dominance exists in expected utility (EU) for all risk-averse investors with no additional restriction on $F(X)$.
(b) However, a dominance in EU for all nondecreasing U exists, if and only if

$$b \leq \frac{\alpha_i - \alpha_j}{\beta_j - \beta_i}. \tag{2.13}$$

If (2.12) holds, no dominance by MV implies no dominance for all nondecreasing U and also no dominance for all nondecreasing concave U. If (2.12) holds and (2.13) does not hold, the MV- and EU-efficient sets are identical when risk aversion is assumed. If both (2.12) and (2.13) hold, the MV- and EU-efficient sets are identical for all nondecreasing preference U.

Next, we study the relationships among the efficient sets for the FSD, SSD, and MV rules for the location-scale family, and the validity of the above property in Levy. Letting $\mathcal{D}_{\text{FSD}}$, $\mathcal{D}_{\text{SSD}}$, and $\mathcal{D}_{\text{MV}}$ be the FSD efficient set, the SSD efficient set, and the MV efficient set, respectively, we obtain the following proposition.

Proposition 2.5. *For any location-scale family,*
(1) $\mathcal{D}_{\text{SSD}} \subset \mathcal{D}_{\text{FSD}}$*;*
(2) $\mathcal{D}_{\text{SSD}} \subset \mathcal{D}_{\text{MV}}$*; and*
(3) (a) $\mathcal{D}_{\text{MV}} - \mathcal{D}_{\text{FSD}} \neq \varnothing$*, and*
(b) $\mathcal{D}_{\text{FSD}} - \mathcal{D}_{\text{MV}} \neq \varnothing$.

Proof. Since $X \succ_1 Y \Rightarrow X \succ_2 Y$, we obtain part (1) of Proposition 2.5. The following is a simple example to show that $\mathcal{D}_{\text{SSD}} \neq \mathcal{D}_{\text{FSD}}$.

Example 2.6. $Y = \beta X$, where $0 < \beta < 1$ and $E(X) = 0$.

In this example, $Y \succ_2 X$ but X and Y do not dominate each other in the sense of FSD. Hence, $(X, Y) \in \mathcal{D}_{\text{FSD}}$ but $(X, Y) \notin \mathcal{D}_{\text{SSD}}$. Thus, part (1) of the proposition holds.

Applying Hadar and Russell [10, Theorem 4], Tesfatsion [36, Theorem 1′], or Li and Wong [20, Theorem 8b], we find that $\mathcal{D}_{\text{SSD}}$ is a subset of $\mathcal{D}_{\text{MV}}$. To show that $\mathcal{D}_{\text{SSD}}$ is a proper subset of $\mathcal{D}_{\text{MV}}$, we use the following example.

Example 2.7. Let X be the seed random variable with support $[a, b] = [0, 1]$, let $Y_i = \alpha_i + \beta_i X$, and let $Y_j = \alpha_j + \beta_j X$, and set $\beta_i > \beta_j > 0$ and $\alpha_i = \alpha_j + \beta_i - \beta_j$.

In this example, $(Y_i, Y_j) \in \mathcal{D}_{\text{MV}}$ but $(Y_i, Y_j) \notin \mathcal{D}_{\text{SSD}}$. Hence, $\mathcal{D}_{\text{SSD}}$ is a proper subset of $\mathcal{D}_{\text{MV}}$ and thus part (2) of the proposition holds.

Example 2.7 can also be used to prove (3a). In this example, $(Y_i, Y_j) \in \mathcal{D}_{\text{MV}}$ but $(Y_i, Y_j) \notin \mathcal{D}_{\text{FSD}}$. Hence, (3a) holds.

One can also easily postulate that Example 2.6 can be used to show (3b) as $(X, Y) \in \mathcal{D}_{\text{FSD}}$ but $(X, Y) \notin \mathcal{D}_{\text{MV}}$. □

It is well established that the FSD efficient set is equivalent to the EU efficient set for all nondecreasing preference structures U, the SSD efficient set is equivalent to the EU efficient set for all nondecreasing concave U; see, for example, Hanoch and Levy [12], Hadar and Russell [10], Meyer [24], and Li and Wong [20]. From part (1) of the above

proposition, we know that the SSD efficient set is a subset of the FSD efficient set. Hence, we can define a complement of the SSD efficient set within the FSD efficient set, denoted by $\mathcal{D}_{\text{SSD}}^{c}$, to be the efficient set for all nondecreasing preference U but not for any nondecreasing concave U. We have

$$\mathcal{D}_{\text{FSD}} = \mathcal{D}_{\text{SSD}} \cup \mathcal{D}_{\text{SSD}}^{c} \tag{2.14}$$

and $\mathcal{D}_{\text{SSD}}^{c}$ is not an empty set. In the proof of parts (2) and (3) in the above proposition, we simply utilize $(Y_i, Y_j) \in \mathcal{D}_{\text{SSD}}^{c}$ such that the results hold.

Lastly, we valuate the validity of Levy's property. It is easy to find that Example 2.7 in the above can be used to show that parts (1) and (2b) in Levy's property may not hold. In this example, we illustrate that $(Y_i, Y_j) \in \mathcal{D}_{\text{MV}}$ but (2.12) does not hold as

$$\frac{\alpha_j - \alpha_i}{\beta_i - \beta_j} = \frac{\alpha_j - \alpha_j - \beta_i + \beta_j}{\beta_i - \beta_j} = -1 < a. \tag{2.15}$$

This shows that part (1) in Levy's property may not hold in $\mathcal{D}_{\text{SSD}}^{c}$. Additionally, we find that $Y_i \succ_1 Y_j$. Applying Li and Wong [20, Theorem 7], we have $E[U(Y_i)] > E[U(Y_j)]$ for any nondecreasing U and thus, there exists a dominance in EU for all nondecreasing U. However, as

$$\frac{\alpha_i - \alpha_j}{\beta_j - \beta_i} = -1 < b, \tag{2.16}$$

thus inequality in (2.13) does not hold, implying that part (2b) in Levy's property may not hold.

We now give another example in which (2.12) holds but $(Y_i, Y_j) \notin \mathcal{D}_{\text{MV}}$ as shown in the following.

Example 2.8. Let X be the seed random variable with support $[a,b] = [0,1]$, let $Y_i = \alpha_i + \beta_i X$, and let $Y_j = \alpha_j + \beta_j X$, and set $\beta_i > \beta_j > 0$ and $\alpha_j = \alpha_i + \beta_i - \beta_j$.

In this example, since $\beta_i > \beta_j > 0$ and $\alpha_j > \alpha_i$, we have $(Y_i, Y_j) \notin \mathcal{D}_{\text{MV}}$. However,

$$\frac{\alpha_j - \alpha_i}{\beta_i - \beta_j} = \frac{\alpha_i + \beta_i - \beta_j - \alpha_i}{\beta_i - \beta_j} = 1 > a \tag{2.17}$$

and hence (2.12) holds. This leads to our conclusion that part (1) of Levy's property does not hold in this example. However, in this example, we find that

$$\frac{\alpha_i - \alpha_j}{\beta_j - \beta_i} = 1 \geq b \tag{2.18}$$

and hence (2.13) holds and it is easy to show that $Y_i \succ_1 Y_j$. In this connection, part (2b) of Levy's property is valid in this example. Another trivial example in which part (2b) does not hold is the following.

Example 2.9. We set $\alpha_i > \alpha_j$ and $\beta_i = \beta_j$.

In this example, $Y_i \succ_1 Y_j$ and hence there exists a dominance in EU for all nondecreasing U but (2.13) does not hold.

3. Concluding remarks

Meyer [25] contributed to the theory of mean-variance criterion by extending the theory to include the comparison among distributions that differ only by location and scale parameters as well as to include the general utility functions with only convexity or concavity restrictions. Levy [16] extended Meyer's results by introducing some relationships between the stochastic-dominance and the mean-variance efficient sets. However, Meyer [26] commented that Levy's findings is an application of the principle that segments of efficient sets cannot have slopes which are greater (smaller) than the highest (least) sloped indifference curve and commented that those portions of the MV-efficient set which are either too flat or too steeply sloped are not EU efficient.

We first make some comments on Meyer's paper and extend the results from Tobin [37] that the indifference curve is convex upward for risk averters, concave downwards for risk lovers, and horizontal for risk neutral investors to include the general conditions as stated in Meyer [25]. We then comment on Levy's findings and show that the relationships in Levy's property do not hold in certain situations. We further explore the relationships among the first- and second-degree stochastic dominance efficient sets and the mean-variance efficient set to show that they are not equal to one another. We check the literature on the subject and conclude that the results in our paper are still new and we hope that our results would be able to contribute to the existing literature.

Further extensions of the theory developed in this paper, future work could extend our efforts to link stochastic dominance to mean-variance criterion developed by Markowitz [21], Tobin [37], and Sharpe [33] for location-scale family. As the theory developed by Meyer and Levy, and in this paper mainly concerns only risk averters, it would also be worthwhile to extend it to risk lovers (see, e.g., Hammond [11], Meyer [24], Hershey and Schoemaker [13], Stoyan [35], Myagkov and Plott [27], Wong and Li [44], Post [28], Anderson [1], and Post and Levy [30]) and to investors with S-shaped or reverse S-shaped utility functions (see, e.g., Kahneman and Tversky [14], Tversky and Kahneman [38], Levy and Wiener [19], and Levy and Levy [17, 18]). Another area of extension is to extend our theory to a variable of loss (see, e.g., Weeks and Wingler [41], Weeks [40], Post and Diltz [29], and Dillinger et al. [5]). In addition, the theory developed in this paper could be applied to many different areas in business, economics, and finance. For example, one could easily incorporate our approach to explain well-known financial anomalies (see, e.g., McNamara [23], Wong and Bian [42], Post [28], Post and Levy [30], Kuosmanen [15], and Fong et al. [9]) and to model investment risk (see, e.g., Matsumura et al. [22], Doumpos et al. [6], Wong and Chan [43], Fong and Wong [8], and Broll et al. [3]).

Acknowledgments

The author is grateful to professor Mahyar Amouzegar and anonymous referees for their substantive comments that have significantly improved this manuscript. My deepest thanks are given to Thomas Kwok Keung Au, Bin Cheng, and Song Yan for their helpful assistance and comments. The author would also like to thank Robert B. Miller and Howard E. Thompson for their continuous guidance and encouragement.

References

[1] G. Anderson, *Toward an empirical analysis of polarization*, Journal of Econometrics **122** (2004), no. 1, 1–26.

[2] D. P. Baron, *On the utility theoretic foundations of mean-variance analysis*, Journal of Finance **32** (1977), no. 5, 1683–1697.

[3] U. Broll, J. E. Wahl, and W.-K. Wong, *Elasticity of risk aversion and international trade*, Economics Letters **91** (2006), no. 1, 126–130.

[4] K. V. Chow, *Marginal conditional stochastic dominance, statistical inference, and measuring portfolio performance*, Journal of Financial Research **24** (2001), no. 2, 289–307.

[5] A. M. Dillinger, W. E. Stein, and P. J. Mizzi, *Risk averse decisions in business planning*, Decision Sciences **23** (1992), no. 4, 1003–1008.

[6] M. Doumpos, S. Zanakis, and C. Zopounidis, *Multicriteria preference disaggregation for classification problems with an application to global investing risk*, Decision Sciences **32** (2001), no. 2, 333–385.

[7] M. S. Feldstein, *Mean-variance analysis in the theory of liquidity preference and portfolio selection*, Review of Economics Studies **36** (1969), no. 1, 5–12.

[8] W. M. Fong and W.-K. Wong, *The modified mixture of distributions model: a revisit*, Annals of Finance **2** (2006), no. 2, 167–178.

[9] W. M. Fong, W.-K. Wong, and H. H. Lean, *International momentum strategies: a stochastic dominance approach*, Journal of Financial Markets **8** (2005), no. 1, 89–109.

[10] J. Hadar and W. R. Russell, *Stochastic dominance and diversification*, Journal of Economic Theory **3** (1971), no. 3, 288–305.

[11] J. S. Hammond, *Simplifying the choice between uncertain prospects where preference is nonlinear*, Management Science **20** (1974), no. 7, 1047–1072.

[12] G. Hanoch and H. Levy, *Efficiency analysis of choices involving risk*, Review of Economic Studies **36** (1969), no. 3, 335–346.

[13] J. C. Hershey and P. J. H. Schoemaker, *Risk taking and problem context in the domain of losses: an expected utility analysis*, Journal of Risk and Insurance **47** (1980), no. 1, 111–132.

[14] D. Kahneman and A. Tversky, *Prospect theory: an analysis of decision under risk*, Econometrica **47** (1979), no. 2, 263–291.

[15] T. Kuosmanen, *Efficient diversification according to stochastic dominance criteria*, Management Science **50** (2004), no. 10, 1390–1406.

[16] H. Levy, *Two-moment decision models and expected utility maximization: comment*, American Economic Review **79** (1989), no. 3, 597–600.

[17] M. Levy and H. Levy, *Prospect theory: much ado about nothing?*, Management Science **48** (2002), no. 10, 1334–1349.

[18] H. Levy and M. Levy, *Prospect theory and mean-variance analysis*, Review of Financial Studies **17** (2004), no. 4, 1015–1041.

[19] H. Levy and Z. Wiener, *Stochastic dominance and prospect dominance with subjective weighting functions*, Journal of Risk and Uncertainty **16** (1998), no. 2, 147–163.

[20] C.-K. Li and W.-K. Wong, *Extension of stochastic dominance theory to random variables*, RO Recherche Opérationnelle **33** (1999), no. 4, 509–524.

[21] H. M. Markowitz, *Portfolio selection*, Journal of Finance **7** (1952), no. 1, 77–91.

[22] E. M. Matsumura, K. W. Tsui, and W.-K. Wong, *An extended multinomial-Dirichlet model for error bounds for dollar-unit sampling*, Contemporary Accounting Research **6** (1990), no. 2, 485–500.

[23] J. R. McNamara, *Portfolio selection using stochastic dominance criteria*, Decision Sciences **29** (1998), no. 4, 785–801.

[24] J. Meyer, *Second degree stochastic dominance with respect to a function*, International Economic Review **18** (1977), no. 2, 477–487.

[25] ———, *Two-moment decision models and expected utility maximization*, American Economic Review **77** (1987), no. 3, 421–430.
[26] ———, *Two-moment decision models and expected utility maximization: reply*, American Economic Review **79** (1989), no. 3, 603.
[27] M. Myagkov and C. R. Plott, *Exchange economies and loss exposure: experiments exploring prospect theory and competitive equilibria in market environments*, American Economic Review **87** (1997), no. 5, 801–828.
[28] T. Post, *Empirical tests for stochastic dominance efficiency*, The Journal of Finance **58** (2003), no. 5, 1905–1931.
[29] G. V. Post and J. D. A. Diltz, *A stochastic dominance approach to risk analysis of computer systems*, MIS Quarterly **10** (1986), no. 4, 362–375.
[30] T. Post and H. Levy, *Does risk loving drive asset prices? a stochastic dominance analysis of aggregate investor preferences and beliefs*, Review of Financial Studies **18** (2005), no. 3, 925–953.
[31] M. Rothschild and J. E. Stiglitz, *Increasing risk. I. A definition*, Journal of Economic Theory **2** (1970), no. 3, 225–243.
[32] ———, *Increasing risk. II. Its economic consequences*, Journal of Economic Theory **3** (1971), no. 1, 66–84.
[33] W. F. Sharpe, *A simplified model for portfolio analysis*, Management Science **9** (1963), no. 2, 277–293.
[34] H.-W. Sinn, *Economic Decisions under Uncertainty*, Studies in Mathematical and Managerial Economics, vol. 32, North-Holland, Amsterdam, 1983.
[35] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Chichester, 1983.
[36] L. Tesfatsion, *Stochastic dominance and the maximization of expected utility*, Review of Economic Studies **43** (1976), no. 2, 301–315.
[37] J. Tobin, *Liquidity preference as behavior towards risk*, Review of Economics Studies **25** (1958), no. 2, 65–86.
[38] A. Tversky and D. Kahneman, *Advances in prospect theory: cumulative representation of uncertainty*, Journal of Risk and Uncertainty **5** (1992), no. 4, 297–323.
[39] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, New Jersey, 1944.
[40] J. K. Weeks, *Stochastic dominance: a methodological approach to enhancing the conceptual foundations of operations management theory*, Academy of Management Review **10** (1985), no. 1, 31–38.
[41] J. K. Weeks and T. R. Wingler, *A stochastic dominance ordering of scheduling rules*, Decision Sciences **10** (1979), no. 2, 245–257.
[42] W.-K. Wong and G. Bian, *Robust estimation in capital asset pricing model*, Journal of Applied Mathematics and Decision Sciences **4** (2000), no. 1, 65–82.
[43] W.-K. Wong and R. H. Chan, *On the estimation of cost of capital and its reliability*, Quantitative Finance **4** (2004), no. 3, 365–372.
[44] W.-K. Wong and C.-K. Li, *A note on convex stochastic dominance*, Economics Letters **62** (1999), no. 3, 293–300.
[45] Y. Zhao and W. Ziemba, *A dynamic asset allocation model with downside risk control*, Journal of Risk **3** (2000), no. 1, 91–113.

Wing-Keung Wong: Risk Management Institute and Department of Economics, National University of Singapore, 1 Arts Link, Singapore 117570
E-mail address: ecswwk@nus.edu.sg

COMPARISON OF TWO COMMON ESTIMATORS OF THE RATIO OF THE MEANS OF INDEPENDENT NORMAL VARIABLES IN AGRICULTURAL RESEARCH

C. G. QIAO, G. R. WOOD, C. D. LAI, AND D. W. LUO

Received 26 October 2005; Revised 24 May 2006; Accepted 5 June 2006

This paper addresses the problem of estimating the ratio of the means of independent normal variables in agricultural research. The first part of the research examines the distributional properties of the ratio of independent normal variables, both theoretically and using simulation. The second part of the research evaluates the relative merits of two common estimators of the ratio of the means of independent normal variables in agricultural research, an arithmetic average and a weighted average, via simulation experiments using normal distributions. The results are then tested using research data from rice breeding multi-environment trials in Jilin Province, China, in 1994. These data are used to demonstrate the diagnostic approach developed for assessing the "safe" use of the arithmetic and the weighted average methods for estimating the ratio of the means of independent normal variables.

1. Introduction

A ratio $R = X/Y$ of independent normal variables is commonly used to capture the relative merits of two contrasting treatments, practices or methodologies in agricultural research. Examples include the ratio of grain yield of a new crop variety to that of the commercial control variety across a range of environments, harvest index (the ratio between economical and biological yields of plants), and relative efficiency (the ratio between error estimates of two biological models in agricultural research). It is important to know how the mean $E(X/Y)$ of the ratio of the two independent normal variables should be estimated when several such ratio estimates are available. Throughout, we assume that X and Y are uncorrelated and that $\mu_Y > 0$.

The motivation for this research lies in the study of relative performance of rice varieties in grain yield in Jilin Province of China in 1994 (see Jilin Provincial Seed Station [4]), where a series of ratio estimates needed to be pooled or averaged over different environments. For this rice breeding multi-environment trial (MET) conducted over eight

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 78375, Pages 1–14
DOI 10.1155/JAMDS/2006/78375

Table 1.1. Grain yields (kg/ha) of three rice varieties (850011, Chang 90–40, and Yan 501) and the percent grain yield of each of these test varieties relative to the control variety (Jiyin 12) by the arithmetic average ($\overline{R}_A$) and weighted average ($\overline{R}_W$) in a multi-environment trial in 1994.

Location	850011 (X)	Control (Y)	Chang 90–40 (X)	Control (Y)	Yan 501 (X)	Control (Y)
Changchun	7353	5498	6304	5498	6753	5498
Gongzhuling	7574	6063	7917	6063	6485	6063
Jilin	9285	8820	8475	8820	5745	8820
Shuangyang	5646	7857	6504	7857	—	—
Lishu	8940	8945	9250	8945	8625	8945
Shulan	8554	8049	9005	8049	9254	8049
Tonghua	6278	5002	—	—	6627	5002
Yanbian	7889	7820	7545	7820	7956	7820
Correlation (X, Y)	0.601 ($p = 0.115$)		0.648 ($p = 0.116$)		0.418 ($p = 0.351$)	
Mean	7690	7257	7857	7579	7349	7171
Standard deviation	1264	1521	1154	1316	1279	1622
Coefficient of variation	0.164	0.210	0.147	0.174	0.174	0.226
$\overline{R}_A$	108.6	—	105.1	—	105.8	—
$\overline{R}_W$	106.0	$CV_{\overline{Y}} = 0.074$	103.7	$CV_{\overline{Y}} = 0.066$	102.5	$CV_{\overline{Y}} = 0.086$

locations, the grain yield data were analysed to quantify the percent increase in grain yield of three varieties over the control variety (Table 1.1). In such studies, a subset of rice varieties are added in or dropped out from the regional variety testing program every year, based on their overall performance (mainly yield) relative to the control. This makes the field evaluation of rice varieties progress in a roll-over pattern. The aim is to estimate the mean percent yield increase of each of the test varieties over the control variety across a range of environments. In Table 1.1, the percent grain yield of each test variety relative to the control variety (Jiyin 12) is used to assess the yield improvement of the new variety at these locations. The ratio of grain yield of each test variety to grain yield of the control (expressed as a percentage), over all possible trials in the MET, is to be estimated.

Since the mean $E(X/Y)$ of the ratio of two independent normal variables does not exist (Lukacs and Laha [10]; Lukacs [9]; Springer [16]; Johnson et al. [5]), this causes a practical problem in its estimation due to the non-existence of $E(1/Y)$, because Y can in theory assume values arbitrarily close to zero. Lai et al. [8] studied a punctured normal distribution, where a small neighbourhood ($|Y| \le \varepsilon$ with ε a small positive number) is removed from consideration, through two left-truncated normal variables. They show that the mean of the inverse of the punctured normal variable exists, whence $E(X/Y \mid |Y| > \varepsilon) = E(X)E(1/Y \mid |Y| > \varepsilon)$ also exists although $E(X/Y)$ fails to exist. They also justify the estimation of μ_X/μ_Y as a surrogate for $E(X/Y)$, because μ_X/μ_Y is a satisfactory measure of centre for X/Y. Hence, as the maximum likelihood estimator of μ_X/μ_Y, $\overline{X}/\overline{Y}$ is naturally

the best estimator of μ_X/μ_Y. The aim of this paper is to explore theoretical and numerical aspects of the estimation of this ratio, leading to the provision of useful advice for the practitioner.

Two methods are widely used for averaging different ratio estimates in agricultural research. The first is the arithmetic average approach, which divides the sum of all the ratio estimates by the total number of estimates (Kaeppler et al. [6]; Moreau et al. [11]; Qiao et al. [13]). The second is known as the weighted average approach, which estimates the true ratio via dividing the sum of all the numerators by the sum of all the denominators of the individual ratio estimates (Robinson et al. [15]; Haque et al. [2]; Witcombe et al. [17]). When used on the same set of data to estimate the mean of the ratio of two independent normal variables, these two approaches may give different results or even reach contradictory conclusions in some circumstances. We have not, however, found any report in the literature comparing these two methods. We note that related research was conducted in Qiao et al. [14], where the corresponding estimators of a binomial proportion using several independent samples in agricultural research were investigated. That work provided the impetus for the current study.

We pause now to describe the two estimators. Suppose a sample of observations (X_i, Y_i), $i = 1,2,\ldots,n$, is taken from a bivariate normal population $N(\mu_X,\mu_Y,\sigma_X,\sigma_Y,\rho)$ and for each observation, the ratio X_i/Y_i is calculated. There are two popular ways in agricultural research to estimate the ratio μ_X/μ_Y, the arithmetic average approach, with $\overline{R}_A = (\sum X_i/Y_i)/n$, and the weighted average approach, with

$$\overline{R}_W = \sum W_i \frac{X_i}{Y_i} = \left[\left(\frac{Y_1}{\sum Y_i}\right)\left(\frac{X_1}{Y_1}\right) + \left(\frac{Y_2}{\sum Y_i}\right)\left(\frac{X_2}{Y_2}\right) + \cdots + \left(\frac{Y_n}{\sum Y_i}\right)\left(\frac{X_n}{Y_n}\right)\right] = \frac{\sum X_i}{\sum Y_i} = \frac{\overline{X}_n}{\overline{Y}_n}. \tag{1.1}$$

Intuition suggests that $\overline{R}_A = (\sum X_i/Y_i)/n$ is a poor estimator of μ_X/μ_Y. This is because Y_i can be small and positive, leading to large and positive X_i/Y_i, thus biasing the final average upwards. It averages after division. In contrast, $\overline{R}_W = \overline{X}_n/\overline{Y}_n$ should be a better estimator of μ_X/μ_Y as very small $\overline{Y}_n$ values are less likely to occur, thus lessening the upward bias. It averages before division. Hence, $\overline{R}_W$ appears generally superior to $\overline{R}_A$.

For the motivation example, a ratio of means of independent normal variables (grain yield in this instance) is to be estimated. The arithmetic and weighted average ratio estimators produced different estimates in Table 1.1 and it is unclear which estimator should be used. This forms the drive for investigations of the theoretical foundation of the difference between the two methods and for evaluation of them in a more general sense in agricultural research.

The paper is presented in five sections. Section 2 explores the distribution of the ratio of two independent normal variables; this is followed by an evaluation of the two estimators of the ratio of normal means, both theoretically and using simulation. Section 4 applies the findings to a data set from an agricultural experiment, while Section 5 contains general recommendations concerning the use of the two estimators in agricultural research.

2. Distribution of the ratio of independent normal variables

2.1. The probability density function of the ratio of independent normal variables. Springer (see [16, pages 139–148]) found the probability density function of $W = (X/\sigma_X)/(Y/\sigma_Y)$ and then $R = X/Y$ through the use of the simple transformation $R = (\sigma_X/\sigma_Y)W$. This result is rather unwieldy for computational purposes. Kamerud [7] gave the probability density function of $R = X/Y$ explicitly. There is an error in her derivation of the density function of W that we rectify in the following, making it necessary to adjust the density function.

Define $U = X/\sigma_X$, $V = Y/\sigma_Y$, and thus $U \sim N(\mu_X/\sigma_X, 1)$, $V \sim N(\mu_Y/\sigma_Y, 1)$. Set $W = U/V$ and let g be its density function. Replacing μ_1 and μ_2 in Kamerud [7] by μ_X/σ_X and μ_Y/σ_Y, respectively, we have

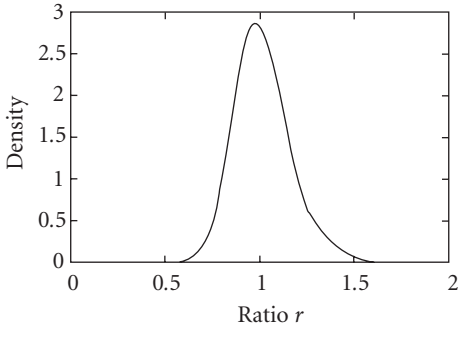
$$g(w) = (2\pi)^{-1} Q \exp(M), \tag{2.1}$$

where $M = -(1/2)((\mu_Y/\sigma_Y)w - \mu_X/\sigma_X)^2 s^2$, $Q = ks(2\pi)^{1/2}[1 - 2\Phi(-k/s)] + 2s^2 \exp(-k^2/2s^2)$, $s = (w^2+1)^{-1/2}$, $k = ((\mu_X/\sigma_X)w + \mu_Y/\sigma_Y)s^2$, and Φ is the standard normal cumulative distribution function. The probability density of R is then given by $f(r) = (\sigma_Y/\sigma_X)g((\sigma_Y/\sigma_X)r)$.

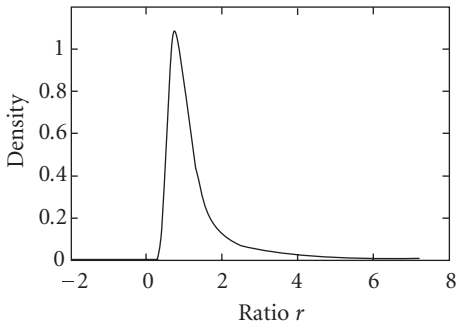
In contrast to the method given in Springer [16], Kamerud's expressions are easy to compute numerically. Hence, Kamerud's probability density function is used to generate graphs of X/Y against its density, shown in Figure 2.1, to assess the distributional properties of the ratio of two independent normal variables. Some typical plots (Figures 2.1(a)–2.1(c)) are drawn using this density function, with varying coefficient of variation (CV) for the denominator variable. From the considerations of Section 2 and Qiao et al. [14], it is evident that the CV of the denominator is of critical importance. In Figure 2.1(a), the CV of both X and Y is small (0.1). Hence, the density function is fairly symmetric around $\mu_X/\mu_Y = 1$, having the bell-shape of a normal distribution. The long tail in Figure 2.1(b) and multiple peaks in Figure 2.1(c), where the CV is small for the numerator but large for the denominator, indicate that the moments, especially the mean of the distribution, may not exist.

For small coefficient of variation of Y (CV_Y), the moments of the ratio appear to exist. This is due to the fact that very small Y values were not sampled in the above graphical presentation and hence we were effectively sampling from $(X/Y) \mid |Y| > \varepsilon$, a punctured normal for the denominator variable (Lai et al. [8]). The moments of X/Y appear to exist in this situation. Both the arithmetic and the weighted average methods involve ratios of independent normal variables. We will demonstrate later that, as far as estimation of μ_X/μ_Y is concerned, both the arithmetic and the weighted average methods can be used when CV_Y is sufficiently small. The circumstances under which the ratio of two independent normal values can be used to safely estimate the ratio of the means are now investigated using simulation.

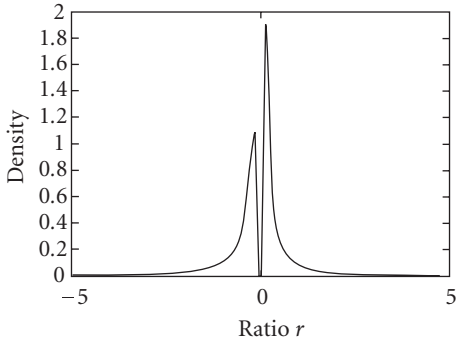
2.2. Simulation of the distribution of the ratio of normal random variables. Software SAS 8.2 was used to simulate the distributional properties of the ratio $R = X/Y$ of two



(a) $\sigma_X = \sigma_Y = 10$, $\mathrm{CV}_X = \mathrm{CV}_Y = 0.1$



(b) $\sigma_X = 10$, $\sigma_Y = 50$, $\mathrm{CV}_X = 0.1$, $\mathrm{CV}_Y = 0.5$



(c) $\sigma_X = 10$, $\sigma_Y = 500$, $\mathrm{CV}_X = 0.1$, $\mathrm{CV}_Y = 5.0$

Figure 2.1. Density functions for the ratio R of two independent normal variables $X \sim N(100, \sigma_X)$ and $Y \sim N(100, \sigma_Y)$, so $\mu_X/\mu_Y = 1$, as CV_Y varies.

normal variables $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. The population means of both variables were fixed at 100 and hence $\mu_X/\mu_Y = 1$. The population standard deviations of both variables took the values 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, and 500, leading to both CV_X and CV_Y taking values 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 2, and 5. For each of these 144 combinations, 500 000 pairs of (X_i, Y_i) were sampled; the mean and standard deviation of the ratios $R_i = X_i/Y_i$ were examined.

Before considering the simulation results, we offer some theoretical reflections. The mean of X/Y does not exist, but under sampling, X/Y and $X/Y \mid |Y| > \varepsilon$ are essentially the same random variable for sufficiently small ε. For example, if $\mu_Y > 0$ and $\mathrm{CV}_Y < 0.2$, it is possible to find an ε such that $0 < \varepsilon < \mu_Y - 5\sigma_Y$. Hence, fewer than one in a million sample values of Y will have absolute value less than ε. As argued in the introduction, $E(X/Y \mid |Y| > \varepsilon) = E(X)E(1/Y \mid |Y| > \varepsilon)$ and this is approximately μ_X/μ_Y (Lai et al. [8, Section 4.2]) as long as $\mathrm{CV}_Y < 0.2$. As pointed out in (Lai et al. [8, Section 5.2]), $\overline{X}/\overline{Y}$ is the maximum likelihood estimator of μ_X/μ_Y, and hence the estimator of choice. In summary,

as long as $CV_Y < 0.2$, theory tells us that $\overline{X}/\overline{Y}$ is a sound estimator for the centre of X/Y. Our simulation results now confirm these findings.

The simulation results, listed in Table 2.1, indicate that the sample mean and standard deviation of the ratio estimates are all strongly influenced by CV_Y. This supports our earlier remark that CV_Y, not CV_X, is a critical parameter. When $CV_Y < 0.2$, the mean of R remains close to $\mu_X/\mu_Y = 1$, while the standard deviation of R increases approximately linearly as CV_X increases (Table 2.1). It appears that the variation of R is almost purely determined by the variation in the numerator variable when CV_Y is small. Evidently, $CV_Y = 0.2$ is an appropriate cut-off point for the denominator; for larger CV_Y values, the mean deviates substantially from $\mu_X/\mu_Y = 1$ and the standard deviation increases accordingly. In contrast, CV_X has no influence on the mean of R, and a relatively small influence on the standard deviation. Hence, the deleterious effect of increasing CV_Y is much stronger than that when increasing CV_X.

The sample mean of the ratios fails to estimate μ_X/μ_Y when $CV_Y > 0.4$, while the standard deviation is extremely large, with erratic behaviour, when $CV_Y > 0.3$. For the sample means to serve as reasonable estimators of μ_X/μ_Y for this sample size (500 000), CV_Y apparently has to be kept sufficiently small ($CV_Y < 0.2$ appears to suffice). In practical applied research, it is rare for the CV of a normal variable to be larger than 5.0. Thus, as long as $CV_Y < 0.2$, it makes empirical sense use the ratio estimator X/Y.

This simulation was repeated first with $\mu_X/\mu_Y = 10/100 = 0.1$, and then with $\mu_X/\mu_Y = 100/10 = 10$. The mean and standard deviation of the ratio behave similarly to the case where $\mu_X/\mu_Y = 1$. This provides circumstantial evidence that the magnitude of μ_X/μ_Y does not influence the manner in which the sample mean estimates μ_X/μ_Y.

2.3. Implications in applied research. The non-existence of moments of the ratio of normal variables presents a problem. In practical applications, as long as we avoid sampling in an interval around $Y = 0$, moments of X/Y will appear to exist. If we let ε be a sufficiently small positive quantity, then X_i/Y_i can be used to estimate the ratio of μ_X/μ_Y, provided $|Y_i| > \varepsilon$. Hall [1] showed that if a positive random variable Y has a normal distribution singly truncated from below, denoted by $N_a(\mu,\sigma)$, where $0 < a < Y$, then the inverse moments $E(Y^{-1})$ and $E(Y^{-2})$ can be approximated accurately by expressions involving Dawson's integral. The expressions are independent of the truncation point a, provided that $(\sigma/\mu)^2 < a/\mu < 1/25$. This will ensure the apparent existence of the expectation of the ratio of two independent normal variables $E(X/Y)$ when $(\sigma/\mu)^2 < 1/25$, or $CV_Y = \sigma/\mu < 1/5 = 0.2$. The central idea behind this and behind our investigations is similar, namely to make the denominator variable nonzero, a condition easily met in practical research.

The findings also suggest that if we want to use the sample mean of ratios X_i/Y_i to estimate μ_X/μ_Y, then the larger the sample we use, the smaller the CV_Y we will need to avoid sample points getting close to zero in the denominator. When CV_Y is sufficiently small, there is almost no chance for a value of Y very close to zero being sampled, thus ensuring the apparent existence of sample moments.

When CV_Y is very small, Y behaves as $Y \mid |Y| \geq \varepsilon$ for some $\varepsilon > 0$, thus the moments of $1/Y$ can be accurately approximated (Hall [1]; Nahmias and Wang [12]). This leads to

Table 2.1. Simulation of the ratio distribution: mean and standard deviation for 500 000 pairs of observations X_i/Y_i, where $X_i \sim N(\mu_X, \sigma_X)$ and $Y_i \sim N(\mu_Y, \sigma_Y)$, under varying coefficients of variation (CV), with $\mu_X/\mu_Y = 100/100 = 1$.

CV_X	CV_Y											
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	2.0	5.0
	Mean											
0.1	1.010	1.046	1.149	1.168	1.738	1.077	3.653	1.199	0.466	−1.810	−0.144	−0.054
0.2	1.011	1.046	1.150	1.167	1.771	0.967	3.485	1.123	0.271	−1.910	−0.143	−0.031
0.3	1.011	1.047	1.152	1.166	1.803	0.857	3.318	1.046	0.076	−2.009	−0.142	−0.007
0.4	1.011	1.047	1.153	1.164	1.835	0.747	3.150	0.970	−0.120	−2.108	−0.140	0.017
0.5	1.011	1.047	1.155	1.163	1.868	0.637	2.982	0.893	−0.315	−2.207	−0.139	0.041
0.6	1.011	1.047	1.156	1.162	1.900	0.527	2.815	0.817	−0.510	−2.307	−0.138	0.065
0.7	1.011	1.047	1.158	1.161	1.932	0.417	2.647	0.740	−0.706	−2.406	−0.137	0.089
0.8	1.011	1.047	1.160	1.159	1.965	0.307	2.479	0.664	−0.901	−2.505	−0.136	0.113
0.9	1.012	1.047	1.161	1.158	1.997	0.197	2.312	0.587	−1.096	−2.605	−0.134	0.137
1.0	1.012	1.047	1.163	1.157	2.029	0.088	2.144	0.511	−1.292	−2.704	−0.133	0.160
2.0	1.013	1.049	1.178	1.145	2.353	−1.012	0.467	−0.255	−3.245	−3.697	−0.121	0.399
5.0	1.017	1.053	1.107	1.225	3.323	−4.311	−4.563	−2.550	−9.105	−6.676	−0.085	1.116
	Standard deviation											
0.1	0.1	0.3	12.4	49.1	548.6	433.7	1606.9	255.3	1020.0	2145.6	552.5	93.4
0.2	0.2	0.3	12.6	50.7	575.8	439.2	1552.3	244.4	1143.8	2175.4	545.2	84.1
0.3	0.3	0.4	12.9	52.3	604.0	459.6	1498.9	240.1	1269.0	2206.5	538.3	79.9
0.4	0.4	0.5	13.4	54.1	633.1	493.2	1446.8	242.7	1395.1	2238.9	531.9	81.5
0.5	0.5	0.6	13.9	55.9	663.0	537.5	1396.2	252.0	1521.9	2272.6	526.0	88.8
0.6	0.6	0.7	14.4	57.8	693.5	590.1	1347.3	267.3	1649.3	2307.5	520.6	100.4
0.7	0.7	0.8	15.1	59.8	724.6	648.9	1300.1	287.6	1777.1	2343.6	515.7	115.1
0.8	0.8	0.9	15.8	61.8	756.2	712.5	1254.9	311.9	1905.2	2380.8	511.3	131.8
0.9	0.9	1.0	16.5	63.9	788.3	779.6	1212.0	339.5	2033.6	2419.0	507.5	149.8
1.0	1.0	1.1	17.3	66.1	820.8	849.4	1171.6	369.5	2162.3	2458.3	504.2	168.8
2.0	2.0	2.2	26.4	89.3	1160.1	1615.1	965.4	730.3	3455.8	2896.7	503.7	376.8
5.0	5.1	5.4	58.4	167.2	2234.0	4055.0	2305.0	1931.0	7357.4	4490.4	779.2	1029.9

apparent existence of the sample moments of X/Y. From our simulations and the results of Hall [1], $CV_Y < 0.2$ can be used as a condition which determines the usefulness of $\overline{R}_A$ and $\overline{R}_W$.

3. Comparison of the two estimators

In this section, we examine the performance of the two estimators of μ_X/μ_Y, first in the light of the conclusion of Section 2, then theoretically, and finally using simulation.

3.1. Estimators and coefficient of variation. From the previous section, it is evident that X_i/Y_i is a reasonable estimate of μ_X/μ_Y provided $CV_Y < 0.2$. This observation will provide

the reason why $\overline{R}_W$ improves as the sample size n increases, while for $\overline{R}_A$, this is not the case; hence, $\overline{R}_W$ will be regarded as a superior estimator. We now examine $\overline{R}_W$ and $\overline{R}_A$ separately and conclude that $\overline{R}_W$ can be used if $\mathrm{CV}_{\overline{Y}_n} < 0.2$, while $\overline{R}_A$ can be adopted if $\mathrm{CV}_Y < 0.2$. The error in $\overline{R}_A$ as an estimator of μ_X/μ_Y does not decrease with sample size, whereas the error in $\overline{R}_W$ as an estimator of μ_X/μ_Y decreases to zero with sample size, hence $\overline{R}_W$ is to be favoured.

3.2. Weighted average ratio estimator. Recall that $\overline{R}_W$ is called the weighted average estimator, named so because it can be written as

$$\overline{R}_W = \frac{\overline{X}_n}{\overline{Y}_n} = \left(\frac{Y_1}{\sum Y_i}\right)\left(\frac{X_1}{Y_1}\right) + \left(\frac{Y_2}{\sum Y_i}\right)\left(\frac{X_2}{Y_2}\right) + \cdots + \left(\frac{Y_n}{\sum Y_i}\right)\left(\frac{X_n}{Y_n}\right). \tag{3.1}$$

Since $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$, it follows that $\overline{X}_n \sim N(\mu_X, \sigma_X/\sqrt{n})$ and $\overline{Y}_n \sim N(\mu_Y, \sigma_Y/\sqrt{n})$. Hence, $\mathrm{CV}_{\overline{X}_n} = (\sigma_X/\sqrt{n})/\mu_X$ and $\mathrm{CV}_{\overline{Y}_n} = (\sigma_Y/\sqrt{n})/\mu_Y$. If $\mu_X, \mu_Y \neq 0$ and $(\sigma_Y/\sqrt{n})/\mu_Y < 0.2$, then our simulations demonstrate that $\overline{R}_W = \overline{X}_n/\overline{Y}_n$ is an acceptable estimator of μ_X/μ_Y. Thus, for practical purposes, we recommend that $\overline{R}_W$ is used to estimate μ_X/μ_Y, since taking a sample of sufficiently large size n will reduce the coefficient of variation of $\overline{Y}_n$.

In designing a research experiment or survey, the sample size n required to provide a reasonably good estimate of μ_X/μ_Y can be determined in the following way. Take a sample of size n from $X \sim N(\mu_X, \sigma_X)$ and $Y \sim N(\mu_Y, \sigma_Y)$. In order for $\overline{R}_W = \overline{X}_n/\overline{Y}_n$ to estimate μ_X/μ_Y, the coefficient of variation for the denominator of $\overline{R}_W$ has to satisfy $\mathrm{CV}_{\overline{Y}} = \sigma_Y/\sqrt{n}/\mu_Y < 0.2$, or $n > 25(\sigma_Y^2/\mu_Y^2)$. Here, $\mathrm{CV}_{\overline{Y}}$, rather than CV_Y, being small is the condition that needs to be fulfilled. In practical situations, the population means and standard deviations of interest are rarely known, but can be estimated by the relevant sample means and standard deviations. Hence, the above inequality can be approximated by $n > 25(s_Y^2/\overline{Y}_n^2)$.

In practical terms, the sample size n is always predetermined. Thus, sample results can be examined to see if they satisfy the requirement $s_Y/\sqrt{n}/\overline{Y}_n < 0.2$. This will provide a general guideline for evaluating the suitability of the weighted average method in estimating the ratio of the means of two normal variables.

3.3. Arithmetic average ratio estimator. Estimator $\overline{R}_A = \sum_{i=1}^{n}(X_i/Y_i)/n$ is an equally weighted average of n ratios X_i/Y_i. We can adopt the same methodology used in evaluating the weighted average method to assess the suitability of $\overline{R}_A$. The coefficient of variation of Y_i, however, is σ_Y/μ_Y in this case, instead of $\sigma_Y/\sqrt{n}/\mu_Y$. If $\sigma_Y/\mu_Y \geq 0.2$, for example, then X_i/Y_i is a poor estimator of μ_X/μ_Y. Taking a larger sample size n is of little use. Naturally, the sample value of $s_Y/\overline{Y}_n$ can be used as a diagnostic tool for the evaluation of the appropriateness of $\overline{R}_A$. Thus, we recommend the use of $\overline{R}_A$ only if the coefficient of variation of Y_i is sufficiently small, that is, $\mathrm{CV}_Y = s_Y/\overline{Y}_n < 0.2$. The simulation results, which follow, support our recommendation.

3.4. Theoretical considerations. Here we prove that $\overline{R}_W$ does converge to μ_X/μ_Y in probability, as the sample size increases. Recall that $X_n \xrightarrow{P} X$, convergence in probability, if for every $\varepsilon > 0$, $P(|X_n - X| \geq \varepsilon) \to 0$ as $n \to \infty$. We now present the following relevant results.

Lemma 3.1 (Lukacs [9, Corollary to Theorem 2.3.3]). *Let $g(x, y)$ be a continuous function of the real variables x and y. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $g(X_n, Y_n) \xrightarrow{P} g(X, Y)$ as $n \to \infty$.*

Theorem 3.2. *Let $\overline{X}_n$ and $\overline{Y}_n$ be means of samples of size n, drawn independently from normal populations. Then $\overline{X}_n/\overline{Y}_n \xrightarrow{P} \mu_X/\mu_Y$.*

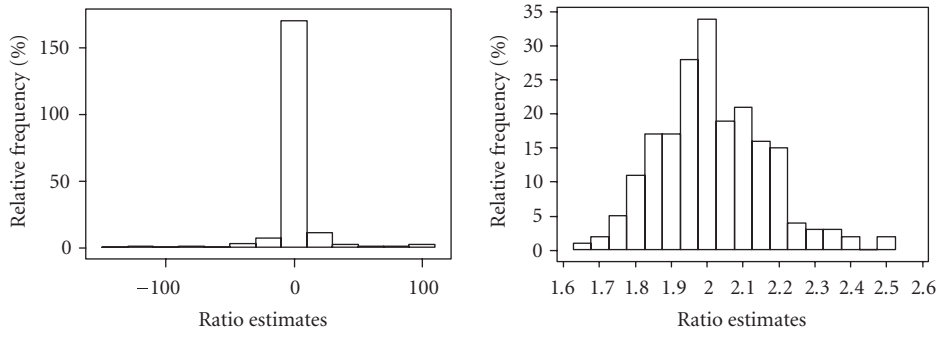
Proof. From the weak law of large numbers, $\overline{X}_n \xrightarrow{P} \mu_X$ and $\overline{Y}_n \xrightarrow{P} \mu_Y$. Take $g(x, y) = x/y$, $X_n = \overline{X}_n$, $Y_n = \overline{Y}_n$, $X = \mu_X$, and $Y = \mu_Y$ in the above lemma and the theorem follows immediately. □

It is the behaviour of X/Y for Y is near zero that permits us only to conclude that $\overline{R}_W$ converges to μ_X/μ_Y in probability. Ensuring that $\mu_Y \neq 0$ and $\mathrm{CV}_Y < 0.2$ in practice allows us to avoid estimation difficulties, when using $\overline{R}_W$.

3.5. Simulation study of the two estimators. Random samples were generated, using software SAS 8.2, from two independent normal distributions, $N(\mu_X, \sigma_X)$ and $N(\mu_Y, \sigma_Y)$, to evaluate the relative merits of the two ratio estimators. The coefficients of variation of the two populations were assumed equal, or $\sigma_X/\mu_X = \sigma_Y/\mu_Y = \mathrm{CV}$. A preliminary simulation was conducted to compare the distributions of the two estimators graphically. Systematic simulations were then conducted for a more in-depth evaluation of the distributions using parameter values typically found in agricultural studies. Mean and standard deviation were examined for each of the two estimators.

3.6. A preliminary simulation. The distributions of $\overline{R}_A$ and $\overline{R}_W$ were simulated from two independent normal populations for the particular case, where $\mu_X = \sigma_X = 200$ and $\mu_Y = \sigma_Y = 100$, hence $\mu_X/\mu_Y = 2$ and there is moderately large population variation ($\mathrm{CV}_X = \mathrm{CV}_Y = 1$). Two hundred samples, each of size $n = 300$, were drawn from each of the numerator and denominator populations. The distributions of $\overline{R}_A$ and $\overline{R}_W$ are graphically compared in Figure 3.1.

The central tendency is different between the two estimators, with the mean of the $\overline{R}_W$ estimates being almost the same as the true ratio of two, and that of the $\overline{R}_A$ estimates further away. Furthermore, the variance of the former is much smaller than that of the latter. This indicates that $\overline{R}_A$ gives some unusually large or extraordinarily small values while $\overline{R}_W$ is concentrated near the true ratio. In this fairly typical example, it is evident that the weighted average is better at estimating the ratio of the two population means. The reason for this contrast is explained as follows. For ratio estimator $\overline{R}_A$, since $X \sim N(200, 200)$ and $Y \sim N(100, 100)$, $\mathrm{CV}_Y = 100/100 = 1 > 0.2$. Thus the $\overline{R}_A$ estimates are meaningless, since the values of $\overline{R}_A = (\sum_{i=1}^{300} X_i/Y_i)/300$ are extremely variable. For ratio estimator $\overline{R}_W$, in contrast, $\overline{X}_{300} \sim N(200, 200/\sqrt{300})$, $\overline{Y}_{300} \sim N(100, 100/\sqrt{300})$, thus $\mathrm{CV}_{\overline{Y}_{300}} = 100/\sqrt{300} = 0.0577 < 0.2$. Hence, the values of $\overline{R}_W = \overline{X}_{300}/\overline{Y}_{300}$ are close to



(a) Distribution of the arithmetic average (b) Distribution of the weighted average

Figure 3.1. A comparison of the distributions of $\overline{R}_A$ and $\overline{R}_W$ using 200 ratio estimates. Note the different scales used in (a) and (b), where the mean and standard deviation over the 200 estimates are 1.430 and 18.465 for $\overline{R}_A$ and 2.015 and 0.152 for $\overline{R}_W$, respectively.

Table 3.1. Impact of sample size on the mean and standard deviation of $\overline{R}_A$ and $\overline{R}_W$.

Sample size	1	4	25	100	400
Mean of $\overline{R}_A$	3.135	0.669	0.744	0.540	0.478
Standard deviation of $\overline{R}_A$	32.968	19.407	7.820	7.328	6.216
Mean of $\overline{R}_W$	3.135	1.042	1.038	1.014	1.001
Standard deviation of $\overline{R}_W$	32.968	1.591	0.302	0.147	0.073
$\mathrm{CV}_{\overline{Y}}$	1	0.5	0.2	0.1	0.05

$\mu_X/\mu_Y = 2$, leading to useful $\overline{R}_W$ values. In conclusion, $\overline{R}_W$ here is a better estimator of μ_X/μ_Y than $\overline{R}_A$.

3.7. Comparison as sample size changes, with coefficient of variation fixed. Here we illustrate the effect of increasing sample size on the two estimators, when $\mathrm{CV}_Y > 0.2$. We use X and Y independently drawn from two $N(100, 100)$ distributions, whence $\mathrm{CV}_X = \mathrm{CV}_Y = 1$; 200 random samples of 1, 4, 25, 100, and 400 pairs of observations (x_i, y_i) were generated. Table 3.1 summarised the distributions of $\overline{R}_A$ and $\overline{R}_W$ for each sample size, where the means and standard deviations are based on 200 samples in each cell of the table and $\overline{R}_A = \overline{R}_W$ when $n = 1$. Results show that the weighted average settles down to the true ratio of one as the sample size increases. The arithmetic average $\overline{R}_A$ always fails to estimate μ_X/μ_Y, whereas with increasing sample size, $\mathrm{CV}_{\overline{Y}}$ falls under 0.2 and the weighted average $\overline{R}_W$ becomes a useful estimator of μ_X/μ_Y.

Note that $\overline{R}_A$, even as the sample size increases, shows no tendency to approach the true ratio of one. In fact, the mean of $\overline{R}_A$ took arbitrary values as sample size increased. On the other hand, the distribution of $\overline{R}_W$ centres on the true ratio as the sample size increases. In particular, for sample sizes of 25 or more (whence $\mathrm{CV}_{\overline{Y}} < 0.2$), $\overline{R}_W$ performs well.

In summary, $\overline{R}_W$ unlike $\overline{R}_A$, improves as an estimator of μ_X/μ_Y under moderate increases in sample size. The major difference between $\overline{R}_A$ and $\overline{R}_W$ is mainly because the latter has a better theoretical basis as an estimator for μ_X/μ_Y. The advantage of $\overline{R}_W$ over $\overline{R}_A$ in reducing the estimation bias, however, depends on the sample size.

4. Application of the two estimators in rice trials

The grain yield data of the rice breeding MET are used in an attempt to evaluate the relative merits of the two estimators of the ratio of independent normal variables in agricultural research. Detailed results of the analyses using both estimators were listed in Table 1.1. An examination of the correlations between the numerator (X) and denominator (Y) variables shows that there was no significant correlation between the yield of each of the three test varieties (X) and that of the control variety (Y). Hence, the following analysis assuming independent normal variables is justified. (Under the assumption of (X, Y) having a bivariate normal distribution, corr $(X, Y) = 0$ implies that X and Y are independent.)

4.1. Estimation of the pooled percent yield improvement over control. Here the ratio of averages $\overline{R}_W$ represents the expected performance of the test variety across the whole region, while the average ratio $\overline{R}_A$ could be regarded as an indicator of what might be expected at any particular location. The choice of the two estimators depends predominantly on the aims of the research, rather than purely on their statistical properties. Since the emphasis was on testing for broad adaptation of the crop varieties, or to summarise information on the overall performance of each cultivar, relative to the control, over the whole range of environments (region), $\overline{R}_W$ is thus a naturally better option than $\overline{R}_A$. As far as specific adaptation is concerned, the $\overline{R}_A$ may have its merit in that it has a better relationship with the expected performance of the variety at a particular location. This, however, is out of the scope of the present study.

The results show that there is a degree of variation in the difference between $\overline{R}_A$ and $\overline{R}_W$ for the three test varieties, ranging from 1.4% to 3.3% (Table 1.1). Estimators $\overline{R}_A$ and $\overline{R}_W$ demonstrate greater difference for the two test varieties 850011 and Yan 501 than for Chang 90–40. From the plant breeding point of view, there is reason (to be discussed in the next subsection) to believe that differences of such magnitude between $\overline{R}_A$ and $\overline{R}_W$ for rice varieties are sufficiently large to change the conclusions of the plant breeding METs.

It is regulated by the Jilin Provincial Crop Variety Evaluation Committee [3] that a new variety of a self-pollinated crop species such as rice has to exceed the control, in grain yield, by at least 5% over three consecutive years before it can be considered for release and commercialisation. The regulation imposed by the committee is most stringent, and it is usually difficult for a test variety to increase grain yield by an extra 1% against the control variety. Thus a 1% difference between the two ways of estimating the pooled ratio of the two rice varieties under comparison can make a real difference in deciding whether a particular variety should be released. Therefore, based on the observed difference between $\overline{R}_A$ and $\overline{R}_W$ for the three varieties, it is evident that the two ways of estimating the ratio of normal variables can influence the decision of plant breeding in terms of recommendation for release and commercialisation. The findings of this paper indicate

that the weighted average ratio estimator $\overline{R}_W$ should be used in practical agricultural research.

4.2. Application of the diagnostic approach in rice trials. The difference between these two estimators ranges from 1.4% to 3.3%, depending on the coefficient of variation for the denominator variable, the grain yield of the control. When the CV of the control is larger than 0.2, as in the case for 850011 and Yan 501, the two estimators differ by a reasonably large amount, 2.6% and 3.3%, respectively. The $\overline{R}_A$ is unreliable in this case, while $\overline{R}_W$ should be used to demonstrate the yield potential of the two varieties relative to the control. In comparison, in the case of Chang 90–40, the CV of the control is only 0.174 (below 0.2) and hence the difference between $\overline{R}_W$ and $\overline{R}_A$ is relatively much smaller. Thus, the difference between $\overline{R}_W$ and $\overline{R}_A$ is dependent on the CV of the grain yields for the control (denominator variable) over the range of environments in which the test variety is being compared with the control. Furthermore, $\mathrm{CV}_{\overline{Y}}$, the CV of the denominator of estimator $\overline{R}_W$, is always much smaller than CV_Y, the CV of the denominator of estimator $\overline{R}_A$, for each of the three comparisons between the test varieties and the control (Table 1.1). This clearly demonstrates the advantage of using the weighted average method in these situations.

Based on the $\overline{R}_W$ estimates of all test rice varieties, only 850011 exceeded the control in grain yield by more than 5% in 1994. By standards commonly adopted in the province, a particular variety will qualify for possible release only if it has outperformed (exceeded) the control in grain yield by 5% or more for all three years of the Provincial Regional Test. Thus, if 850011 continued to outperform the control by 5% or more in grain yield for another two years in the Regional Test, it would be recommended for release, as long as its other agronomic traits have reached the relevant levels of standards. The other two test varieties (Chang 90–40 and Yan 501) have both failed to exceed the control in grain yield by the threshold of 5%. Hence, both varieties were regarded as having no potential for future release from this round of regional trials.

Further studies will focus on a comparison of weighted and arithmetic average estimators under assumption of dependence. Another potential estimator of μ_X/μ_Y, the geometric mean of the X/Y ratios, may prove useful under this circumstance, since it may possess some potentially valuable attributes. A comprehensive investigation of these estimators is thus justified.

5. Conclusions

The mean of the ratio X/Y of two independent normal variables does not exist. The mean appears to exist, however, and is close to μ_X/μ_Y, if we avoid sampling points for which $|Y| \leq \varepsilon$, with ε being a small positive quantity. This favourable situation is approximated in practice when the coefficient of variation of the denominator variable is sufficiently small (less than 0.2). In such circumstances, the ratio of two independent variables can be used to estimate μ_X/μ_Y.

The coefficient of variation of the denominator should thus be considered when estimating a ratio of independent normal variables; the weighted average method automatically reduces denominator coefficient of variation as sample size increases and so is

better than the arithmetic average method. We recommend the use of the weighted average approach for estimating the true ratio from a series of ratio estimates in agricultural research. The arithmetic average approach, however, has to be adopted when only the individual ratios are recorded.

Using the weighted average estimates of all test rice varieties in the motivation example, we concluded that only rice variety 850011 exceeded the control in grain yield by more than 5% in 1994. If 850011 continued to outperform the control by 5% or more in grain yield for another two years in the three-year Provincial Regional Test, it would be recommended for release, as long as its other agronomic traits have reached the relevant levels of standards.

The empirically determined critical coefficient of variation value (0.2) for the denominator of the ratio of independent normal variables can be used to evaluate the suitability of both estimators. A practical diagnostic formula has been proposed to assess the reliability of the weighted average ratio estimator, namely that the coefficient of variation for the denominator mean $\overline{Y}_n$ is smaller than 0.2. The arithmetic average ratio estimator is of less use and should be employed only when the coefficient of variation for the denominator is smaller than 0.2. The development of a satisfactory estimator of the ratio when X and Y are dependent remains an area for future research.

Acknowledgments

The authors are grateful to the two referees for their valuable comments, which helped to improve the quality of the manuscript.

References

[1] R. L. Hall, *Inverse moments for a class of truncated normal distributions*, Sankhyā. Series B **41** (1979), no. 1-2, 66–76.

[2] A. K. M. A. Haque, N. H. Choudhury, M. A. Quasem, and J. R. Arboleda, *Rice post-harvest practices and loss estimates in Bangladesh - part III: parboiling to milling*, Agricultural Mechanization in Asia, Africa and Latin America **28** (1997), 51–55.

[3] Jilin Provincial Crop Variety Evaluation Committee, *Crop Variety Evaluation Regulations of Jilin Province*, Jilin Provincial Science and Technology Press, Changchun, 1995.

[4] Jilin Provincial Seed Station, *Report on Regional Crop Variety Tests of Jilin Province in 1994*, Jilin Provincial Science and Technology Press, Changchun, 1995.

[5] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed., Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, vol. 1, John Wiley & Sons, New York, 1994.

[6] S. M. Kaeppler, J. L. Parke, S. M. Mueller, L. Senior, C. Stuber, and W. F. Tracy, *Variation among maize inbred lines and detection of quantitative trait loci for growth at low phosphorus and responsiveness to arbuscular mycorrhisal fungi*, Crop Science **40** (2000), 358–364.

[7] D. B. Kamerud, *Solution to problem 6104: the random variable X/Y, X, Y normal*, American Mathematical Monthly **85** (1978), no. 3, 206–207.

[8] C. D. Lai, G. R. Wood, and C. G. Qiao, *The mean of the inverse of a punctured normal distribution and its application*, Biometrical Journal **46** (2004), no. 4, 420–429.

[9] E. Lukacs, *Stochastic Convergence*, 2nd ed., Academic Press, New York, 1975.

[10] E. Lukacs and R. G. Laha, *Applications of Characteristic Functions*, Griffin's Statistical Monographs & Courses, no. 14, Hafner, London, 1964.

[11] L. Moreau, H. Monod, A. Charcosset, and A. Gallais, *Marker-assisted selection with spatial analysis of unreplicated field trials*, Theoretical and Applied Genetics **98** (1999), no. 2, 234–242.
[12] S. Nahmias and S. S. Wang, *Approximating partial inverse moments for certain normal variates with an application to decaying inventories*, Naval Research Logistics Quarterly **25** (1978), no. 3, 405–413.
[13] C. G. Qiao, K. E. Basford, I. H. DeLacy, and M. Cooper, *Evaluation of experimental designs and spatial analyses in wheat breeding trials*, Theoretical and Applied Genetics **100** (2000), no. 1, 9–16.
[14] C. G. Qiao, G. R. Wood, and C. D. Lai, *Estimating a binomial proportion from several independent samples*, New Zealand Journal of Crop and Horticultural Science **33** (2005), 293–302.
[15] D. L. Robinson, C. D. Kershaw, and R. P. Ellis, *An investigation of two dimensional yield variability in breeders' small plot barley trials*, Journal of Agricultural Science **111** (1988), 419–426.
[16] M. D. Springer, *The Algebra of Random Variables*, John Wiley & Sons, New York, 1979.
[17] J. R. Witcombe, R. Petre, S. Jones, and A. Joshi, *Farmer participatory crop improvement. IV. The spread and impact of a rice variety identified by participatory varietal selection*, Experimental Agriculture **35** (1999), no. 4, 471–487.

C. G. Qiao: Centre for Social Research and Evaluation, Ministry of Social Development,
P.O. Box 1556, Wellington, New Zealand
E-mail address: chungui.qiao001@msd.govt.nz

G. R. Wood: Department of Statistics, Macquarie University, NSW 2109, Australia
E-mail address: gwood@efs.mq.edu.au

C. D. Lai: Institute of Information Sciences and Technology, Massey University, Private Bag 11 222,
Palmerston North, New Zealand
E-mail address: c.lai@massey.ac.nz

D. W. Luo: Analytical Development, Fonterra Marketing & Innovation, Private Bag 11029,
Palmerston North, New Zealand
E-mail address: dongwen.luo@fonterra.com

EFFECTIVENESS OF HIGH INTEREST RATE POLICY ON EXCHANGE RATES: A REEXAMINATION OF THE ASIAN FINANCIAL CRISIS

TIM BRAILSFORD, JACK H. W. PENM, AND CHIN DIEW LAI

Received 25 December 2005; Revised 29 June 2006; Accepted 30 June 2006

One of the most controversial issues in the aftermath of the Asian financial crisis has been the appropriate response of monetary policy to a sharp decline in the value of some currencies. In this paper, we empirically examine the effects on Asian exchange rates of sharply higher interest rates during the Asian financial crisis. Taking account of the currency contagion effect, our results indicate that sharply higher interest rates helped to support the exchange rates of South Korea, the Philippines, and Thailand. For Malaysia, no significant causal relation is found from the rate of interest to exchange rates, as the authorities in Malaysia did not actively adopt a high interest rate policy to defend the currency.

1. Introduction

Many countries at the centre of the Asian financial crisis adopted a high interest rate policy in an attempt to defend their currencies. This action is consistent with the traditional view in which tight monetary policy is believed to be necessary for supporting a currency, as higher interest rates increase the return for investing in a country and hence reduce capital outflows, and discourage speculative attacks on the currency concerned.

However, many economists have argued a revisionist view. They believe that when balance of payment crises occur simultaneously with financial crises, as is the case of the Asian financial crisis, a tightening of monetary policy may be counter productive. This is because, they argue, sharply higher interest rates will adversely affect economic activity and financial market confidence. Consequently, such a policy response will lead to further currency depreciation such as Feldstein [6].

Empirical testing of this issue has so far yielded mixed results. Most of the previous studies are not supportive of the use of sharply higher interest rates to defend the currency during financial crises. However, these studies also do not produce findings that support

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 35752, Pages 1–9
DOI 10.1155/JAMDS/2006/35752

the revisionist view. Many studies fail to identify any significant relationship between interest and exchange rates in the crisis-affected countries.

Recently, Dekle et al. [4] have provided results that empirically support the traditional view. For South Korea, Malaysia, and Thailand, their results indicate changes in interest rates "Granger cause" movements in their respective exchange rates during the Asian financial crisis. Dekle et al. adopt the approach introduced by Hsiao [10] in which a parsimonious vector autoregressive (VAR) specification is determined that allows the presence of zero and nonzero patterned coefficients. The presence of zero and nonzero patterned coefficients appears to have significantly contributed to their findings.

Although Hsiao's approach allows the presence of zero and nonzero patterned coefficients in a VAR system, the model specification is determined by applying an order selection algorithm to each single equation separately, rather than to the system as a whole. As demonstrated by Penm and Terrell [15], the so-determined specification can lead to misleading conclusions on the presence of Granger causality in the system. To overcome this shortfall, Penm and Terrell [15] provide a robust algorithm to select the optimal VAR specification with zero and nonzero patterned coefficients (if the underlying system has such a structure). Brailsford et al. [1] develop an adjustment to this algorithm, which ensures that the resultant variance-covariance matrix of the white noise disturbance process is symmetric for the determined VAR.

In this paper, we reexamine the existence of Granger causality from interest to exchange rates for four Asian countries that were at the centre of the Asian financial crisis, namely Thailand, Malaysia, the Philippines, and South Korea. We apply the algorithm developed by Brailsford et al. [1] to daily observations during the crisis period. An innovative approach that has been adopted in this study is that the relationship is examined allowing for the presence of contagion effects from movements in other crisis-affected Asian currencies. The presence of contagion effects during the Asian financial crisis has been well documented (e.g., Nagayasu [13]). Given the strong contagion effect during the Asian financial crisis, misleading results can be obtained if such an effect is not accounted for in the model.

This paper is organised as follows. In Section 2, brief reviews of the previous studies and interest and exchange rate movements during the Asian financial crisis are presented. Due to the nature of this study, the literature review is limited to those that focus on the testing of Granger causality. In Section 3, we discuss the VAR specifications determined by the procedure developed by Brailsford et al. [1]. We also employ the method presented by Geweke [7] to measure the linear dependence in the systems. Out-of-sample forecasting is then undertaken using the determined VAR models. These results are presented in Section 4, and a summary is given in Section 5.

2. Previous empirical evidence

The nature of the interest and exchange rate relationships in the Asian financial crisis has been subjected to a significant debate among international organisations and researchers. For example, the International Monetary Fund argues that sharp rises in interest rates are helpful in stabilising Asian exchange rates (IMF [11]). On the other hand, the World

Bank believes that significantly higher interest rates destabilised the Asian currencies by markedly increasing the risks of business bankruptcy and economic contraction (Caporale et al. [2]).

Numerous studies have employed Granger causality testing to investigate whether sharply higher interest rates supported or weakened Asian exchange rates during the Asian financial crisis. These studies present mixed results about the effectiveness of using sharply higher interest rates to support Asian exchange rates. Based on the full-order VAR techniques, Goldfajn and Baig [8] estimate the relationship between interest and exchange rate data for a number of Asian countries and find little evidence supporting the use of higher interest rates. Similarly, Kaminsky and Schmukler [12] estimate full-order VAR models using daily nominal interest and exchange rates to calculate the corresponding impulse response functions. Their results also indicate little interaction between interest and exchange rates in either direction. Using full-order VARs in levels, Choi and Park [3] reexamine this issue. They include spot and forward exchange rates and interest rate differentials in their study and conclude that no causal relationship from interest rate differentials to spot exchange rates exists for the countries they investigate.

Attempts have also been made to examine this issue using error correction modelling techniques. For example, Gould and Kamin [9] estimate, in the VECM framework, the relationship between the real exchange rate and domestic interest rates for a number of Asian countries and Mexico. They also include international credit spreads and domestic stock prices, as proxies for creditworthiness and country specific risk, in order to improve the estimation. However, changes in domestic interest rates are still found to be insignificant in influencing movements in the exchange rates.

Notwithstanding the above results, Park et al. [14] use daily observations to test for causal relations between interest and exchange rates in South Korea. They report evidence of Granger causality from higher interest rates to exchange rate movements during the crisis period. As discussed above, Dekle et al. [4] report similar results. Using weekly observations, higher interest rates are found in Granger cause movements in the exchange rates of a number of Asian countries, including South Korea, Thailand, and Malaysia.

Interest and exchange rate movements during the crisis. The Asian financial crisis started in Thailand in mid-1997, with the Thai baht under significant pressure due to speculative currency attacks. The initial responses from the Thai government were intervention in the foreign exchange market and introduction of capital controls. Following a significant worsening of the foreign reserve position, the baht was floated in early July 1997. As these measures failed to stem the sharp decline in the value of the baht, the Thai government sought assistance from the IMF in early August 1997. After an agreement was reached with the IMF, interest rates in Thailand were raised sharply and kept relatively high for the remainder of 1997 and early 1998. Toward mid-1998, interest rates were gradually reduced, following a gradual return of stability in the currency market.

Following the speculative attacks on Thailand's currency, Malaysia's ringgit and the Philippines'peso were also under significant downward pressure as a result of the contagion effect. In Malaysia, the initial response from the government was a sharp increase in the official interest rate. However, this increase lasted only for a short while before interest rates were reduced to the preshock level. Because of a relatively sound foreign reserve

position, Malaysia did not seek assistance from the IMF and interest rates in that country remained relatively stable.

At the beginning of the Asian financial crisis, the overnight interest rate differential between the Philippines and the United States was the widest in the region. Significant downward pressure on the peso emerged in August 1997. Domestic interest rates in the Philippines became unstable in the second half of 1997. For example, in early October 1997, the overnight interbank call rate increased from around 12 per cent to 102 per cent within a few days, before falling back to the preshock level in late October. In early 1998, the peso exhibited some stability against the US dollar. Consequently, movements in domestic interest rates became less volatile.

Korea's currency, the won, depreciated gradually between July and September 1997, partly reflecting the contagion effect of the currency instability in South-East Asia. The overnight interest rate differential with the United States also gradually widened over this period. Between late October and early December 1997, a crisis of debt financing in Korea emerged, leading to significant downward pressure on the Korean won. In response, domestic interest rates were raised significantly. The Korean government also sought assistance from the IMF in early December 1997.

In the first few months of 1998, the Korean exchange rate was volatile and so was the overnight call rate. A solution emerged, after an agreement was reached with foreign banks to roll over most of Korea's short-term debts, with stability gradually returning to the foreign exchange market.

3. Empirical test results

In this section, we present the empirical results of testing for Granger causality between interest and exchange rates for the above-mentioned four Asian countries. Since the debate has focused on the effectiveness of using higher interest rates to defend a sharp decline in currency, we have therefore concentrated this testing over the Asian financial crisis period (defined as from 1 July 1997 to 1 July 1998).

For ease of comparison with previous studies, we have adopted a similar model to Dekle et al. [4], which includes the variables, daily overnight interest rate differential with the United States, exchange rate against the US dollar, and producer price differential with the United States (approximated by the monthly index movements). To capture the currency contagion effect during the crisis period, we also include the exchange rate of the Malaysian ringgit against the US dollar in the models for Thailand, the Philippines, and South Korea. In the case of Malaysia, the Thai baht against the US dollar is used as a proxy.

In the calculation of interest rate differentials, we use the overnight interbank rates for the Asian countries and the daily repo rate for the United States. Following Dekle et al. [4], we employ observations over the whole crisis period. Data of interest and exchange rates were obtained from Datastream. Producer price indexes were from *International Financial Statistics*.

To undertake this testing, a pretest strategy is followed by first examining for the presence of unit roots and also, where applicable, for cointegration in the models. Based on the ADF test (Dickey and Fuller [5]), all the exchange rates over the crisis period can be

characterised as integrated of order 1. While the series of overnight interest rate differential for South Korea is found integrated of order 1, those for Thailand, Malaysia, and the Philippines are stationary. The producer price differentials are also found to be stationary. The hypothesis of cointegration is rejected for each individual system using the Stock and Watson [16] test.

The zero and nonzero patterned VAR specifications are determined using the algorithm developed by Brailsford et al. [1] (together with the Schwarz criterion). No discussion is given on this procedure for the reason of brevity. Interested readers are referred to Brailsford et al. [1] and Penm and Terrell [15] for details.

The estimation results based on the Zellner [17] SUR are presented in Table 3.1. We also apply the Brailsford et al. procedure to the estimated residuals to ensure that they can be characterised as white noise.

As demonstrated by the determined VAR specifications (see Table 3.1), changes in overnight interest rate differentials are found to have affected the exchange rates of Thailand, the Philippines, and South Korea during the Asian financial crisis. For Malaysia, however, the determined zero and nonzero patterned VAR specification indicates that the variable, overnight interest rate differential, is independent of the rest of the system. Consequently, we omit this variable from the system and present the estimation results for Malaysia without this variable.

This finding for Malaysia suggests that over the Asian financial crisis period, interest rate movements in that country do not significantly influence movements in the Malaysian ringgit against the US dollar. This is in contrast to the finding of Dekle et al. [4], but consistent with a priori expectations, as the Malaysian government did not actively adopt a high interest rate policy to defend its currency during the Asian financial crisis.

As mentioned above, we also include the exchange rate of the Thai baht against the US dollar in the system for Malaysia, as a proxy for the contagion effect. As presented in Table 3.1, the one-period lagged Thai exchange rate variable is selected to explain movements in the Malaysian ringgit. The estimated relationship is statistically significant at the 5 per cent level. The coefficient estimate indicates that a depreciation of the Thai baht against the US dollar Granger causes a depreciation of the Malaysian ringgit against the US dollar during the crisis period.

In the case of Thailand, the one-period lagged differential in overnight interest rates is selected as an explanatory variable for the exchange rate. Although the estimated coefficient has a sign consistent with a priori expectations, the associated t-statistics is not significant at the 5 per cent level, casting doubts on the test results. In an attempt to improve the estimation, we also first difference the variable, interest rate differential, and repeat the selection procedure. The one-period lagged interest rate differential is again selected in the exchange rate equation, but the coefficient estimate remains insignificant at the 5 per cent level. Consequently, we conclude that, for Thailand, only weak evidence is obtained for the presence of Granger causality from domestic interest rate movements to the currency.

The one-period lagged Malaysian ringgit is also selected as an explanatory variable for movements in the Thai baht, with a t-statistics significant at the 5 per cent level. The

Table 3.1. VAR estimates. X^1 denotes units of domestic currency per unit of the US dollar, X^2 denotes differential between domestic and US overnight interest rates, X^3 denotes differential between domestic and US producer price indexes, and X^4 denotes another regional exchange rate and d first difference. The zero and nonzero patterned model specifications are determined using the Schwarz criterion.

Thailand	$d\ln X_t^1 = \underset{(1.30)}{0.0150} + \underset{(3.13)}{0.2118} d\ln X_{t-1}^1 - \underset{(2.63)}{0.1401} d\ln X_{t-2}^1 - \underset{(1.50)}{0.0053}\ln X_{t-1}^2 + \underset{(2.48)}{0.1499} d\ln X_{t-1}^4,$ $\ln X_t^2 = \underset{(4.17)}{0.3969} + \underset{(11.31)}{0.6956}\ln X_{t-1}^2 + \underset{(2.50)}{0.1530}\ln X_{t-2}^2,$ $X_t^3 = \underset{(0.95)}{0.0376} + \underset{(64.56)}{0.9720} X_{t-1}^3,$ $d\ln X_t^4 = \underset{(1.24)}{0.0014} + \underset{(2.06)}{0.1578} d\ln X_{t-1}^1 + \underset{(1.45)}{0.1008} d\ln X_{t-1}^4$
Malaysia	$d\ln X_t^1 = \underset{(1.22)}{0.0014} + \underset{(1.53)}{0.1049} d\ln X_{t-1}^1 + \underset{(1.98)}{0.1447} d\ln X_{t-1}^4,$ $X_t^3 = \underset{(1.61)}{0.0665} + \underset{(56.91)}{0.9553} X_{t-1}^3,$ $d\ln X_t^4 = \underset{(1.01)}{0.0010} + \underset{(2.35)}{0.1398} d\ln X_{t-1}^1 + \underset{(2.87)}{0.1834} d\ln X_{t-1}^4$
The Philippines	$d\ln X_t^1 = \underset{(1.18)}{0.0061} - \underset{(2.65)}{0.1441} d\ln X_{t-2}^1 + \underset{(2.41)}{0.0075} d\ln X_{t-4}^1 - \underset{(4.24)}{0.0240} d\ln X_{t-6}^2 + \underset{(3.02)}{0.0143} d\ln X_{t-7}^2 + \underset{(7.25)}{0.3923} d\ln X_{t-1}^4,$ $\ln X_t^2 = \underset{(5.23)}{0.3608} + \underset{(20.41)}{1.1824}\ln X_{t-1}^2 - \underset{(5.81)}{0.3348}\ln X_{t-2}^2 - \underset{(1.67)}{0.0138} X_{t-1}^3,$ $X_t^3 = \underset{(1.51)}{0.0506} + \underset{(521.97)}{0.9584} X_{t-1}^3,$ $d\ln X_t^4 = \underset{(1.38)}{0.0016} + \underset{(2.66)}{0.1655} d\ln X_{t-1}^4$
South Korea	$d\ln X_t^1 = \underset{(0.45)}{0.0008} + \underset{(5.77)}{0.3298} d\ln X_{t-1}^1 - \underset{(2.32)}{0.0690} d\ln X_{t-1}^2 + \underset{(1.86)}{0.0549} d\ln X_{t-3}^2 + \underset{(1.95)}{0.1782} d\ln X_{t-1}^4,$ $d\ln X_t^2 = \underset{(-0.87)}{-0.0033} + \underset{(1.72)}{0.1963} d\ln X_{t-2}^1 - \underset{(3.07)}{0.1824} d\ln X_{t-3}^2 + \underset{(3.92)}{0.0255} X_{t-2}^3 - \underset{(3.30)}{0.0215} X_{t-3}^3,$ $X_t^3 = \underset{(0.59)}{0.0211} + \underset{(3.48)}{3.6846} d\ln X_{t-3}^1 - \underset{(3.02)}{1.6718} d\ln X_{t-1}^2 + \underset{(77.84)}{0.9782} X_{t-1}^3,$ $d\ln X_t^4 = \underset{(1.17)}{0.0013} + \underset{(3.93)}{0.1467} d\ln X_{t-1}^1 + \underset{(2.61)}{0.1563} d\ln X_{t-1}^4$

sign of this coefficient estimate is consistent with a priori expectations, reinforcing the significance of currency contagion during the Asian financial crisis.

For the Philippines, the estimated relationship between interest rate differential and movements in the exchange rate appears dynamic. In the equation of exchange rate, the six-period lagged interest rate differential is selected with a negative coefficient, and the seven-period lagged interest rate differential is selected with a positive coefficient. Based on the coefficient estimates, the net effect of a widening of interest rate differential in the Philippines causes an appreciation of the peso against the US dollar. In contrast to the results for Thailand, these coefficient estimates are statistically significant at the 5 per

cent level, which give strong support for the presence of Granger causality from interest rate movements to the exchange rate. In addition to the effects of interest rate changes, the lagged movements in the Malaysian ringgit significantly influence the peso during the crisis.

For South Korea, the one-period lagged interest rate differential is selected in the equation of exchange rate with a negative coefficient, and the three-period lagged interest rate differential is selected with a positive coefficient. The estimation results indicate that during the Asian financial crisis, higher interest rates adopted by the Korean government help to support the won. The variable, one-period lagged movements in the Malaysian ringgit, is also selected as an explanatory variable for movements in the Korean won. The coefficient estimate indicates that a depreciation of the Malaysian ringgit also results in a decline in the value of the won against the US dollar over the crisis period.

4. Measurement of linear dependence

To further understand the effects of interest rate movements on the Asian currencies over the crisis period, we measure the linear dependence in the above VAR systems using the Geweke [7] approach. Two cases of interest are presented in Table 4.1. Testing at the 5 per cent level, 95 per cent confidence intervals are shown parenthetically. In the first case, the linear dependence on the interest rate differential is calculated. In the second case, the linear dependence on the contagion effect is measured. Because of the determined specifications, these measures effectively indicate the linear dependence of the exchange rate on changes in interest rate differential and the contagion effect.

In Table 4.1, the measures indicate that the interest rate effects varied among the Asian exchange rates. In the case of the Philippines, the linear dependence of its peso on interest rate differential is stronger than that for Thailand and South Korea. This effect is also higher than the impact on the currency of the contagion effect.

However, in the cases of Thailand and South Korea, the linear dependence of their exchange rates on changes in interest rate differentials is less significant than the contagion effects. These results indicate that, for these two countries, the exchange rate movements during the Asian financial crisis are more significantly influenced by currency contagion. Despite sharply higher interest rates imposed by the authorities, such a policy response is unable to prevent their currencies from declining against the US dollar.

An important question raised by these results is the appropriateness of using a high interest rate policy to defend the currency, especially in the presence of significant currency contagion. There are economic consequences associated with sharply higher interest rates. For example, sharply higher interest rates, if sustained, will lead to a marked slowdown in economic activity.

To further demonstrate the impact of changes in interest rate differentials and currency contagion on movements in the Asian exchange rates during the crisis period, we also examine the forecasting performance of our models. To undertake this exercise, we divide the sample into two periods. The first period consists of data from 1 July 1997 to 18 June 1998 and the second period consists of data from 19 June to 1 July 1998. We use data from the first period to reestimate the VAR specification for each country and then produce the

Table 4.1. Measurement of linear dependence. Confidence intervals in brackets. Nonzero measurement indicates the existence of Granger causality.

	$X_t = [x_t^1 \quad x_t^3 \quad x_t^4]$ $Y_t = [x_t^2]$		$X_t = [x_t^1 \quad x_t^2 \quad x_t^3]$ $Y_t = [x_t^4]$	
Thailand	0.007	[−0.002 0.015]	0.032	[0.002 0.653]
Malaysia	—	Not available	0.025	[0.001 0.580]
The Philippines	0.176	[0.002 0.345]	0.059	[0.002 0.131]
South Korea	0.013	[0.001 0.035]	0.045	[0.005 0.094]

Table 4.2. Forecasting performance.

	AR	VAR including interest rate and currency contagion	Improvement
Thailand	1.53%	1.49%	2.85%
Malaysia	2.44%	2.33%	4.52%
The Philippines	1.43%	0.75%	47.3%
South Korea	3.18%	2.46%	22.5%

forecasts for the second period. To examine the forecasting performance, we calculate the root mean squared error (RMSE) for the respective exchange rate over the forecast period, expressed as the percentage of the sample mean over the forecast period (see Table 4.2).

For the purpose of comparison, we also construct a set of univariate autoregressive (AR) systems for the four Asian exchange rates using the Brailsford et al. procedure. Similarly, these AR systems are first estimated using observations from the first period. Forecasts for the second period are then produced and the RMSEs are calculated.

Table 4.2 presents the improvement in forecasting performance of our models. The results indicate that interest rate movements and currency contagion are two important factors in the determination of Asian exchange rates during the crisis period. Consistent with the estimation results presented in Table 3.1, the improvement in forecasting performance is particularly significant for the Philippines and South Korea.

5. Summary

In this paper, we have reexamined the effects on Asian exchange rates of higher interest rates during the Asian financial crisis. In contrast to most previous studies, we find that higher interest rates provided support for many Asian exchange rates during the crisis. This finding is consistent with the traditional view about this relationship. We find no evidence to support the revisionist view, in which sharply higher interest rates are argued to lead to a weaker exchange rate during financial crises.

Currency contagion is found to be significant in the Asian financial crisis. In the cases of Thailand and South Korea, the contagion effects on their currencies are deemed to be more significant than the impacts of sharply higher interest rates. This finding raises questions about the appropriateness of using a high interest rate policy to defend an exchange rate, especially in the presence of contagion.

References

[1] T. Brailsford, J. H. W. Penm, and R. D. Terrell, *The adjustment of the Yule-Walker relations in VAR modelling: the impact of the Euro on the Hong Kong stock market*, Multinational Finance Journal **5** (2001), no. 1, 35–58.

[2] G. Caporale, A. Cipollini, and P. Demetriades, *Monetary policy and the exchange rate during the Asian crisis: identification through heteroscedasticity*, working paper, Department of Economics, University of Leicester, Leicester, 2000.

[3] I. Choi and D. Park, *Causal relation between interest and exchange rates in the Asian currency crisis*, working paper, Kookmin University, Seoul, 2000.

[4] R. Dekle, C. Hsiao, and S. Wang, *High interest rates and exchange rate stabilization in Korea, Malaysia, and Thailand: an empirical investigation of the traditional and revisionist views*, Review of International Economics **10** (2002), no. 1, 64–78.

[5] D. A. Dickey and W. A. Fuller, *Distribution of the estimators for autoregressive time series with a unit root*, Journal of the American Statistical Association **74** (1979), no. 366, part 1, 427–431.

[6] M. Feldstein, *Refocusing the IMF*, Foreign Affairs **77** (1998), no. 2, 20–33.

[7] J. Geweke, *Measurement of linear dependence and feedback between multiple time series*, Journal of the American Statistical Association **77** (1982), no. 378, 304–324.

[8] I. Goldfajn and T. Baig, *Monetary policy in the aftermath of currency crises: the case of Asia*, working paper, International Monetary Fund, Washington, DC, 1998.

[9] D. Gould and S. Kamin, *The impact of monetary policy on exchange rate during financial crises*, working paper, Board of Governors of the Federal Reserve System, Washington, DC, 1999.

[10] C. Hsiao, *Autoregressive modelling of Canadian money and income data*, Journal of the American Statistical Association **74** (1979), no. 367, 553–560.

[11] International Monetary Fund, *The role of monetary policy in responding to currency crises*, World Economic Outlook, , Washington, DC, 1998.

[12] G. Kaminsky and S. Schmukler, *The relationship between interest rates and exchange rates in six Asian countries*, working paper, The World Bank, Washington, DC, 1998.

[13] J. Nagayasu, *Currency crisis and contagion: evidence from exchange rates and sectoral stock indices of the Philippines and Thailand*, Journal of Asian Economics **12** (2001), no. 4, 529–546.

[14] Y. C. Park, C.-S. Chung, and Y. Wang, *Exchange rate policies in Korea: has exchange rate volatility increased after the crisis?*, working paper, Korean Institute for Economic Policy, East Asian Bureau of Economic Research, 1999.

[15] J. H. W. Penm and R. D. Terrell, *Multivariate subset autoregressive modelling with zero constraints for detecting overall causality*, Journal of Econometrics **24** (1984), no. 3, 311–330.

[16] J. H. Stock and M. W. Watson, *Testing for common trends*, Journal of the American Statistical Association **83** (1988), no. 404, 1097–1107.

[17] A. Zellner, *An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias*, Journal of the American Statistical Association **57** (1962), no. 298, 348–368.

Tim Brailsford: UQ Business School, The University of Queensland, Brisbane QLD 4072, Australia
E-mail address: t.brailsford@business.uq.edu.au

Jack H.W. Penm: School of Finance & Applied Statistics, College of Business and Economics, The Australian National University, Canberra ACT 0200, Australia
E-mail address: jack.penm@anu.edu.au

Chin Diew Lai: Institute of Information Sciences and Technology, Massey University, Palmerston North, New Zealand
E-mail address: c.lai@massey.ac.nz

LOSS PROTECTION IN PAIRS TRADING THROUGH MINIMUM PROFIT BOUNDS: A COINTEGRATION APPROACH

YAN-XIA LIN, MICHAEL McCRAE, AND CHANDRA GULATI

Received 4 September 2005; Revised 10 May 2006; Accepted 15 May 2006

Pairs trading is a comparative-value form of statistical arbitrage designed to exploit temporary random departures from equilibrium pricing between two shares. However, the strategy is not riskless. Market events as well as poor statistical modeling and parameter estimation may all erode potential profits. Since conventional loss limiting trading strategies are costly, a preferable situation is to integrate loss limitation within the statistical modeling itself. This paper uses cointegration principles to develop a procedure that embeds a minimum profit condition within a pairs trading strategy. We derive the necessary conditions for such a procedure and then use them to define and implement a five-step procedure for identifying eligible trades. The statistical validity of the procedure is verified through simulation data. Practicality is tested through actual data. The results show that, at reasonable minimum profit levels, the protocol does not greatly reduce trade numbers or absolute profits relative to an unprotected trading strategy.

1. Introduction

Pairs trading is a statistical arbitrage strategy with a long history of modest but persistent profits on Wall St (Peskin and Boudreau [10]; Gatev et al. [3]). The strategy identifies pairs of shares whose prices are driven by the same economic forces, then trades on any temporary deviations of those two-share prices from their long-run average relationship (Gillespie and Ulph [4]). The arbitrage or risk-free nature of the strategy arises from the opening of opposing positions for each trade—shorting the over-valued share and buying the under-valued share.

The simple statistical techniques used for share pairs selection and trading decisions make pairs trading an appealing arbitrage strategy. But simplicity comes at a cost. Correlation, covariance, and regression analysis of share price associations provide an imprecise, simplistic statistical definition of a long-run equilibrium relationship between share

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 73803, Pages 1–14
DOI 10.1155/JAMDS/2006/73803

prices. Moreover, they do not necessarily imply mean reversion to a long-run equilibrium price spread.

This paper assumes that such deficiencies of the statistical techniques are best dealt with by systematic improvement within the underlying statistical modeling itself, rather than left to costly hedging and conditional order techniques. We use cointegration theory to provide a statistically precise foundation for the decisions involved in pairs trading and then use these principles to derive a loss limiting rule that ensures that each eligible trade will return some preset minimum profit, subject, of course, to the previously mentioned endemic market risks which are always present.

In this paper we use the principle of cointegrated series to derive a precise, dynamic definition of long-run equilibrium price spread that inherently implies mean reversion in component series. We then use cointegration principles to establish a protocol for ensuring that any selected trade will satisfy preset minimum profit conditions.

The paper extends the work on cointegration in pairs trading by Gatev et al. [3], Gillespie and Ulph [4], and Alexander and Dimitriu [1] to integrate a minimum nominal profit requirement into other trading strategy decisions such as the choices about share pairs, dollar weighting of long/short positions, trade opening and closing criteria and total dollar investment. We first use cointegration coefficient weighting (CCW) principles to derive the necessary conditions that will ensure that a trade delivers a preset minimum nominal profit per trade (MNPPT). These conditions are then incorporated into a practical, five step procedure for achieving any given MNPPT.

The analysis proceeds in six sections. Section 2 summarizes pairs trading fundamentals and identifies the main parameter estimates required for a pairs trading strategy. Section 3 introduces the concept of cointegration-based dollar weighting of long/short positions as the theoretical foundation for deriving the necessary conditions of a preset minimum profit per trade. A five step procedure for putting the necessary conditions into practice is presented at the end of Section 3. Section 4 uses the simulated data series to investigate the procedure's sensitivity to alternative opening trade hurdle values under two trading conditions: un-constrained and constrained total investment dollars. Actual daily share price data is used in Section 5 to examine the effect of investment dollar constraints on the number of valid trades for six preset minimum profit levels, with decreasing open condition values at each level. Section 6 discusses the risk minimization implications of our results in the context of arbitrage trading strategies.

The data simulation exercise in Section 4 indicates that while all trades can be immunized in a theoretical sense, the crucial factor that determines the number of eligible trades is the allowable investment dollar maximum, since some trades require large outlays. However, real data are needed to test the practical limitations of the capital requirement.

The practicality of imposing minimum profit conditions is tested on daily closing price data for two Australian Stock Exchange quoted bank shares—the Australia New Zealand Bank (ANZ) and the Adelaide Bank (ADB) over the period January 2, 2001 to August 30, 2002. The results show that trading strategy profit potential is not unduly constrained by adding a reasonable minimum profit condition to protect against losses.

2. Pairs trading and statistical arbitrage

Pairs trading relies on the principle of equilibrium pricing for near-equivalent shares. In efficient markets, capital asset pricing model-based valuation theory and the law of one price require price equality for equivalent financial assets over time (Reilly and Brown [11]; Sharpe et al. [12]). The price spreads of near-equivalent assets should also conform to a long-term stable equilibrium over time. Hendry and Juselius [6] use this principle to show that short-term deviations from these equivalent pricing conditions may create opportunities for arbitrage profits depending upon the size and duration of the price shock.

When a sufficiently large deviation of price spread from the long-run norm is identified, a trade is opened by simultaneously buying (go long) the under-valued share and selling (short) the over-valued share. The trade is closed out when prices return to their equilibrium price spread levels by selling the long position and off-setting the short position. Net trading profit sums the profits from the long and short positions, calculated as the difference between the opening and closing prices (net of trading costs less interest on short sale receipts). See Gillespie and Ulph [4] and L'Habitant [8].

The "risk free" characteristic of pairs trading arises from the simultaneous long-short (buy-sell) opening market positions. The opposing positions ideally immunize trading outcomes against systematic market-wide movements in prices that may work against uncovered positions (see Jacobs and Levy [7]).

But arbitrage trading of the "convergence trade" type is rarely risk-less. Market events, persistent pricing inefficiencies or structural price changes may invalidate statistical pricing models, confound future price expectations or require parameter reestimation. Price spreads after position opening may escalate rather than revert, or the equilibrium position may shift. The inherent nature of losses were dramatically demonstrated by the unraveling of long-term capital management's highly leveraged long/short sovereign bond positions in the late 90s (Lowenstein [9]).

Pairs trading is also exposed to risk from the inherent limitations in the statistical techniques used to identify and extract profit potential. Traditional techniques may appeal in their simplicity but suffer severe limitations as a foundation for trading decision choices that determine arbitrage profit potential and extraction.

The profit reduction consequences of these risks may be offset by loss limitation strategies including stop loss and time limit orders and derivatives hedging. But these strategies are costly and only limit rather than prevent loss. With regard to statistical inefficiency, a preferable situation is integrating loss protection into the statistical modeling itself. This paper develops and tests such a procedure by using cointegration theory to define the necessary conditions for ensuring a minimum nominal profit before a trade is opened. The next section describes the foundation for this analysis.

3. Cointegration-based strategies

Alexander et al. [2] demonstrate that the arbitrage profit potential between two shares depends critically on the presence of a long-term equilibrium spread between share prices, the existence of short-run departures (price shocks) from that equilibrium and

re-convergence to equilibrium. In this situation, the statistical technique used for pairs trading must be able to provide an effective model of share price time behavior; detect equilibrium value relationships, and provide a measure of the extent and size of short-term variations from that equilibrium relationship. Gatev et al. [3], Gillespie and Ulph [4] and Alexander and Dimitriu [1] all suggest that cointegration theory offers a more integrative framework for statistical arbitrage strategies than current techniques.

Definition 3.1. A time series X_t is called an $I(1)$ series if the first difference of the time series forms a stationary series, denoted by $I(0)$.

Many share price series are $I(1)$ series. Therefore, the following cointegration definition is given based on $I(1)$ series.

Definition 3.2. Let $X_{1t}, X_{2t}, \ldots, X_{nt}$ be a sequence of $I(1)$ time series. If there are nonzero real numbers $\beta_1, \beta_2, \ldots, \beta_n$ such that

$$\beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_n X_{nt} \tag{3.1}$$

becomes an $I(0)$ series, then $X_{1t}, X_{2t}, \ldots, X_{nt}$ are said to be cointegrated.

Cointegrated price series possess a stationary long-run stable equilibrium relationship with the associated property of mean reversion. By the definition, the linear combination of cointegrated price series is stationary and will always revert back to the mean of the stationary series. This is an important fact, which will ensure that the pair trading technique developed in this paper becomes practicable. Further details on cointegration analysis can be found in Harris [5].

We now use the concept of cointegrated share price series to derive the necessary conditions for ensuring a given minimum profit per trade over a selected trading horizon. The selection of share pairs and all subsequent trading decisions is made on cointegration principles. Long and short positions are weighted by their cointegration coefficients rather than being equally weighted. Trade opening boundaries are defined in terms of deviations from the linear combination of cointegrated series rather than deviations from some absolute mean-spread value.

The following assumptions simplify the analysis:

(A1) the two-share price series are always cointegrated over the relevant time horizon;

(A2) long and short positions always apply to the same shares in the share pair. For any trade, S_1 always represents the short position while S_2 represents the long position;

(A3) at the opening of any trade, the price for the shorted share S_1 is always higher than the price of the share in long position S_2.

Remarks 3.3. (1) Since divergence from equilibrium pricing is random, any one share in a pair is just as likely to be over-priced as under-priced. However, since we are only concerned with profit per trade and since any one's trade must be concluded before the next trade is started, assumption (A2) does not affect either the validity of the simulation or empirical tests in relation to the ability of the necessary conditions to generate greater than minimum required profits per trade.

(2) Assumption (A3) is quite general since a linear transformation of a price series has no effect on its cointegrating properties.

(3) The above simplifying assumptions do not affect the validity or generality of the "necessary conditions" tests, except that assumption (A3) under-estimates the number of possible trades in any period, since, on standardized prices, the shorted share, S_1, is just as likely to be below S_2 as above.

3.1. Profit produced by a completed pair trade. The first step is to derive a profit formula for a pair of shares S_1 and S_2.

Let t_o and t_c be the times of opening and closing out a trade position, respectively. A trade is opened at t_o when a preset open trade condition (OTC) is met. The trade is closed out at t_c when a preset close trade condition (CTC) is met.

The following notations are used in the analysis. Denote by

$N_{S_1}(t_o)$ the number of shares in short position at t_o;
$N_{S_2}(t_o)$ the number of shares in long position at t_o;
$P_{S_1}(t_o)$ the price of share S_1 at t_o;
$P_{S_1}(t_c)$ the price of share S_1 at t_c;
$P_{S_2}(t_o)$ the price of share S_2 at t_o;
$P_{S_2}(t_c)$ the price of share S_2 at t_c.

A trade is opened when the OTC is met at time t_0. The over-valued share, S_1, is shorted (sold), so that $N_{S_1}(t_o)$ shares are sold for the receipt of $N_{S_1}(t_o)P_{S_1}(t_o)$ dollars. A long position on $N_{S_2}(t_o)$ shares is taken in the under-valued share S_2 at a cost of $N_{S_1}(t_o)P_{S_2}(t_o)$ dollars.

The trade is then closed out when the CTC is met at time t_c, by simultaneously selling the long position shares for the receipt of $P_{S_2}(t_c)N_{S_2}(t_0)$ dollars and buying back the $N_{S_1}(t_o)$ of S_1 shares at a cost of $N_{S_1}(t_o)P_{S_1}(t_c)$ dollars.

Thus, the total profit from the trade is

$$TP_{t_c} = N_{S_2}(t_o)\left[P_{S_2}(t_c) - P_{S_2}(t_o)\right] + N_{S_1}(t_o)\left[P_{S_1}(t_o) - P_{S_1}(t_c)\right]. \tag{3.2}$$

A trade is profitable if and only if $TP_{t_c} > 0$. So a loss prevention strategy equates to ensuring that $TP_{t_c} > 0$ or, more generally, that $TP_{t_c} > K > 0$ for any preset value K.

3.2. The conditions of minimum profit per trade under the CCW rule. We now establish a cointegration coefficient weighting (CCW) rule and derive the conditions necessary to ensure a minimum profit per trade (MPPT).

Under (A1), the prices of shares S_1 and S_2 are cointegrated; say

$$P_{S_1}(t) + \beta P_{S_2}(t) = \varepsilon_t, \quad t \geq 1 \tag{3.3}$$

where ε_t is an $I(0)$ series.

The following study is restricted to the situation where the cointegration coefficient $\beta < 0$. This condition is not restrictive since previous studies show that most cointegrated share price series conform to this condition.

To ensure that the money gained from S_1 at t_o will cover the outlay to buy S_2 at t_o, we need the following condition for opening a trade:

$$N_{S_1}(t_o)P_{S_1}(t_o) \geq N_{S_2}(t_o)P_{S_2}(t_o). \tag{C1}$$

In general, a trade can be opened at any time as long as (C1) is satisfied. Here we introduce an open trade criterion by the following.

Open trade condition (OTC(a)). Let a be a positive real number. A time t_o can be considered as an open trading time if t_o satisfies the following condition:

$$P_{S_1}(t_0) + \beta P_{S_2}(t_0) = \varepsilon_{t_0} > a > 0. \tag{3.4}$$

To ensure that both conditions OTC(a) and (C1) are true, a condition on $N_{S_1}(t_o)$ and $N_{S_2}(t_o)$ needs to be imposed. If a trader decides to buy n shares, that is, $N_{S_2}(t_o) = n$, then, the trader should sell at least $n/|\beta|$ shares in the short position. For simplicity, fractional share holdings are permitted. In this situation we will have $P_{S_1}(t_o)N_{S_1}(t_o) > P_{S_2}(t_o)N_{S_2}(t_o)$. So (C1) holds. After manipulation, under OTC(a) with $N_{S_2}(t_o) = n$ and $N_{S_1}(t_o) = n/|\beta|$, the total profit made at time t_c can be calculated below:

$$\begin{aligned} TP_{t_c} &= N_{S_2}(t_0)\left[P_{S_2}(t_c) - P_{S_2}(t_0)\right] + N_{S_1}(t_0)\left[P_{S_1}(t_0) - P_{S_1}(t_c)\right] \\ &= \frac{n}{\beta}\left\{\left[\varepsilon_{t_c} - P_{S_1}(t_c)\right] - \left[\varepsilon_{t_o} - P_{S_1}(t_o)\right]\right\} + \frac{n}{|\beta|}\left[P_{S_1}(t_0) - P_{S_1}(t_c)\right] \\ &= -\frac{n}{|\beta|}\left[P_{S_1}(t_o) - P_{S_1}(t_c)\right] + \frac{n}{|\beta|}\left[P_{S_1}(t_o) - P_{S_1}(t_c)\right] \\ &\quad + \frac{n}{|\beta|}\left(\varepsilon_{t_o} - \varepsilon_{t_c}\right) = \frac{n\left(\varepsilon_{t_0} - \varepsilon_{t_c}\right)}{|\beta|}. \end{aligned} \tag{3.5}$$

This derivation shows that for any pair of cointegrated shares, if at open time t_o the number of shares in the long and short positions are $N_{S_2}(t_o) = n$ and $N_{S_1}(t_o) = n/|\beta|$, respectively, the total profit from long/short trading can be expressed solely in terms of β, ε_{t_o}, ε_{t_c}, and n.

We now need to define an appropriate close time t_c such that a trader, who opened a trade under OTC(a) with $N_{S_2}(t_o) = n$ and $N_{S_1}(t_o) = n/|\beta|$, will be able to gain a minimum of \$$K$ when the trader closes the trade.

From (3.5), to ensure the minimum gain requirement, ε_{t_c} has to satisfy the following inequality:

$$\frac{n\left(\varepsilon_{t_0} - \varepsilon_{t_c}\right)}{|\beta|} > K. \tag{3.6}$$

In other words, the value of ε_{t_c} has to be lower than ε_{t_o} and the difference between ε_{t_o} and ε_{t_c} has to be greater than $|\beta|K/n$. Thus, to ensure a minimum profit of \$$K$, we use the following closing condition.

Close trade condition (CTC(a), (b)). If a trade is opened at OTC(a) with $N_{S_2}(t_o) > K|\beta|/(a-b)$ and $N_{S_1}(t_o) = N_{S_2}(t_o)/|\beta|$, where $a > b$, then the trade needs to be closed at t_c when $\varepsilon_{t_c} < b$.

In practice, $N_{S_1}(t_o)$ can take value $[N_{S_2}(t_o)/|\beta|] + 1$, in case $N_{S_2}(t_o)/|\beta|$ is not an integer, where "$[d]$" denotes the maximum integer less than d.

3.3. Five-step trading strategy. We now use the above necessary conditions to build a five-step procedure for obtaining the required minimum profit \$$K$ on any completed trade.

Step 1. Choose an opening condition a and closing condition b such that $a > b$. Usually b is assigned as the mean of ε_1 and a is assigned as $k\sigma$ where k is a positive real number and σ is the standard deviation of ε_1 (recall that ε_t is a stationary time series).

Step 2. Choose an integer $n > K|\beta|/(a-b)$.

Step 3. Open a trade at t_o when $P_{S_1}(t_o) > P_{S_2}(t_o)$ and condition OTC(a) is satisfied.

Step 4. Buy n shares of S_2 and sell $[n/|\beta|] + 1$ shares of S_1 at time t_0.

Step 5. Close out the trade at t_c when $\varepsilon_{t_c} < b$.

Following the above steps, we have

$$\frac{n(\varepsilon_{t_0} - \varepsilon_{t_c})}{|\beta|} > \frac{n(a-b)}{|\beta|} > K, \tag{3.7}$$

which will ensure a given MPPT of \$$K$ for the trade.

In the above strategy, the proportion of shares assigned to the long and short positions is determined by the cointegration coefficient β rather than by the more traditional equal weighting strategy. We label this the cointegration coefficient weighting strategy (CCW).

Remark 3.4. In practice, the CCW weighting strategy will always work if the total dollar investment is permitted by the broker. This is because the open and close conditions are based on the movement of the stationary time series ε_t. To ensure an appropriate frequency of trades, the opening condition (a) and the closing condition (b) should be chosen such that they are regularly crossed by the process ε_t, thus ensuring the frequent opening and closing of trades.

4. The application of minimum profit conditions

The preceding analysis derived the theoretical conditions for achieving a given MPPT and formulated a five step procedure to implement the trading strategy. We now examine practical application issues of constraints imposed by the procedure on numbers of trades and sensitivity to maximum investment levels.

The theoretical derivation of the necessary conditions for achieving a given level of minimum profit may be enhanced if the procedure is a practical one in terms of its impact on trade numbers and trading profitability. The current analysis concentrates on profit per trade, trade numbers, and dollar investment implications.

Table 4.1. Total profit and trades for varying MPPT levels under CCW strategy: simulated data.

K	Open condition $*$ (a)	Close condition $*$ (b)	Average total profit	Average total trades
10	$m+\sigma$	m	1350	51.8
50	$m+\sigma$	m	6006	51.8
100	$m+\sigma$	m	11751	51.8
10	$m+(\sigma/2)$	m	3248	81.7
50	$m+(\sigma/2)$	m	15180	81.7
100	$m+(\sigma/2)$	m	30032	81.7

$*m$ *and* σ *denote the mean and standard deviations, respectively, of the* $I(0)$ *series* ε_t

.

We investigate the trade number and profit constraint issues using data generated from a known cointegration model. The simulation study has two purposes. First, to demonstrate how the CCW strategy works for the simulated data and whether altering the values for (a) and (b) affects the number of trades for any given MPPT level. Second, to demonstrate the effect on trade numbers of introducing a constraint on the total dollar investment allowed in any trade.

As an investigative technique, simulation enhances control over the data generation process by ensuring that sample data conform to a given cointegration model with the prescribed parameters. This filters out data "noise" that may complicate results on the effects of the treatment variable on the target variable/s.

The sample price data are simulated from a cointegration model:

$$\begin{aligned} P_{S_1}(t) + \beta P_{S_2}(t) &= \varepsilon_t, \\ P_{S_2}(t) - P_{S_2}(t-1) &= e_t, \end{aligned} \tag{4.1}$$

where $e_t - 0.1e_{t-1} = 13 + \delta_{1,t}$ and $\delta_{1,t}$ are iid normally distributed, $N(0, 0.5)$; $\beta = -0.2$; ε_t follows model $\varepsilon_t - 0.2\varepsilon_{t-1} = 13 + \delta_t$ and $\{\delta_t\}$ are iid with standard normal distribution $N(0,1)$. 100 independent samples are simulated from this model and each sample has 500 data points equally spaced over the trading horizon to permit calculation of profit per time period over a horizon of 500 periods. Simulations are run with \$$K$ equal to \$10, \$50, and \$100, respectively.

4.1. Application to the CCW strategy: unconstrained investment. Following the five step trading strategy, the simulation output in Table 4.1 is given by setting $N_{S_2}(t_o) = [K|\beta|/(a-b)] + 1$ and $N_{S_1}(t_o) = N_{S_2}(t_o)/|\beta|$ for each trade.

Table 4.1 shows that the average total trade numbers per sample is just over 51. When the value (a) is closer to the mean of ε_t, the average total trade numbers increase to over 81 trades per sample.

Under the CCW rule, the number of trades in a trading horizon is largely determined by the open and close criterion values. Since both criteria now relate to the stationary time series ε_t, reconvergence to the long-run equilibrium value m is more frequent.

Table 4.2. Total profit and trades for varying MPPT under CCW strategy with constrained investment dollars: simulated data.

K	W	Open condition (a)	Close condition (b)	Average total profit	Average total trades
10	90000	$m+\sigma$	m	523	15.25
10	100000	$m+\sigma$	m	875	25.37
10	250000	$m+\sigma$	m	1932	51.78
10	100000	$m+(\sigma/2)$	m	0	0
10	250000	$m+(\sigma/2)$	m	3999	81.75
50	250000	$m+\sigma$	m	0	0
50	400000	$m+\sigma$	m	5864	46.50
50	250000	$m+1.5\sigma$	m	1523	14.16
50	100000	$m+1.5\sigma$	m	0	0
100	400000	$m+1.5\sigma$	m	1114	5.69

4.2. Application to the CCW strategy: constrained investment dollars. In the previous unconstrained CCW simulation, the total dollar investment in long/short positions cannot be preset. They depend on the price of shares at each open trade position. So while minimum profit \$$K$ requirement is met, the total dollars investment required to produce this result may be large. In this section, another simulation study is considered. This simulation constrains the total dollar investment permitted per trade. Trades that require \$$W$ investment above the indicated values are now deleted. Table 4.2 presents the results.

The results indicate that the size of the average dollar commitment per trade necessary to meet the MPPT condition can make the rate of return on investment very small at the given entry hurdle—even when set below the prevailing risk free rate. However, recall that we are not deriving a profit maximizing strategy, but a strategy ensuring a given minimum profit. The simulations emphasize the sensitivity of required capital outlay to the other decision parameter values. Keeping outlays feasible implies the selection of realistic parameter values and reasonably priced shares relative to intended outlay. Expensive shares require more capital.

The low rate of return on investment may reflect the lack of price shocks in the simulated model of this section. A more realistic test of returns requires actual data. The next section details an empirical investigation of these results.

5. Application of the CCW strategy to empirical data

We now use empirical data to examine how alternative levels of maximum investment affect trade numbers for a given MPPT value K. Maximum investment limits (W) are set at \$5.000 through to \$100.000 for given alternative MPPT levels K of \$10 through to \$2000. Within each investment level, the opening condition is varied from $m+1.5\sigma$ to $m+\sigma/5$ and the closing condition is always set at m, where m and σ are the mean and standard deviations of ε_t in (5.1), respectively.

The data are daily closing prices from January 2, 2001 to August 30, 2002 for two Australian Stock Exchange quoted bank shares—the Australia New Zealand Bank (ANZ) and the Adelaide Bank (ADB). The cointegration parameters are estimated on the first year's data, that is, from January 2, 2001 to January 1, 2002. Both price series are $I(1)$ processes with a stationary cointegrated spread of the form:

$$P_{\mathrm{ADB}}(t) + \beta P_{\mathrm{ANZ}}(t) = \varepsilon_t, \tag{5.1}$$

where ε_t is an $I(0)$ series. The estimate of β is $-1/2.0237 = -0.4941$. The model is then applied to the data from January 2, 2002 to August 30, 2002, which are 167 trade days. The outputs are presented by Tables 5.1 and 5.2.

Several patterns emerge from the tables. First, at least one valid trade is generated at all MPPT levels, except where the open condition becomes too low to allow potential trades to develop at the given investment levels. Predictably, the number of valid trades yielding a given MPPT increases with increased investment dollars. Second, a reduction in open trade boundary values increases the number of valid trades and then falls to zero trades as the spread becomes too small to generate trades within the given investment levels. This pattern reflects the functional relationship between the open condition level and the level of MPPT.

Third, the number of valid trades may appear low for all MPPT levels. But the restrictive nature of the second analytical assumption makes the results conservative. The restriction of valid trades to those situations where S_2 is the shorted share will eliminate number of potential trades. So the actual number of trades in an unrestricted trading situation is probably higher than those reported here at all MPPT levels.

The purpose of the analysis did not include an examination of the effects of the MPPT procedure on the total profit levels of pairs trading. However, the total profit figures for the trading during the 167 days are included in Tables 5.1 and 5.2. At all MPPT levels the rate of return on investment increases as open condition boundaries are lowered, until they become too low to generate eligible trades at the MPPT level within the given investment levels. The pattern and level of increases in the rate of return on investment appear consistent across increasing levels of MPPT levels and invariant to that level.

6. Discussion

In this paper we derived a cointegration-based procedure that would always return at least a given minimum profit level. We then tested the feasibility of the procedure in terms of the number of possible trades that could be immunized at different MPPT levels for several combinations of open trade values, and investment dollars. The results of the empirical analysis suggest that the five-step strategy is feasible for commonly used parameter values.

Pairs trading strategies involve several decision choices. Taken together, these choices determine how much arbitrage profit potential is actually extracted from each pairs trade. Our cointegration-based analysis provides exploitable information on the long-run time

Table 5.1. Total profit and trades under varying MPPT for three levels of investment: ANZ and ADB share pairs.

K	W	Open condition (a)	Close condition (b)	Number of trades	Total profit P	TP/W
10	5000	$m+1.5\sigma$	m	1	12.76	0.00255
		$m+\sigma$		1	14.08	0.00282
		$m+0.75\sigma$		2	26.26	0.00525
		$m+(\sigma/2)$		2	37.26	0.00745
		$m+(\sigma/3)$		3	74.38	0.01487
50	5000	$m+1.5\sigma$	m	1	63.83	0.01276
		$m+\sigma$		1	68.42	0.01368
		$m+0.75\sigma$		2	131.30	0.02626
		$m+(\sigma/2)$		2	186.31	0.03726
		$m+(\sigma/3)$		0	0	0
50	10000	$m+1.5\sigma$	m	1	63.83	0.00638
		$m+\sigma$		1	68.42	0.00684
		$m+0.75\sigma$		2	131.30	0.01313
		$m+(\sigma/2)$		2	186.31	0.01863
		$m+(\sigma/3)$		3	368.26	0.03683
100	5000	$m+1.5\sigma$	m	1	127.65	0.02553
		$m+\sigma$		1	135.84	0.02716
		$m+0.75\sigma$		0	0	0
		$m+(\sigma/2)$		0	0	0
		$m+(\sigma/3)$		0	0	0
100	10000	$m+1.5\sigma$	m	1	127.65	0.01277
		$m+\sigma$		1	135.84	0.01358
		$m+0.75\sigma$		2	262.60	0.02626
		$m+(\sigma/2)$		2	372.61	0.03726
		$m+(\sigma/3)$		0	0	0
100	50000	$m+1.5\sigma$	m	1	127.65	0.00255
		$m+\sigma$		1	135.84	0.00272
		$m+0.75\sigma$		2	262.60	0.00525
		$m+(\sigma/2)$		2	372.61	0.00745
		$m+(\sigma/3)$		3	734.70	0.01469

Table 5.2. Total profits and trades under varying MPPT with constant constrained investment: ANZ and ADB share pairs.

K	W	Open condition (a)	Close condition (b)	Number of trades	Total profit P	TP/W
		$m+1.5\sigma$		1	636.84	0.00637
		$m+\sigma$		1	678.18	0.00678
		$m+0.75\sigma$		2	1310.06	0.01310
500	100000	$m+(\sigma/2)$	m	2	1858.92	0.01859
		$m+(\sigma/3)$		3	3666.25	0.03666
		$m+(\sigma/4)$		4	5998.04	0.05998
		$m+(\sigma/5)$		0	0	0
		$m+1.5\sigma$		1	1273.67	0.01274
		$m+\sigma$		1	1355.36	0.01355
		$m+0.75\sigma$		2	2620.11	0.02620
1000	100000	$m+(\sigma/2)$	m	2	3717.84	0.03718
		$m+(\sigma/3)$		0	0	0
		$m+(\sigma/4)$		0	0	0
		$m+(\sigma/5)$		0	0	0
		$m+1.5\sigma$		1	2547.34	0.02547
		$m+\sigma$		1	2710.72	0.02710
		$m+0.75\sigma$		0	0	0
2000	100000	$m+(\sigma/2)$	m	0	0	0
		$m+(\sigma/3)$		0	0	0
		$m+(\sigma/4)$		0	0	0
		$m+(\sigma/5)$		0	0	0

series behavior of share pairs that is not available through currently used statistical methods. Unlike these current techniques, cointegration also offers a technique for systematically analyzing the interdependence of strategic choices. Our analysis shows that the profitability of a pairs trading strategy depends upon using weighting rules, minimum profit hurdles, and open/close criterion that reflect traders' preferences and are appropriate to the short and long-run price behavior of the component shares. Unrealistic values imply low trading rates, excessive trade durations, and low profits per share trade. Through cointegration the trader has a tool for investigating the statistical relationship between parameters.

Our analysis also emphasizes a range of other fundamental issues in statistical arbitrage strategies that require further study.

(1) The contribution to arbitrage profit of each share depends upon relative price volatilities and mean-reversion characteristics of the component shares.
(2) "Success" in pairs trading is a compromise between arbitrage levels and profit levels. Alternative weighting rules may optimize one objective but not both. For example, at the extreme, the most profitable strategy is to weight investment in the more volatile share at 100 percent and zero weight the other share, but this strategy offers minimal systematic risk protection.
(3) Any trading strategy is a compromise between trading frequency, duration, and per trade profitability. Arbitrage profit levels depend on achieving a suitable mix relevant to the price series behavior of a given pair of financial assets.
(4) For cointegrated share pairs, the latent profit potential relates directly to both the size and the frequency of short-term shocks characterizing each price series. Exploiting that potential depends on strategic choices.

Pairs trading, although limited to the simplest long/short case of two shares, is directly congruent with the much wider case of *n-share* long/short portfolios. Moreover, since there is no reason why pairs trading should not use put and call options rather than the underlying shares, our statistical analysis also translates to the derivatives portfolio context. It also reflects the statistical equivalent of the economic maxim that there are no "*free lunches*."

Acknowledgments

The paper has benefited from comments of Professor S. Gupta, Professor B. D. Sharma, and participants of the SCRA Conference, University of Maine, 2003, and Dr. W. Friesling, Director of Research, Commonwealth Bank, Australia, and participants at the 2003 IMMACS Workshop on Mathematical Finance, University of Wollongong. The authors are thankful to the anonymous referees for useful comments and remarks that led to a better exposition of the results.

References

[1] C. Alexander and A. Dimitriu, *The cointegration alpha: enhanced index tracking and long-short equity market neutral strategies*, 2002, Discussion Papers in Finance ISMA Center 2002-08, University of Reading.
[2] C. Alexander, I. Giblin, and W. Weddington III, *Cointegration and asset allocation: a new active hedge fund strategy*, 2001, Discussion Papers in Finance ISMA Center 2001-03, University of Reading.
[3] E. Gatev, W. Goetzmann, and G. Rouweunhorst, *Pairs trading: performance of a relative value arbitrage rule*, Working Paper 7032, National Bureau of Economic Research, Washington DC, 1999.
[4] T. Gillespie and C. Ulph, *Pair trades methodology: a question of mean reversion*, Proceedings of International Conference on Statistics, Combinatorics and Related Areas and the 8th International Conference of Forum for Interdisciplinary Mathematics, NSW, December 2001, unpublished paper.
[5] R. Harris, *Using Cointegration Analysis in Econometric Modelling*, Prentice Hall, London, 1995.
[6] D. Hendry and K. Juselius, *Explaining cointegration analysis: part II*, Energy Journal **22** (2001), no. 1, 75–120.

[7] B. Jacobs and K. Levy, *Long/short equity investing*, Journal of Portfolio Management **20** (1993), no. 1, 52–63.
[8] F.-S. L'Habitant, *Hedge Funds: Myths and Limits*, John Wiley & Sons, Chichester, 2002.
[9] R. Lowenstein, *When Genius Failed: The Rise and Fall of Long-Term Capital Management*, Random House, New York, 2000.
[10] M. Peskin and B. Boudreau, *Why hedge funds make sense*, 2000, http://www.thehfa.org/articles/1.pdf.
[11] F. Reilly and K. Brown, *Investment Analysis and Portfolio Management*, 6th ed., Harcourt College, New York, 2000.
[12] W. F. Sharpe, G. J. Alexander, and J. V. Bailey, *Investments*, 6th ed., Prentice-Hall, New Jersey, 1999.

Yan-Xia Lin: School of Mathematics and Applied Statistics, University of Wollongong, Northfields Avenue, Wollongong, NSW 2500, Australia
E-mail address: yanxia@uow.edu.au

Michael McCrae: School of Finance and Accounting, University of Wollongong, Northfields Avenue, Wollongong, NSW 2500, Australia
E-mail address: mccrae@uow.edu.au

Chandra Gulati: School of Mathematics and Applied Statistics, University of Wollongong, Northfields Avenue, Wollongong, NSW 2500, Australia
E-mail address: cmg@uow.edu.au

MAPPING THE CONVERGENCE OF GENETIC ALGORITHMS

ZVI DREZNER AND GEORGE A. MARCOULIDES

Received 29 August 2005; Revised 25 April 2006; Accepted 6 June 2006

This paper examines the convergence of genetic algorithms using a cluster-analytic-type procedure. The procedure is illustrated with a hybrid genetic algorithm applied to the quadratic assignment problem. Results provide valuable insight into how population members are selected as the number of generations increases and how genetic algorithms approach stagnation after many generations.

1. Introduction

Hybrid genetic algorithms have recently become very popular metaheuristic methods (Beasley [6]). Most genetic algorithms produce offspring by mating parents and attempt to improve the population makeup by replacing existing population members with superior offspring. In contrast, hybrid genetic algorithms, sometimes called memetic algorithms (Moscato [28]), incorporate some heuristic improvement on every offspring before considering its inclusion into the population. For a plethora of introductory expositions published on the topic, see Salhi [33] or Beasley [6].

This paper examines the convergence of genetic algorithms using a cluster-analytic-type technique called the "MD procedure" (Marcoulides and Drezner [26]). The proposed procedure is illustrated with a hybrid genetic algorithm applied to the solution of the quadratic assignment problem (QAP). For a review of the QAP, see Rendl [31]. Because population members form clusters as progress is made to a solution, the clustering structure can provide a better implementation of genetic algorithms. For example, clustering structure can be used to develop better stopping criteria, for instance when the population clusters become stagnant.

In the next section we describe the MD procedure. In Section 3 we describe the quadratic assignment problem and the hybrid genetic algorithm used for its solution. In

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 70240, Pages 1–16
DOI 10.1155/JAMDS/2006/70240

Section 4 we present an analysis of the procedures. Finally, Sections 5 and 6 present the results of some computational experiments and conclusions.

2. The MD procedure

The MD procedure is used to display k-dimensional data in two dimensions so that clusters can be easily observed. The procedure is successful in preserving distances between the various data points, thus retaining the structure of the set of points. It is based on the proposed solution for the layout problem (Drezner [12]), which is a variant of the DISCON (dispersion-concentration) procedure Drezner [11].

The layout problem is very similar to the QAP except that there are no specific locations for the facilities. While the QAP is concerned with finding the best permutation of the facilities among the *given* sites, the layout problem is concerned with the location of facilities of a given size anywhere in the plane. A set of weights $\{w_{ij}\}$, $w_{ij} = w_{ji}$ associated with facility pairs is given. As in the QAP, we wish that pairs of facilities with larger weights be closer to one another in the final configuration.

Drezner [12] proposed to minimize the function

$$\frac{\sum_{i\neq j=1}^{n} w_{ij} d_{ij}^2}{\sum_{i\neq j=1}^{n} d_{ij}^2} \tag{2.1}$$

which is equivalent to

$$\min\left\{ \sum_{i\neq j=1}^{n} w_{ij} d_{ij}^2 \right\} \quad \text{subject to} \quad \sum_{i\neq j=1}^{n} d_{ij}^2 = 1, \tag{2.2}$$

where d_{ij} is the Euclidean distance between the unknown locations of facilities i and j.

Since $d_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2$,

$$\begin{aligned} \sum_{i\neq j=1}^{n} w_{ij} d_{ij}^2 &= \sum_{i\neq j=1}^{n} w_{ij} (x_i - x_j)^2 + \sum_{i\neq j=1}^{n} w_{ij} (y_i - y_j)^2 \\ &= 2\sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} w_{ij} \right\} x_i^2 - 2 \sum_{i\neq j=1}^{n} w_{ij} x_i x_j + 2\sum_{i=1}^{n} \left\{ \sum_{j=1}^{n} w_{ij} \right\} y_i^2 - 2 \sum_{i\neq j=1}^{n} w_{ij} y_i y_j. \end{aligned} \tag{2.3}$$

Define the matrix $S = \{s_{ij}\}$ by $S_{ii} = \sum_{j=1}^{n} w_{ij}$ and $S_{ij} = -w_{ij}$, for $i \neq j$. Our problem is equivalent to minimizing

$$\frac{x^T S x + y^T S y}{x^T E x + y^T E y}, \tag{2.4}$$

where E is the matrix S with all weights equal to 1.

The matrix S is singular and therefore one of its eigenvalues is 0 with an associated eigenvector of $(1,\ldots,1)$. Note that adding a constant to all weights does not change (2.1) or (2.2), and thus we can guarantee that all eigenvalues of S (except the zero eigenvalue) are positive. As is shown by Drezner in [12] and by Marcoulides and Drezner in [26],

the solution to (2.4) is $x = y$, where x or y is the eigenvector associated with the smallest positive eigenvalue. This solution is on a line. To get a two-dimensional solution, we select for the y-coordinates the best solution that is orthogonal to the first solution. This second solution is the eigenvector associated with the *second* smallest positive eigenvalue. These (x, y) coordinates provide a solution to the layout problem in the plane.

Marcoulides and Drezner [26] suggested the use of this layout algorithm to transform k-dimensional data to a two-dimensional scatter plot, while retaining the special structure of the data. They proposed to use the reciprocal of the distances between the points in the k-dimensional space as weights. In this way, points that are close to each other in the k-dimensional data will tend to be close in the solution of the layout problem. Let D_{ij} be the k-dimensional distances between points i and j. Marcoulides and Drezner [26] suggested to use $w_{ij} = D_{ij}^{-p}$ with a positive p as the weights in (2.1), and search for the best p using the golden section search. For each value of p, the correlation coefficient between the original distances D_{ij} and the calculated two-dimensional distances by the procedure d_{ij} is found, and the best p in $[0, 10]$ that maximizes this correlation coefficient is selected for implementation.

Marcoulides and Drezner [27] suggested the application of the MD procedure for cluster analysis with excellent results. They proposed to use the solution on a line (which is the projection of the scatter diagram on the x-axis). Clusters are identified as follows: the distances between successive points on the line are calculated and large distances between successive points constitute separators between clusters.

3. The quadratic assignment problem

The quadratic assignment problem (QAP) is considered to be one of the most difficult combinatorial optimization problems to solve. The problem is defined as follows. A set of n possible sites is given and n facilities are to be located on these sites, one facility at a site. Let c_{ij} be the cost per unit distance between facilities i and j and let d_{ij} be the distance between sites i and j. A high cost between two facilities means that we wish the two facilities to be close to one another. The cost f to be minimized over all possible permutations, calculated for an assignment of facility i to site $p(i)$ for $i = 1, \ldots, n$, is

$$f = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} d_{p(i)p(j)}. \tag{3.1}$$

Optimal algorithms can solve relatively small problems. Recently, Anstreicher et al. [3], Hahn and Krarup [23], Nystrom [30], and Anstreicher and Brixius [2] report optimal solutions for problems with $n = 30$ to 36 facilities. Such optimal solutions are based on branch-and-bound algorithms which require "good" lower bounds. Gilmore [20] and Lawler [24] proposed the first lower bound based on the simple assignment problem. Anstreicher and Brixius [2] proposed a lower bound based on quadratic programming. Two lower bounds used by Hahn and Grant [21] and Hahn et al. [22] are based on a dual formulation. A dual formulation was suggested by Drezner in [13] and its implementation reported by Resende et al. in [32].

Since optimal algorithms can solve only relatively small problems, considerable effort has been devoted to constructing heuristic algorithms. The first heuristic algorithm proposed for the solution of the QAP was CRAFT (Armour and Buffa [4]) which is a descent heuristic. More recent algorithms use metaheuristics such as Tabu search (Battiti and Tecchiolli [5]; Skorin-Kapov [34]; and Taillard [35]), simulated annealing (Burkard and Rendl [8]; Wilhelm and Ward [38]; and Connolly [10]), genetic algorithms (Ahuja et al. [1]; Fleurent and Ferland [18]; Tate and Smith [37]; and Drezner [15–17]), ant colony search (Gambardella et al. [19]), or specially designed heuristics (Drezner [14]; Li et al. [25]). For a complete discussion and list of references, see Burkard [7], Çela [9], Rendl [31], and Taillard [36].

3.1. The hybrid genetic algorithm. Genetic algorithms maintain a population of solutions. In order to create each generation, two parents are selected and merged to produce an offspring. If the produced offspring is better than the worst population member, the offspring replaces that member. The process continues for a prespecified number of generations. The best population member at the conclusion of the process is the solution of the genetic algorithm. Hybrid genetic algorithms apply an improving procedure on each offspring before considering it for inclusion in the population. Such an improvement procedure produces better offspring and the algorithm usually requires fewer generations to obtain quality solutions. The important components of a hybrid genetic algorithm are the merging process of two parents, and the postmerging improvement algorithm.

For the implementation of the genetic algorithm for the solution of the QAP, each solution (chromosome) is defined by the facilities assigned to sites $\#1, \#2, \ldots, \#n$. The Hamming distance between two solutions is the *number* of facilities located at *different* sites. We use a population of 100 solutions. As the merging procedure, the "cohesive merging procedure" (Drezner [15]) is used. The idea behind the cohesive merging procedure is to select about half of the sites that are close to one another "(cohesive)" and assigning the facilities from the first parent to this cohesive set, and to assign the facilities from the second parent to the rest of the sites. For a complete description, the reader is referred to Drezner [15]. As the postmerging procedure, we use the "short" concentric Tabu search, modified by selecting a random number of levels. The concentric Tabu search was first presented by Drezner in [14] and was used as a postmerging procedure in hybrid genetic algorithms (Drezner [15–17]). The short version was used by Drezner in [16, 17] and gave excellent results.

The postmerging procedure is summarized below. The procedure starts with a solution termed the "center" solution and attempts to find a better solution by checking solutions at increasing Hamming distance from the center solution. This process can be viewed as searching in concentric circles centered at the center solution. The concentric Tabu search (Drezner [14]) stops once one application of the radial search fails to find a better solution. In the short concentric Tabu search, the maximum radius of the concentric searches is randomly generated at $[0.3n, 0.9n]$ which is less than the maximum possible Hamming distance between two solutions (n). The algorithm below applies between 3 and 9 "levels." Each level is a concentric Tabu search, but if the search fails to produce a better solution, a new center solution is selected for the next level.

3.2. The postmerging procedure for the QAP. The procedure starts with a so-called "center" solution. The Hamming distance between permutation p and the center solution is Δp. The procedure proceeds by checking solutions with increasing Hamming distance.

(1) Set a counter $c = 0$. Randomly generate the number of levels L in $[3,9]$ (with probability of 1/7 for each level).
(2) Select R (the radius of the search) randomly in $[0.3n, 0.9n]$.
(3) Set $\Delta p = 0$. sol_0 is the center solution. Empty the solutions sol_1 and sol_2 (the best found solutions for $\Delta p + 1$ and $\Delta p + 2$, resp.).
(4) All pair exchanges of sol_0 are evaluated.
(5) If the exchanged solution is better than the best found solution, the best found solution is updated and the rest of the exchanges are evaluated.
(6) If the distance of an exchanged solution is Δp or lower, it is in the Tabu list. Therefore, it is ignored and the rest of the exchanges are evaluated. (In this way, we force the search away from the center solution.)
(7) If its distance is $\Delta p + 1$ or $\Delta p + 2$, sol_1 or sol_2 is updated, if necessary.
(8) If a new best found solution is found by scanning all the exchanges of sol_0, the starting (center) solution is set to the new best found solution. Go to step (1).
(9) Otherwise, $\text{sol}_0 = \text{sol}_1$, $\text{sol}_1 = \text{sol}_2$, and sol_2 is emptied. Set $\Delta p = \Delta p + 1$.
(10) If $\Delta p \le R$, go to step (4).
(11) If $\Delta p = R + 1$, advance the counter $c = c + 1$, and
  (i) if $c \le L$ and is odd, use the best solution with depth R as the new center solution and go to step (2);
  (ii) if $c \le L$ and is even, use the best solution found throughout the scan (the previous center solution is not considered) as the new center solution and go to step (2);
  (iii) if $c = L + 1$ stop and report the best found solution.

4. Analysis

Hybrid genetic algorithms start with a population of random solutions (each improved by a postmerging procedure) and keep improving the population members by entering better offspring and removing poorer population members. As the number of generations increases, the population members tend to cluster into groups, such that group members are "close" to one another.

In order to analyze this phenomenon, we first define a distance between population members. The Hamming distance is used. The distance between two population members is the number of variables which are different in the two solutions. Thus, two population members are at distance zero from one another if they are identical. Note that this distance measure satisfies the triangle inequality.

Suppose we perform a cluster analysis on a given population. The distance between every pair of population members is calculated and a scatter plot is generated using the MD procedure. The distance matrix is given as input to the MD procedure, and the result is a two-dimensional scatter diagram of the population members. Pairs of population members that are "close" to one another tend to be close to one another in the scatter

diagram. We employ the weights $[D_{ij} - D_{\min} + 1]^{-p}$, where D_{ij} is the Hamming distance between population members i and j, and $D_{\min} = \min_{i \neq j}\{D_{ij}\}$.

We implemented this idea in the analysis of a hybrid genetic (memetic) algorithm for the solution of the quadratic assignment problem. Each solution of the quadratic assignment problem is a permutation of n facilities. Therefore, the distance between two solutions (permutations) can be between 0 (when the permutations are identical) and n. Note that a distance of 1 is impossible.

4.1. Properties of the Hamming distance for the QAP

THEOREM 4.1. *The expected distance between two random permutations is equal to* $n - 1$.

Proof. Let $P_n(k)$ be the probability that two random permutations of n elements have k identical elements. The number of permutations that have k members identical to permutation #1 is $n!P_n(k)$. It is clear that $n!P_n(k) = \binom{n}{k}(n-k)!P_{n-k}(0)$ leading to

$$P_n(k) = \frac{P_{n-k}(0)}{k!}. \tag{4.1}$$

By (4.1),

$$P_{n-1}(k-1) = \frac{P_{n-k}(0)}{(k-1)!} = kP_n(k). \tag{4.2}$$

Therefore,

$$\sum_{k=1}^{n} kP_n(k) = \sum_{k=1}^{n} P_{n-1}(k-1) = 1. \tag{4.3}$$

We showed that the expected number of identical elements in two random permutations is 1, which proves the theorem. □

Theorem 4.1 provides us with a reference for comparison between the average distance among all population members and the expected distance if the population members were random. Thus, if the average distance between all pairs of population members is lower than $n - 1$, then the population is not random.

5. Computational experiments

We selected three QAP problems for analysis: Nug30 (Nugent et al. [29]) of 30 facilities for which the optimum solution of 6124 is known (Anstreicher et al. [3]), Sko56 and Sko100a (Skorin-Kapov [34]) of 56 and 100 facilities, respectively, for which the best known solutions of 34458 and 152002 are not proven optimum yet. We used a population of 100 and therefore the scatter diagram consists of 100 points. Each problem was solved using $50n$ generations, and the results after multiples of $10n$ generations were recorded and analyzed.

In Table 5.1 we report for each problem the minimum and average distances among population members, and the minimum and average values of the objective function for

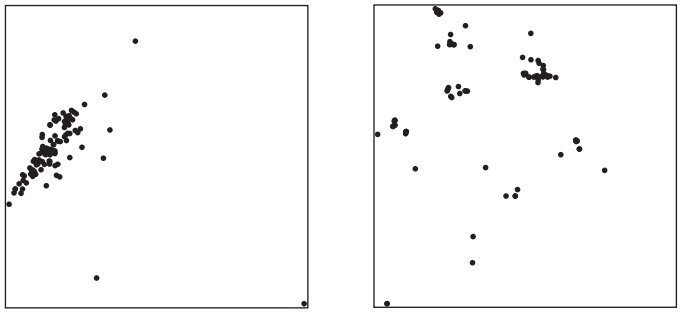
Table 5.1. Distances between population members and objective function values.

Gen.	Nug30				Sko56				Sko100a			
	Distance		Objective		Distance		Objective		Distance		Objective	
	Min.	Aver.	Min.	Aver.	Min.	Aver.	Min.	Aver.	Min.	Aver.	Min.	Aver.
0	0	27.73	6134	6211.48	36	53.84	34512	34906.84	84	97.55	152368	153441.30
$10n$	0	26.98	6124	6160.00	2	45.35	34458	34501.96	2	41.95	152026	152097.70
$20n$	0	26.87	6124	6155.24	2	41.92	34458	34481.68	2	33.32	152026	152075.44
$30n$	0	26.83	6124	6152.54	2	41.36	34458	34479.04	2	30.32	152026	152072.46
$40n$	0	26.70	6124	6151.04	2	41.27	34458	34477.54	2	29.84	152026	152071.74
$50n$	0	26.72	6124	6149.94	2	41.55	34458	34477.24	2	29.77	152026	152071.58

all population members. The starting population of Nug30 (consisting of 100 population members) includes three pairs of identical population members (i.e., at a distance of zero from one another). These pairs of population members have an objective function values of 6146, 6172, and 6190, respectively. Since the hybrid genetic algorithm (Drezner [15–17]) does not allow into the population, the offspring that are identical to existing population members, no more identical population members are added to the population. After $50n$ generations, the worst population member has an objective function value of 6160, and thus two of the identical pairs were removed from the population, and only one of the three pairs is still a member of the population at the end of the process. The optimum solution of 6124 was obtained before $10n$ (300) generations are completed. It seems that the populations do not improve much after 300 iterations. The best known solution for Sko56 was also reached before $10n$ (560) iterations. The populations do not improve much after $20n$ generations. For Sko100a, the procedure obtained the value of 152026 which is slightly higher than the best known solution of 152002 after $10n$ generations as well. It should be noted that the best known solution for Sko100a was obtained frequently with other random seeds (see Section 5.2). The population also seems to have stabilized after $20n$ generations.

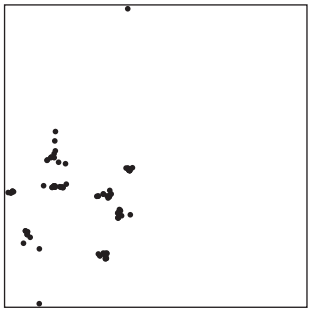
The average distance between population members generally declines as the number of generations increases. However it stabilizes after $20n - 40n$ generations. A random initial population is expected to have an average distance of $n - 1$ by Theorem 4.1. Since a postmerging procedure is applied on the initial population, the initial population is already somewhat clustered (average distance of 27.73 compared with expected of 29 for Nug30, 53.84 compared with 55 for Sko56, and 97.55 compared with 99 for Sko100a). In Figures 5.1, 5.2, and 5.3, we depict the scatter diagrams obtained by the MD procedure. The averages also confirm the scatter diagrams depicted in these figures. The scatter diagram of Nug30 (Figure 5.1) is the most scattered. Therefore, their average distance is not much lower than the expected distance for random populations. On the other hand, the scatter diagram of Sko100a (Figure 5.3) has one cluster of 97 population members. As expected, its average distance is the lowest when compared with the expected average of $n - 1$.

The starting population for Nug30 does not exhibit clear clustering. The successive scatter diagrams indicate "convergence" to five clusters. Note that the problem has exactly

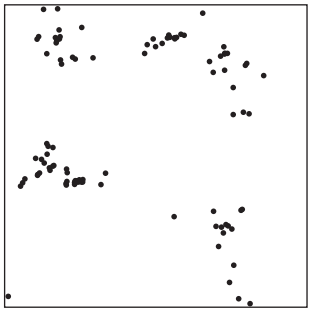


(a) $G = 0$

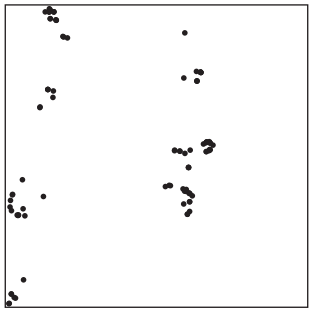
(b) $G = 10n$



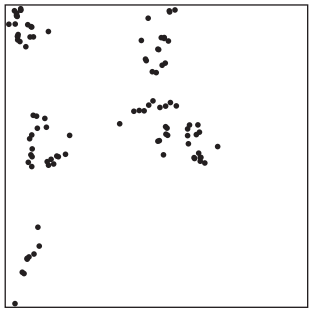
(c) $G = 20n$



(d) $G = 30n$



(e) $G = 40n$

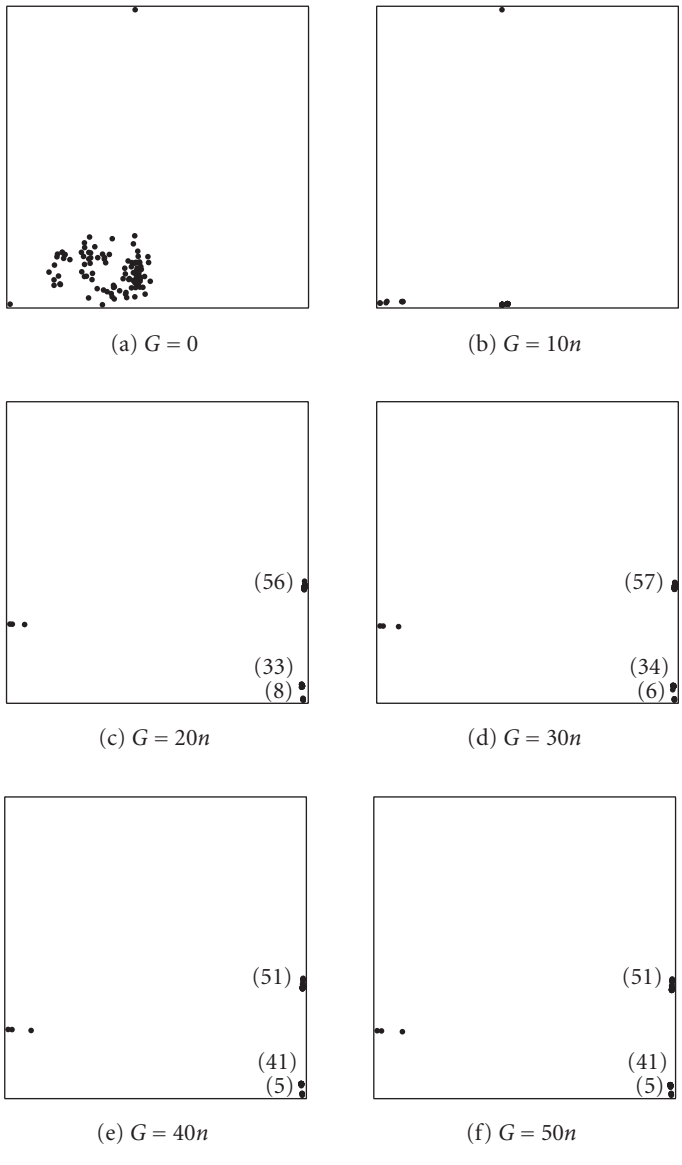


(f) $G = 50n$

Figure 5.1. The scatter plots for Nug30.

four optimum solutions that are mirror images of one another and the Hamming distance between two optimum solutions is 30. It is important to note that if the scatter diagrams are projected on the x-axis, there are only two clusters. The separation between the two clusters is best for $G = 40n$.

Different diagrams are obtained for Sko56 and Sko100a. In Figure 5.2, we observe no clusters at the starting population (with two outliers). The amorphous "cloud" is no longer observed even for $G = 10n$. A projection on the x-axis for $G = 10n$ indicates that



(a) $G = 0$ (b) $G = 10n$

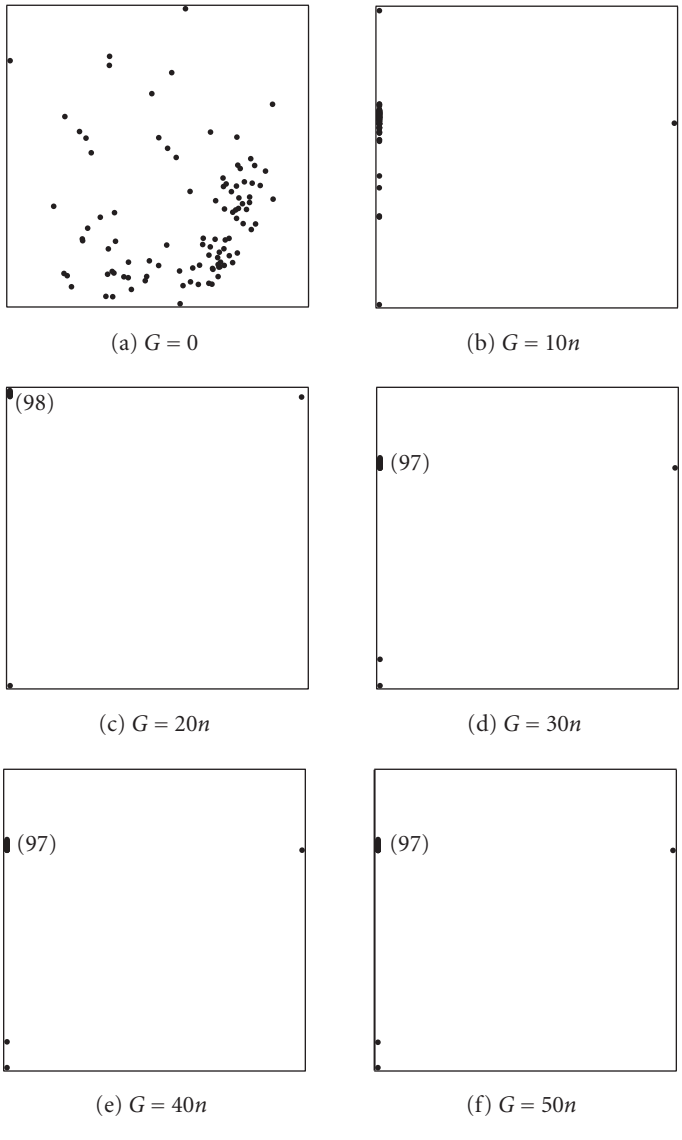
(c) $G = 20n$ (d) $G = 30n$

(e) $G = 40n$ (f) $G = 50n$

Figure 5.2. The scatter plots for Sko56.

there are two clusters. The general structure obtained for $G = 20n$ remains almost unchanged until $G = 50n$ generations are reached. The projections on the x-axis indicate two distinct clusters of 3 and 97 population members, respectively. The second cluster of 97 population members is divided into three clusters in the second dimension.

In Figure 5.3, we depict the scatter diagrams for Sko100a. The starting population does not exhibit any clusters. The projection on the x-axis indicates two clusters of 99 and 1



(a) $G = 0$ (b) $G = 10n$

(c) $G = 20n$ (d) $G = 30n$

(e) $G = 40n$ (f) $G = 50n$

Figure 5.3. The scatter plots for Sko100a.

members each starting with $G = 10n$. From $G = 30n$ and upward, a cluster of 97 with 3 outliers is evident.

5.1. Further investigation of Sko56. The scatter plot for Sko56 (Figure 5.2) shows three main clusters and three outliers. We further investigated the Sko56 problem by recreating the scatter diagram by removing the 3 outliers and running the MD procedure on the remaining 97 population members for $G = 50n$ so that the internal structure of the three

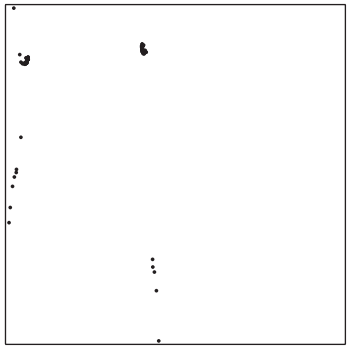


Figure 5.4. The final scatter of Sko56 without the 3 outliers.

Table 5.2. Objective function values for various clusters.

			$G = 10n$		
Cluster size	1	3	4	36	56
Average	34524.0	34497.3	34513.5	34502.4	34500.7
Minimum	34524	34472	34508	34462	34458
Maximum	34524	34512	34518	34526	34526
			$G = 20n$		
Cluster size	3	8	33	56	—
Average	34466.0	34478.0	34483.7	34481.9	—
Minimum	34462	34462	34458	34458	—
Maximum	34472	34490	34494	34494	—
			$G = 50n$		
Cluster size	3	5	41	51	—
Average	34466.0	34471.2	34478.2	34477.7	—
Minimum	34462	34462	34458	34458	—
Maximum	34472	34480	34486	34486	—

main clusters can be observed. The resulting scatter diagram is depicted in Figure 5.4. The projection on the x-axis indicates two distinct clusters. However, the clusters of 41 and 5 members appear as two clusters in the second dimension and the cluster of 51 members reveals a "core" of 42 population members and 9 members in its vicinity with 7 of them possibly defining another cluster.

Another interesting experiment is the analysis of the values of the objective functions for the different clusters. In Table 5.2, we report the statistics for the members of each cluster for $G = 10n, 20n, 50n$. The clusters are depicted in Figure 5.2. For $G = 10n$, the cluster of one is at the top of the scatter diagram. The cluster of 56 is depicted as two or three close points at the bottom-left corner, and the close-by cluster is the cluster of 36, followed by clusters of 3 and 4.

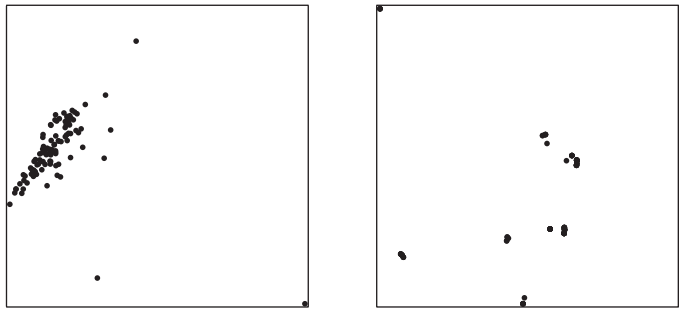
It is clear that the cluster of one has almost the worst value of the objective function in the population (34524 compared with the worst value of 34526, see Table 5.2). It is removed from the population in a few generations. It is interesting that the three outliers have the best average of the value of the objective function of all clusters. Many more iterations are required before any of the members of this cluster are removed from the population. Fortunately, the best known solution is in the bigger clusters. However, it is conceivable that the best solution could fall near the cluster of three. If so, the algorithm will miss it, because it is unlikely (probability of 0.0006 per generation) that both parents will be selected from this cluster to augment it. Once the structure of the clusters is known, one can modify the parent selection accordingly in order to generate better offspring.

5.2. Is avoiding identical population members helpful? Most genetic algorithms do not check whether newly generated offsprings are identical to existing population members before considering them for inclusion in the population. Such a provision is proposed and applied by Drezner in [15–17] with good results.

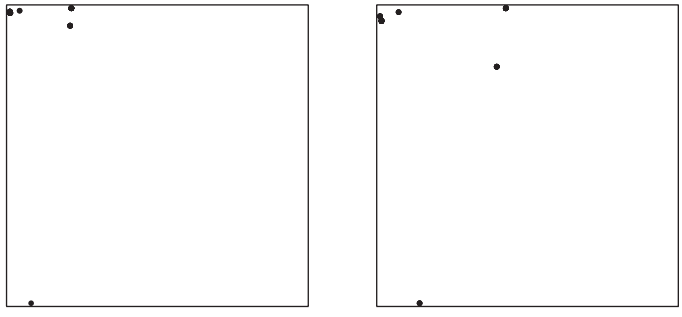
An offspring generated by two identical population members is identical to its parents (regardless of the merging process). The postmerging procedure cannot improve it (it could not further improve its parents at the time they were generated), and therefore the offspring joins the population and more identical members are added to the population. As the group of identical members increases in number, the likelihood of merging two identical parents increases. After some generations, more and more identical parents are merged and the population may consist of all identical members and no improvement is possible. We believe that the reason other genetic algorithms do not employ this provision is that researchers are under the impression that generating an identical offspring is very unlikely and one can ignore such a possibility.

The new tool proposed in this paper can be used to analyze the effect of such a provision. We ran the hybrid genetic 10 times each for Nug30, Sko56, and Sko100a with and without the provision (of not adding offspring identical to existing population members). With this provision in place, the optimum solution for Nug30 and the best known solution for Sko56 were found in all 10 runs. The best known solution for Sko100a was found 3 times out of 10 with the average solution being just 0.015% over the best known solution. The same hybrid genetic algorithm without the provision also found the optimum solution to Nug30 in all 10 runs, but found it only 4 times out of 10 for Sko56 with the average solution being only 0.008% above the best known solution. The best known solution of Sko100a was found four times out of ten but with the average solution being 0.022% over the best known solution.

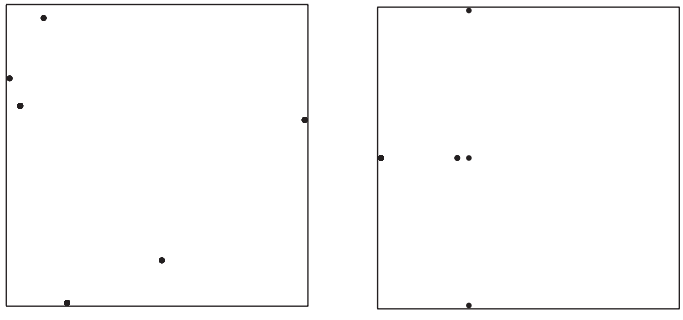
In many of these runs, the population after $50n$ generations consisted of 100 identical members. In many of these cases, all population members are inferior to the best known solution. This clearly indicates an early convergence to an inferior local minimum. In Figure 5.5 we depict the clusters for Nug30. For $G = 50n$, all population members are optimal. The average Hamming distances are 27.73 for $G = 0$, 25.32 for $G = 10n$, 19.03, 19.54, 21.55, and 3.42 for the next checkpoints, respectively. Contrary to the scatter diagrams in Figure 5.1, convergence is observed to four clear clusters, each with identical optimal members. The cluster on the left consists of 94 members, the cluster in the



(a) $G = 0$ (b) $G = 10n$



(c) $G = 20n$ (d) $G = 30n$



(e) $G = 40n$ (f) $G = 50n$

Figure 5.5. The scatter plots for Nug30 without the provision.

middle-right consists of 4 members, and the two clusters (one on top and one at the bottom) have one member each. We are "lucky" in this case that early convergence was to the optimum and not to an inferior local optimum. However, we were not that lucky in six runs for Sko56 and six runs of the Sko100a problem.

6. Conclusions

In this paper we proposed to investigate the structure of the population in genetic algorithms by applying the MD cluster-analytic-type procedure. By analyzing the results, valuable information and insight can be gained into the behavior and characteristics in the population as the genetic algorithm progresses. As an illustration, we analyzed the inclusion of the provision that an offspring identical to an existing population member is ignored rather than being added to the population. The resulting scatter plots show the early convergence of the algorithm, when this provision is not implemented. We also observed that the population becomes stagnant after about $20n$ generations and there is no need to perform $50n$ generations for these test problems.

In future research, we advocate use of this tool in order to construct better and more efficient genetic algorithms. Since the calculations involved in this procedure are very quick, the parameters controlling the genetic procedures can be modified during the progression of the genetic algorithm according to the results of such analyses. As we observed in our test problems, a stopping criterion based on the scatter diagrams can be established for genetic algorithms.

References

[1] R. K. Ahuja, J. B. Orlin, and A. Tiwari, *A greedy genetic algorithm for the quadratic assignment problem*, Computers & Operations Research **27** (2000), no. 10, 917–934.

[2] K. M. Anstreicher and N. W. Brixius, *A new bound for the quadratic assignment problem based on convex quadratic programming*, Mathematical Programming **89** (2001), no. 3, 341–357.

[3] K. M. Anstreicher, N. W. Brixius, J.-P. Goux, and J. Linderoth, *Solving large quadratic assignment problems on computational grids*, Mathematical Programming **91** (2002), no. 3, 563–588.

[4] G. C. Armour and E. S. Buffa, *A heuristic algorithm and simulation approach to relative location of facilities*, Management Science **9** (1963), 294–309.

[5] R. Battiti and G. Tecchiolli, *The reactive tabu search*, ORSA Journal on Computing **6** (1994), 126–140.

[6] J. E. Beasley, *Population heuristics*, Handbook of Applied Optimization (P. M. Pardalos and M. G. C. Resende, eds.), Oxford University Press, Oxford, 2002, pp. 138–157.

[7] R. E. Burkard, *Locations with spatial interactions: the quadratic assignment problem*, Discrete Location Theory (P. B. Mirchandani and R. L. Francis, eds.), Wiley-Intersci. Ser. Discrete Math. Optim., Wiley, New York, 1990, pp. 387–437.

[8] R. E. Burkard and F. Rendl, *A thermodynamically motivated simulation procedure for combinatorial optimization problems*, European Journal of Operational Research **17** (1984), no. 2, 169–174.

[9] E. Çela, *The Quadratic Assignment Problem: Theory and Algorithms*, Combinatorial Optimization, vol. 1, Kluwer Academic, Dordrecht, 1998.

[10] D. T. Connolly, *An improved annealing scheme for the QAP*, European Journal of Operational Research **46** (1990), no. 1, 93–100.

[11] Z. Drezner, *DISCON: a new method for the layout problem*, Operations Research **28** (1980), 1375–1384.

[12] ———, *A heuristic procedure for the layout of a large number of facilities*, Management Science **33** (1987), 909–915.

[13] ———, *Lower bounds based on linear programming for the quadratic assignment problem*, Computational Optimization & Applications **4** (1995), no. 2, 159–165.

[14] ———, *Heuristic algorithms for the solution of the quadratic assignment problem*, Journal of Applied Mathematics and Decision Sciences **6** (2002), no. 3, 163–173.

[15] ———, *A new genetic algorithm for the quadratic assignment problem*, INFORMS Journal on Computing **15** (2003), no. 3, 320–330.
[16] ———, *Compounded genetic algorithms for the quadratic assignment problem*, Operations Research Letters **33** (2005), no. 5, 475–480.
[17] ———, *The extended concentric tabu for the quadratic assignment problem*, European Journal of Operational Research **160** (2005), no. 2, 416–422.
[18] C. Fleurent and J. Ferland, *Genetic hybrids for the quadratic assignment problem*, Quadratic Assignment and Related Problems (New Brunswick, NJ, 1993), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 16, American Mathematical Society, Rhode Island, 1994, pp. 173–187.
[19] L. M. Gambardella, E. D. Taillard, and M. Dorigo, *Ant colonies for the quadratic assignment problem*, Journal of the Operational Research Society **50** (1999), no. 2, 167–176.
[20] P. C. Gilmore, *Optimal and suboptimal algorithms for the quadratic assignment problem*, Journal of the Society of Industrial and Applied Mathematics **10** (1962), no. 2, 305–313.
[21] P. M. Hahn and T. L. Grant, *Lower bounds for the quadratic assignment problem based upon a dual formulation*, Operations Research **46** (1998), no. 6, 912–922.
[22] P. M. Hahn, W. L. Hightower, W. P. Adams, and M. Guignard-Spielberg, *A level-2 reformulation-linearization technique bound for the quadratic assignment problem*, to appear in European Journal of Operational Research.
[23] P. M. Hahn and J. Krarup, *A hospital facility problem finally solved*, Journal of Intelligent Manufacturing **12** (2001), no. 5-6, 487–496.
[24] E. L. Lawler, *The quadratic assignment problem*, Management Science **9** (1963), 586–599.
[25] Y. Li, P. M. Pardalos, and M. G. C. Resende, *A greedy randomized adaptive search procedure for the quadratic assignment problem*, Quadratic Assignment and Related Problems (P. M. Pardalos and H. Wolkowicz, eds.), DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 16, American Mathematical Society, Rhode Island, 1994, pp. 237–261.
[26] G. A. Marcoulides and Z. Drezner, *A procedure for transforming points in multi-dimensional space to two-dimensional*, Educational and Psychological Measurement **53** (1993), 933–940.
[27] ———, *A procedure for detecting pattern clustering in measurement designs*, Objective Measurement: Theory into Practice (M. Wilson, K. Draney, and G. Engelhard Jr., eds.), vol. 5, Ablex, New Jersey, 2000, pp. 287–302.
[28] P. Moscato, *Memetic algorithms*, Handbook of Applied Optimization (P. M. Pardalos and M. G. C. Resende, eds.), Oxford University Press, Oxford, 2002.
[29] C. E. Nugent, T. E. Vollman, and J. Ruml, *An experimental comparison of techniques for the assignment of facilities to locations*, Operations Research **16** (1968), 150–173.
[30] M. Nystrom, *Solving certain large instances of the quadratic assignment problem: Steinberg's examples*, Working paper, California Institute of Technology, California, 1999.
[31] F. Rendl, *The quadratic assignment problem*, Facility Location: Applications and Theory (Z. Drezner and H. Hamacher, eds.), Springer, Berlin, 2002, pp. 439–457.
[32] M. G. C. Resende, K. G. Ramakrishnan, and Z. Drezner, *Computing lower bounds for the quadratic assignment problem with an interior point algorithm for linear programming*, Operations Research **43** (1995), no. 5, 781–791.
[33] S. Salhi, *Heuristic search methods*, Modern Methods for Business Research (G. A. Marcoulides, ed.), Lawrence Erlbaum Associates, New Jersey, 1998.
[34] J. Skorin-Kapov, *Tabu search applied to the quadratic assignment problem*, ORSA Journal on Computing **2** (1990), no. 1, 33–45.
[35] E. D. Taillard, *Robust taboo search for the quadratic assignment problem*, Parallel Computing **17** (1991), no. 4-5, 443–455.
[36] ———, *Comparison of iterative searches for the quadratic assignment problem*, Location Science **3** (1995), no. 2, 87–105.

[37] D. M. Tate and A. E. Smith, *A genetic approach to the quadratic assignment problem*, Computers & Operations Research **22** (1995), no. 1, 73–83.
[38] M. R. Wilhelm and T. L. Ward, *Solving quadratic assignment problems by simulated annealing*, IIE Transactions **19** (1987), 107–119.

Zvi Drezner: College of Business and Economics, California State University-Fullerton, Fullerton, CA 92834, USA
E-mail address: zdrezner@fullerton.edu

George A. Marcoulides: College of Business and Economics, California State University-Fullerton, Fullerton, CA 92834, USA
E-mail address: gmarcoulides@fullerton.edu

A MEASURE OF THE VARIABILITY OF REVENUE IN AUCTIONS: A LOOK AT THE REVENUE EQUIVALENCE THEOREM

FERNANDO BELTRÁN AND NATALIA SANTAMARÍA

Received 30 August 2005; Revised 6 June 2006; Accepted 7 June 2006

One not-so-intuitive result in auction theory is the revenue equivalence theorem, which states that as long as an auction complies with some conditions, it will on average generate the same revenue to an auctioneer as the revenue generated by any other auction that complies with them. Surprisingly, the conditions are not defined on the payment rules to the bidders but on the fact that the bidders do not bid below a reserve value—set by the auctioneer—the winner is the one with the highest bidding and there is a common equilibrium bidding function used by all bidders. In this paper, we verify such result using extensive simulation of a broad range of auctions and focus on the variability or fluctuations of the results around the average. Such fluctuations are observed and measured in two dimensions for each type of auction: as the number of auctions grows and as the number of bidders increases.

1. Introduction

In the early 1980s, a series of papers appeared in the economics literature on auctions, dealing specially with the issue of the expected revenue to an auctioneer in a single-object buyer's auction. The pioneer work of Vickrey offered the first insights into the expected revenues of four different auctions finding them to be equivalent (Milgrom [4]). The main result, appearing in [6] by Riley and Samuelson, and Myerson [5] became known as the revenue equivalence theorem. The theorem states that as long as an auction complies with some conditions, it will on average generate the same revenue to an auctioneer as the revenue generated by any other auction with the same conditions. Surprisingly the conditions are not defined on the payment rules but on the facts that bidders do not bid below a reserve value—defined by the auctioneer—the winner is the one with the highest bid and there is a common equilibrium bidding function used by all bidders.

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 27417, Pages 1–14
DOI 10.1155/JAMDS/2006/27417

More specifically, as Klemperer [3] puts it: "each of a given number of risk-neutral potential buyers of an object has a privately known signal independently drawn from a common, strictly increasing, atomless distribution. Then any auction mechanism in which

(i) the object always goes to the buyer with the highest signal, and
(ii) any bidder with the lowest-feasible signal expects zero surplus

yields the same expected revenue (and results in each bidder making the same expected payment as a function of her signal)."

The result applies both to private-value models—every player's value is independently drawn from the same continuous distribution on a finite interval—and to more general common-value models—the value of the object is the same for all players, but it is unknown at the time of the bidding—provided that bidders' signals are independent.

2. Is an auctioneer interested in the variability of the mean revenue?

The revenue equivalence theorem has been a remarkable piece in the construction of a theory of auctions. Under the stated conditions, such seemingly different auctions as the all-pay or the second-price sealed-bid yield the same expected revenue. As Milgrom [4] affirms, one practical use of the revenue equivalence theorem is as a benchmark for the analyses of revenues in auctions, when the assumptions of the theorem do not hold or cannot be verified properly.

A main concern to be addressed in this paper is that of an auctioneer trying to decide which auction to use. Suppose an auctioneer has an object to sell. If he knew that such an object represented a private value to all potential bidders, bidders values were independent and any bid placed for the auction was larger than a reserve value—which would happen in the case of at least one bidder informed about such price and willing to participate in the auction—then the auctioneer should be indifferent among several different auctions he could choose from. For instance, he could use a first-price sealed-bid auction or a "sad losers" auction (Riley and Samuelson [6]). The latter is an auction in which every bidder, except the winner, pays his/her bid. There could, however, be a very practical concern that the auctioneer needs to deal with: the revenue equivalence theorem states its result in terms of the expected revenue to the seller but the seller not always likes or needs to run a large number of auctions of the same object—or type of object. Maybe, what is being sold is not ordinary merchandise but a right for the exploitation of a public good. Assuming the auctioneer will award the object to the highest bidder, would the design of the auction—that is, the payment from the bidders—matter to the auctioneer? The theorem would ease the auctioneer's worries with a categorical "it would not." Well, "it would not" if the auctioneer ran a sufficient number of auctions so that on average his revenue from each auction was the one predicted by the theorem.

If the auctioneer is not running many auctions or if he is just auctioning one object, his attention may shift to find a measure of the variability of such average or mean value. For instance, in [7] by Waehrer et al., it is shown that a risk-averse auctioneer prefers a first-price auction to a second-price auction, and in turn he prefers a second-price auction to an English auction. In this paper, we use a simulator to better understand how large around the mean are the variations of running several auctions for at least six different

auctions, which under the assumptions of the theorem should yield the same (expected) revenue.

In this paper, firstly, we verify the results of the theorem running simulations of a broad range of auctions and, secondly, we focus on the variability or fluctuation about the average revenue of an auction with a given number of bidders, we attempt to find a criterion that helps the auctioneer to decide about the type of auction to be used. The fluctuations are observed and measured in two dimensions for each type of auction: as the number of instances of a given auction grows and as the number of bidders in the auction increases.

3. The revenue equivalence theorem

Theorem 3.1 (Klemperer [3]). *In an auction of a single object, suppose there are n risk-neutral potential bidders with privately known independent signals drawn from a common distribution $F(v)$. Then any auction mechanism in which* (i) *the object always goes to the buyer with the highest signal, and* (ii) *any bidder with the lowest-feasible signal expects zero surplus yields the same expected revenue.*

For the proof, see Klemperer [3, Chapter 1, page 17].

4. Optimal bids

In all auctions considered here, the winner is the bidder with the highest bid; ties are broken randomly. In an *all-pay auction*, every bidder pays his/her bid; in *sad losers auction*, all but the winner pay their bids; in *last-pays auction*, only the bidder with the lowest bid pays. In *Santa Claus auction*, the auctioneer takes the payment from the winner and gives back a portion of it to all bidders, including the winner (Riley and Samuelson [6]). First-price and second-price are the so-called traditional auctions where the amount paid by the winner is the highest bid or the second highest bid, respectively.

We have used the result above to calculate the optimal bids in several auctions which comply with the conditions of the theorem. Starting with basic results for two bidders presented in [6] by Riley and Samuelson, we previously calculated (Beltrán et al. [1]) the optimal bids for n bidders in *all-pay, sad losers, last-pay*s and *Santa Claus*. To the latter, we have added the first-price auction, whose optimal bid expression is found in [2] by Gibbons, and the second-price auction where it is optimal for a bidder to bid his true value (Klemperer [3]). Optimal bid functions for n users in the auctions mentioned can be found in Appendix A.

5. Simulating the auctions

In order to perform the computational experiments, we used a random number generator to determine the bidders' valuations; the valuations are uniformly drawn from the interval $[0,1]$. Every run consists of a number of auctions or scenarios of the auction, for a predefined number of bidders; the bids are calculated according to the optimal bid functions obtained in the preceding section. The simulator determines the optimal allocation and the revenue for the seller, repeating this procedure until the number of desired scenarios is completed. The runs are conducted while varying the number of bidders and

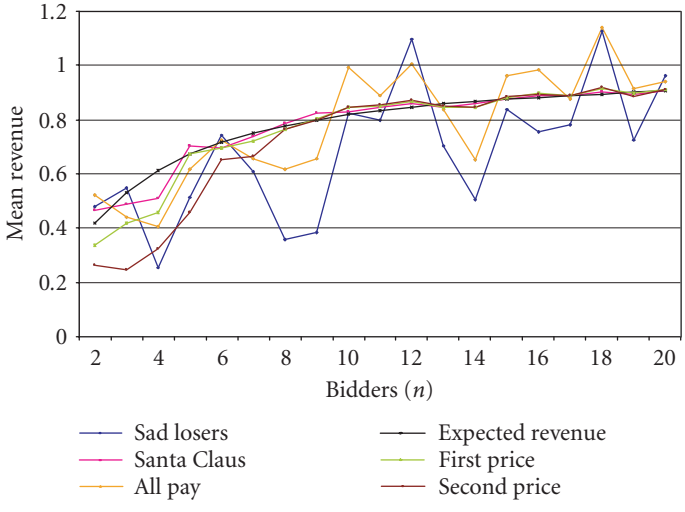


Figure 5.1. Mean revenue for 5 auction types.

Figure 5.2. Coefficient of variation for 5 auction types.

the number of scenarios. For each run, we calculated the mean revenue for the auctioneer and the coefficient of variation of the revenue, defined as the ratio of the variance with respect to the square of the expected value. This was done for each of six auctions: *first-price, second-price, all-pay, Santa Claus, sad losers*, and *last-pays*. The analysis that follows uses data from all auctions except last-pays, which because of its particular design deserves a special analysis in a subsequent section. Figures 5.1 and 5.2 show, for up to 20 bidders, the average and coefficient of variation for 20 scenarios.

It can be observed that as the number of bidders increases, the mean revenue approaches the theoretical expected revenue and the variation around the mean decreases for most of the auction types. However, this variation is significantly different for *sad losers* and *all-pay*. This is also confirmed if the number of scenarios is increased. Appendix B illustrates this fact, where simulations results are reported in which 50 and 100 scenarios were performed for auctions with up to 20 bidders.

In the traditional auction formats (*first-price* and *second-price auctions*), fluctuations around the mean revenue are less than those of the other auctions, except for the *Santa Claus auction*. By the central limit theorem, increasing the number of scenarios, the degree of variability around the mean revenue decreases. Runs with 100 and 500 scenarios were also done. Those results show that as a function of the number of bidders, the coefficient of variation converges to zero for *first-price, second-price*, and *Santa Claus*, and seems to settle around 0.65 for *all-pay*. However the variability of *sad losers* remains high when compared to the others and does not seem to converge to any value.

6. A real experiment

Our previous experimentation with auctions in a broader setting has included the development of SUBASTIN (http://subastas.uniandes.edu.co), a web application for the administration of auctions over the Internet. SUBASTIN collects the bids from the players and determines the winner in a fairly large family of auctions (SUBASTIN administers all the auctions described in this paper plus several dynamic auctions such as ascending English, descending Dutch, German, and simultaneous ascending auctions. SUBASTIN is also capable of administering single-bid combinatorial auctions). Using SUBASTIN, we ran a real *all-pay auction* where bidders were students of a Game Theory Class (Universidad de Los Andes, Departamento de Ingeniería Industrial, Game Theory Course, January–May 2004). When bidding to get the object being auctioned, the bidders used their SUBASTIN Web windows; the results are summarized in Table 6.1. (Bids are stated in Colombian pesos (COP). In April 2004, the exchange rate was US $1 = COP $2700. This illustrates that the object auctioned did not mean a high expense to any bidder.)

The market value of the auctioned object was about $15000. So, the auctioneer was not only able to recover the cost of purchasing the object, but also able to make quite a bit of a profit. It is clear that at least three bidders were not interested at all; some others bid a very low value. It is tempting to say that each of these bidders thought of winning the auction expecting others to bid low as well. Perhaps they disliked the idea that the auctioneer could profit excessively. However, quite a few bid high, even close to the market value. This behavior contrasts the behavior of those who bid low.

This experiment is just a sample of what could happen in a nontraditional auction, even though such type is one that satisfies the assumption of the theorem, at least in regard to who wins the auction and the seeming independence of the bidders' valuations. Simulations of *all-pay* show a larger variability of the expected revenue than that of *first-price, second-price*, and *Santa Claus auctions*. The results from the experiment shed some light on the possibility that an auctioneer prefers using one auction over other.

Table 6.1. Bids in a real all-pay auction.

Bidder ID	Bid
El Coyote	0
Ricky Ricon	1000
Carmedelgad3	10000
Juangalind	12000
Andrevasque1	10000
Andresantac	500
Diego Martin	100
Diegodiazm	1000
Rubenjacome	100
Javieguarin	0
Florbetanc	100
Mauriescoba	0
Ricarpedraz	15000
Paulabarrie	20000
J2zp	13000
Francovoyageur	14000
Sebassalaza	100
Maurisuarez	2000
Juanredond	1000
El Mani	500

7. Some experimental difficulties of last-pays

Results shown in Appendix B for simulation runs of *last-pays* are not quite encouraging. Expected revenue in auctions in which a few bidders are simulated is close to the theoretical value calculated in Appendix A. However, results no longer seem to hold as long as more bidders are included.

In *last-pays*, the auctioneer has positive revenue only if the valuation for all the bidders is greater than the auctioneer's reserve value. If at least one bidder's valuation is less than the reserve value, the revenue for the auctioneer will be zero as such a bidder is the one who should be paying. The probability that all valuations are greater than the reserve value decreases when the number of bidders increases; this also increases the probability that the auctioneer's revenue is zero. The results of the simulation runs performed on last-pays show that when the number of bidders increases, the expected revenue for the auctioneer goes to zero. Appendix B shows the difference in expected revenue obtained when a 5000-scenario simulation is compared to a 50000-scenario simulation.

8. Conclusions

For each run, that is, an auction type simulated several times with a given number n of users, we have found the expected revenue to the auctioneer and a measure of the variability of such result using its coefficient of variation. When the number of bidders is fixed, we have then compared such measure across several auction types. If an auctioneer

does not have the time or the need to run a large number of auctions, would the result provided by the theorem influence his decision as to which auction to use? If he is interested in maximizing his revenue, *all-pay* or *sad losers* seem to provide some greater degree of variability of the expected revenue. From the results, we can argue that an auctioneer seeking to improve his revenue may prefer one auction to another, if he is willing to bear the risk implied in the variance of the revenue.

For the real auction we performed, if we believed that the assumptions of the theorem held, in particular, that the students' signals were independent, then we might assert that the auctioneer could have used a *first-price* or *second-price auction* instead of the *all-pay auction*. In the context of the main result of the theorem, we would have expected the same revenue for the auctioneer without worrying about the type of auction administered. However, as the results of simulations showed, the variability of the revenue is quite different in *all-pay* when compared to the more traditional *first-price* and *second price*. It is in this sense that the result from the real experiment becomes relevant to the inquiry about the auctioneer's question posed at the beginning and the risk he incurs when answering such a question.

Appendices

A. Bid functions (Beltrán et al. [1])

Let π represent the bidder expected revenue, v the bidder's value, b the bid function, and $F(v)$ the distribution of the bidder's value.

Optimal bidding function in *Santa Claus* auction with n bidders is

$$\pi = F^{n-1}(b)\cdot(v-b)+\int_{v*}^{b} F^{n-1}(v)dv, \tag{A.1}$$

$$\pi = b^{n-1}v - b^n + \frac{b^n}{n} - \frac{(v^*)^n}{n}, \tag{A.2}$$

$$\frac{\partial \pi}{\partial b} = (n-1)\cdot b^{n-2}\cdot v - n\cdot b^{n-1} + b^{n-1} = 0, \tag{A.3}$$

$$b^{n-2}(v-b) = 0, \tag{A.4}$$

$$b = v. \tag{A.5}$$

Santa Claus' gift to every bidder in a *Santa Claus auction* is

$$S(b) = \int_{v*}^{b} F^{n-1}(v)dv = \frac{b^n}{n} - \frac{(v^*)^n}{n} = \frac{b^n - (v^*)^n}{n}. \tag{A.6}$$

Optimal bid function in *all-pay auction* with n bidders is

$$\begin{aligned} b(v) &= vF^{n-1}(v) - \int_{v*}^{v} F^{n-1}(x)dx, \\ b(v) &= v\cdot v^{n-1} - \int_{v*}^{v} x^{n-1}dx, \\ b(v) &= v^n - \frac{v^n}{n} + \frac{(v^*)^n}{n}. \end{aligned} \tag{A.7}$$

Optimal bidding function in *sad losers auction* with n bidders is as follows.

(i) Any bidder's expected revenue is $\pi = F^{n-1}(x) \cdot v - (1 - F^{n-1}(x)) \cdot b(x) - c$.

(ii) Let $H(x) = (1 - F^{n-1}(x)) \cdot b(x)$.

(iii) Revenue is maximized when $\partial\pi/\partial x = v \cdot (\partial/\partial x)F^{n-1}(x) - H'(x) = 0$.

(iv) $H'(x) = v \cdot (\partial/\partial x)F^{n-1}(x)$.

At equilibrium, a bidder bids its valuation, that is, $x = v$,

$$\begin{aligned} H'(v) &= v \cdot \frac{\partial}{\partial v} F^{n-1}(v), \\ H(v) &= \int_{v*}^{v} x \cdot d\left(F^{n-1}(x)\right) + k(\text{A.1}). \end{aligned} \tag{A.8}$$

Applying the expression of the expected revenue to v^*, then $\pi_{v*} = v^* \cdot F^{n-1}(v^*) - H(v^*) - c^* = 0$, and so $H(v^*) = v^* \cdot F^{n-1}(v^*) - c^*$ (A.2). Equating (A.1) and (A.2),

$$\begin{aligned} k &= v^* \cdot F^{n-1}\left(v^*\right) - c^*, \\ H(v) &= \int_{v*}^{v} x \cdot d\left(F^{n-1}(x)\right) + v^* \cdot F^{n-1}\left(v^*\right) - c^* \\ &= vF^{n-1}(v) - \int_{v*}^{v} F^{n-1}(x)dx - c^*. \end{aligned} \tag{A.9}$$

If $F(x) \approx U[0,1]$, then

$$H(v) = \frac{1}{n}\left((n-1)v^n + \left(v^*\right)^n - nc^*\right). \tag{A.10}$$

Using the original definition of $H(v)$,

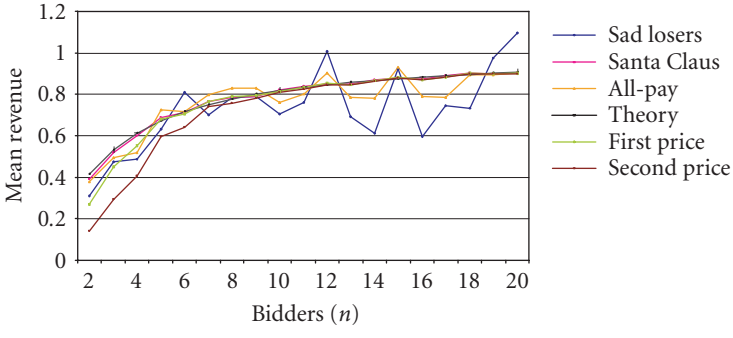
$$b(v) = \frac{H(v)}{1 - v^{n-1}} = \frac{(n-1)v^n + \left(v^*\right)^n - nc^*}{n\left(1 - v^{n-1}\right)}. \tag{A.11}$$

Replacing $c^* = v^* F^{n-1}\left(v^*\right) = \left(v^*\right)^n$,

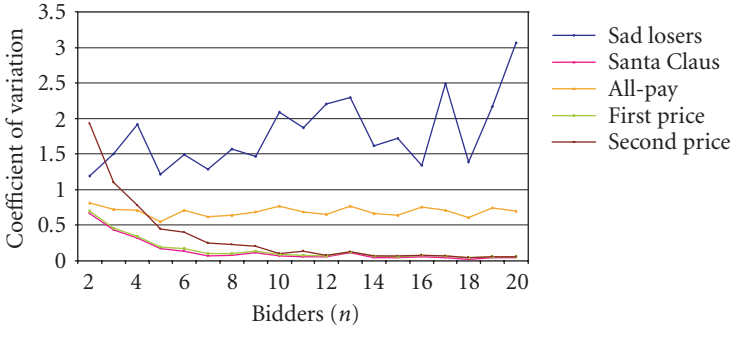
$$b(v) = \frac{(n-1)\left(v^n - \left(v^*\right)^n\right)}{n\left(1 - v^{n-1}\right)}. \tag{A.12}$$

Optimal bid function in *last-pays auction* with n bidders is

$$b(v) = \frac{vF^{n-1}(v) - \int_{v*}^{v} F^{n-1}(x)dx}{\left(1 - F(v)\right)^{n-1}}, \qquad b(v) = \frac{v \cdot v^{n-1} - \int_{v*}^{v} x^{n-1}dx}{(1-v)^{n-1}},$$
$$b(v) = \frac{\left(v^*\right)^n + (n-1)v^n}{n(1-v)^{n-1}}. \tag{A.13}$$



(a)



(b)

Figure B.1. Results for 20 bidders and 100 scenarios.

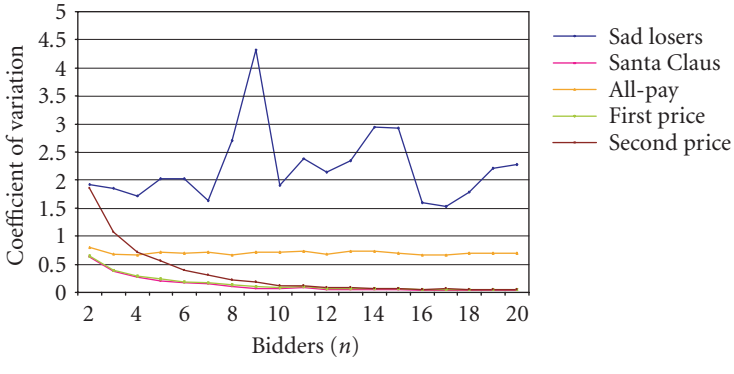
Auctioneer's expected revenue is

$$\begin{aligned}
\pi &= n\int_{v*}^{\overline{v}} \left(vF'(v) + F(v) - 1\right) \cdot F(v)^{n-1} dv, \\
\pi &= n\int_{v*}^{\overline{v}} \left(v + v - 1\right) \cdot v^{n-1} dv, \\
\pi &= n\left[\frac{2v^{n+1}}{n+1} - \frac{v^n}{n}\right]_{v*}^{\overline{v}}, \\
\pi &= \frac{2n\left((\overline{v})^{n+1} - (v^*)^{n+1}\right) - (n+1)\left((\overline{v})^n - (v^*)^n\right)}{n+1}.
\end{aligned} \tag{A.14}$$

B. Simulation results

Figures B.1 to B.11 show results for different numbers of bidders and scenarios.

(a)



(b)

Figure B.2. Results for 20 bidders and 500 scenarios.

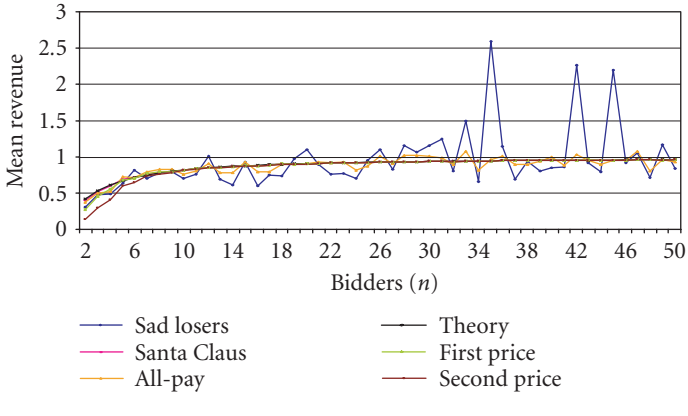


Figure B.3. Results for 50 bidders and 100 scenarios.

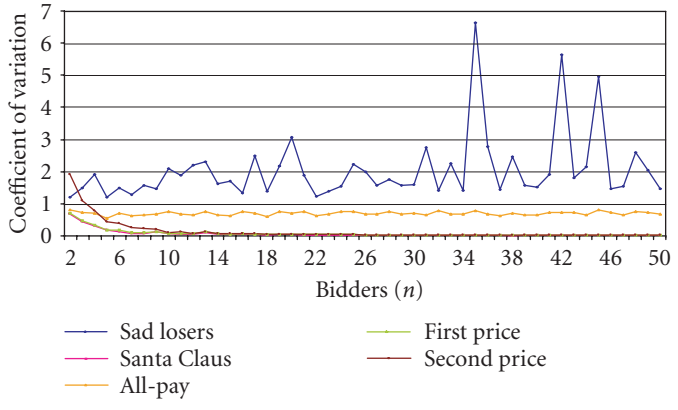


Figure B.4. Results for 50 bidders and 100 scenarios.

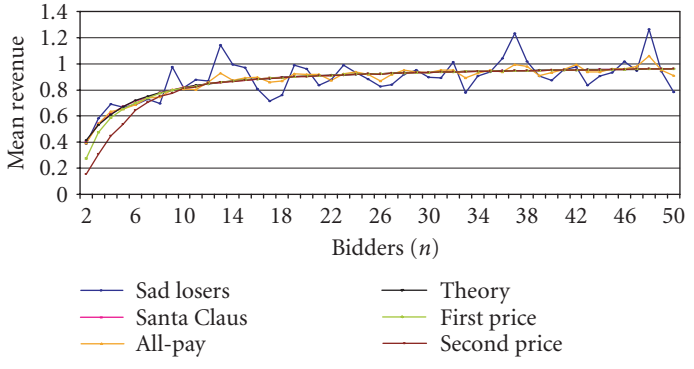


Figure B.5. Results for 50 bidders and 500 scenarios.

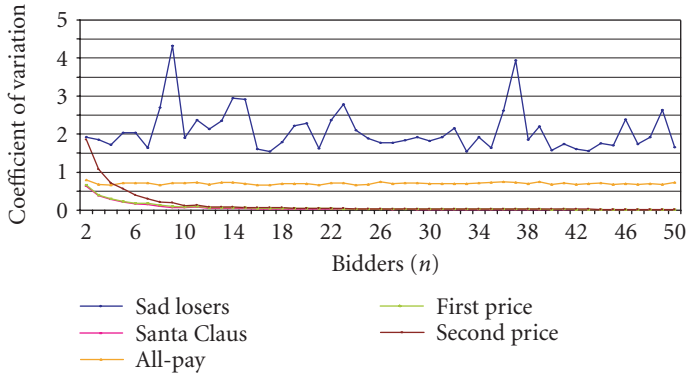


Figure B.6. Results for 50 bidders and 500 scenarios.

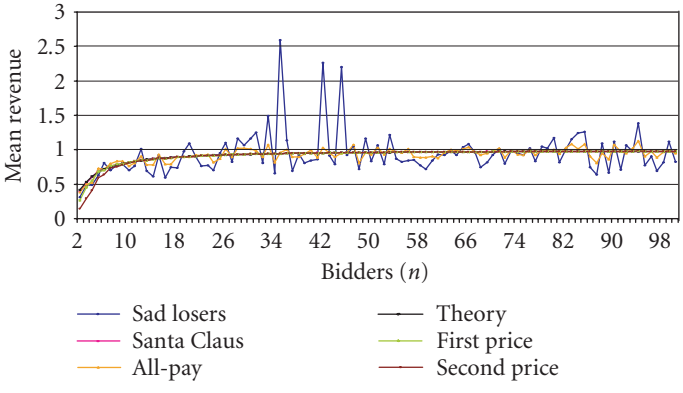


Figure B.7. Results for 100 bidders and 100 scenarios.

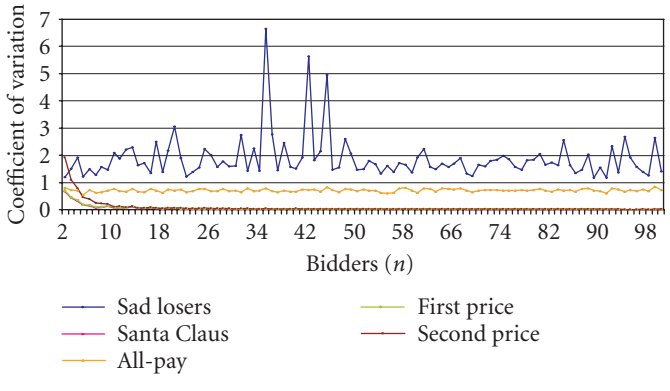


Figure B.8. Results for 100 bidders and 100 scenarios.

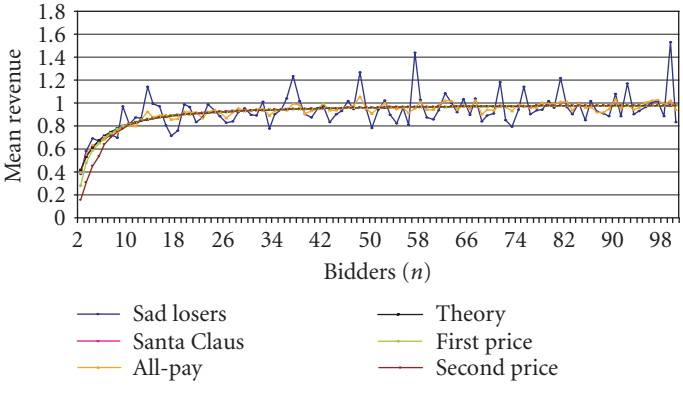


Figure B.9. Results for 100 bidders and 500 scenarios.

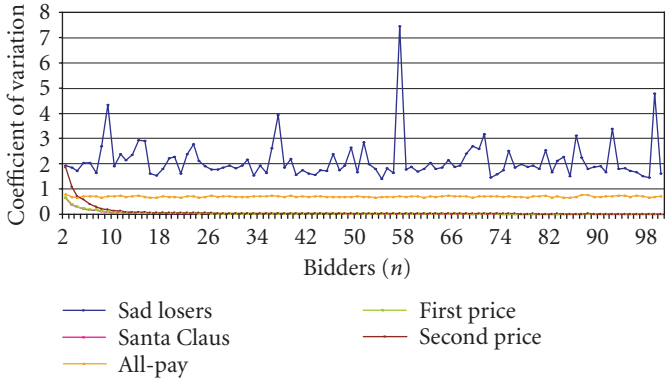


Figure B.10. Results for 100 bidders and 500 scenarios.

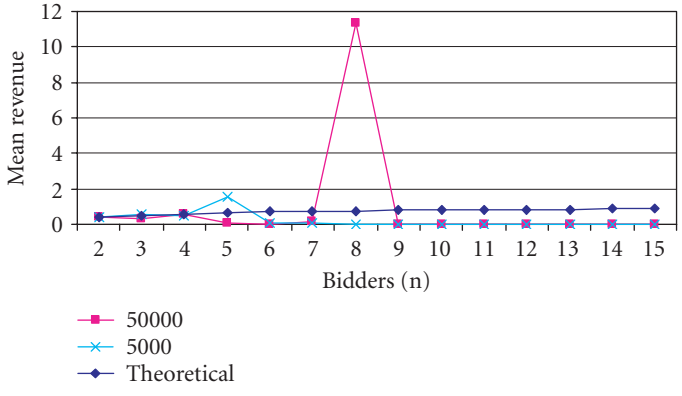


Figure B.11. Last-pay: 15 bidders, results from 5000 and 50000 scenarios.

Acknowledgments

A first version of this paper was presented at the 39th Annual Meeting, Operations Research Society of New Zealand, Auckland, November 2004. The authors want to especially thank the participation of Pamela Cardozo and Camilo Restrepo in a previous work that led to the ideas presented here. Camilo Restrepo provided an interesting explanation, based on his experimentation, of the strange behavior of the last-pays auction.

References

[1] F. Beltrán, N. Santamaría, C. Restrepo, and P. Cardozo, *El teorema de ingreso equivalente para subastas de un objeto: Aproximación experimental*, Revista de Ingeniería, Universidad de Los Andes. Marzo **17** (2003), 12–18.

[2] R. Gibbons, *Game Theory for Applied Economists*, Princeton University Press, New Jersey, 1992.

[3] P. Klemperer, *Auctions: Theory and Practice*, Princeton University Press, New Jersey, 2004.

[4] P. Milgrom, *Putting Auction Theory to Work*, Cambridge University Press, Cambridge, 2004.
[5] R. B. Myerson, *Optimal auction design*, Mathematics of Operations Research **6** (1981), no. 1, 58–73.
[6] J. G. Riley and W. F. Samuelson, *Optimal auctions*, American Economic Review **71** (1981), 381–392.
[7] K. Waehrer, R. M. Harstad, and M. H. Rothkopf, *Auction form preferences of risk-averse bid takers*, RAND Journal of Economics **29** (1998), no. 1, 179–192.

Fernando Beltrán: Information Systems and Operations Management Department,
University of Auckland Business School, Auckland 1142, New Zealand
E-mail address: f.beltran@auckland.ac.nz

Natalia Santamaría: Departamento de Ingeniería Industrial, Universidad de los Andes,
P.O. Box 4976, Bogotá, Colombia
E-mail address: n-santam@uniandes.edu.co
Current address: Rutgers Center for Operations Research (RUTCOR), Rutgers University,
Piscataway, NJ 08854-8003, USA
E-mail address: ntobar@rutcor.rutgers.edu

A SIMULATION FRAMEWORK FOR NETWORKED QUEUE MODELS: ANALYSIS OF QUEUE BOUNDS IN A G/G/c SUPPLY CHAIN

MAHYAR AMOUZEGAR AND KHOSROW MOSHIRVAZIRI

Received 4 April 2006; Accepted 18 May 2006

Some limited analytical derivation for networked queue models has been proposed in the literature, but their solutions are often of a great mathematical challenge. To overcome such limitations, simulation tools that can deal with general networked queue topology must be developed. Despite certain limitations, simulation algorithms provide a mechanism to obtain insight and good numerical approximation to parameters of networked queues. This paper presents a closed stochastic simulation network model and several approximation and bounding schemes for G/G/c systems. The analysis was originally conducted to verify the integrity of simulation models used to develop alternative policy options conducted on behalf of the US Air Force. We showed that the theoretical bounds could be used to approximate mean capacities at various queues. In this paper, we present results for a G/G/8 system though similar results have been obtained for other networks of queues as well.

1. Introduction

In this paper we consider a closed stochastic simulation system model used in the analysis of aircraft engines maintenance and repair options. In this analysis, we evaluated the cost and benefits of centralized maintenance versus a decentralized option. This analysis was prompted by the ongoing reorganization of the Air Force into an Air and Space Expeditionary Force (AEF). The main objective of this reorganization is to replace the forward presence of air power with a force that can deploy quickly from the continental United States (CONUS) in response to a crisis, commence operations immediately upon arrival, and sustain those operations as needed. To support the expeditionary force, support processes such as munitions, fuels, and maintenance also need to be transformed.

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 87514, Pages 1–13
DOI 10.1155/JAMDS/2006/87514

AEF requires a combat support system capable of supporting an expanded range of operations from humanitarian and disaster relief to major combat and peacekeeping operations, which could take place in any of a number of different locations.

One of the critical processes for the Air Force is the intermediate maintenance for jet engines. This so-called intermediate maintenance facility (IMF) consists of several components, including the maintenance (repair and service) shop, the module shop, and the assembly and test cell. IMF is one of three levels of maintenance used by the Air Force to repair jet engines, especially those powering fighter aircraft.

(i) Flightline maintenance consists mostly of inspections, diagnostics, engine removals, and some quick repairs that do not involve engine teardown.

(ii) Service at IMF includes disassembly of the engines; substantial repairs to parts such as fans, low pressure turbines, and afterburners; and engine test cell runs.

(iii) Depot maintenance involves the complete teardown and refurbishment of any repairable part in an engine. The rebuilding of an engine at the depot allows the engine's use of parameters (flight time, cycles, etc.) effectively to be reset at zero.

Traditionally, the IMF has been located at the operating base with the aircraft. This policy was reinforced by the planning for major wars in Europe and Korea: a unit would be moved to existing bases in theater in preparation for immediate action and could expect little resupply during the first few weeks of combat. Under traditional planning for wing deployment, therefore, the IMF is prepared to move along with the rest of the wing support, although not with the combat units themselves, who will use spares to replace engines until the IMF arrives and is up and running.

1.1. Current practices and trends. In recent years, the question of whether or not IMF operations should be centralized has been the subject of frequent discussion in the engine community. Many factors have favored centralization, including the increased complexity of engines and the large investment required for repair facilities. Other factors have mitigated against centralization, particularly the fact that, unlike other commodities such as avionics components, engines are heavy and bulky and thus require special packing to ship. Over the years, the Air Force has experienced a pattern of alternation between the partial centralization of the maintenance operations—in certain regions and for certain engine types—and the subsequent restoration of IMF to operating units. The requirements associated with expeditionary operations—including the ability to move quickly and the need to keep initial transportation requirements down—have raised new questions about the policy of locating the IMF at the operating base. This research aims to provide insights into this issue by determining whether engine maintenance support can best be provided from decentralized shops at the supported bases or from a centralized, off-base facility.

The operation and maintenance of engines comprise the sequence of events illustrated as an aggregate in Figure 1.1: planes fly (sorties) from main bases and remote operating locations across the globe to meet training and other requirements. After each sortie, aircraft engines are inspected on the flight line, and, depending on the accumulated flying hours and other factors, are given minor maintenance. Engines may also be removed from aircraft and sent to an intermediate maintenance facility for major service and repair. At

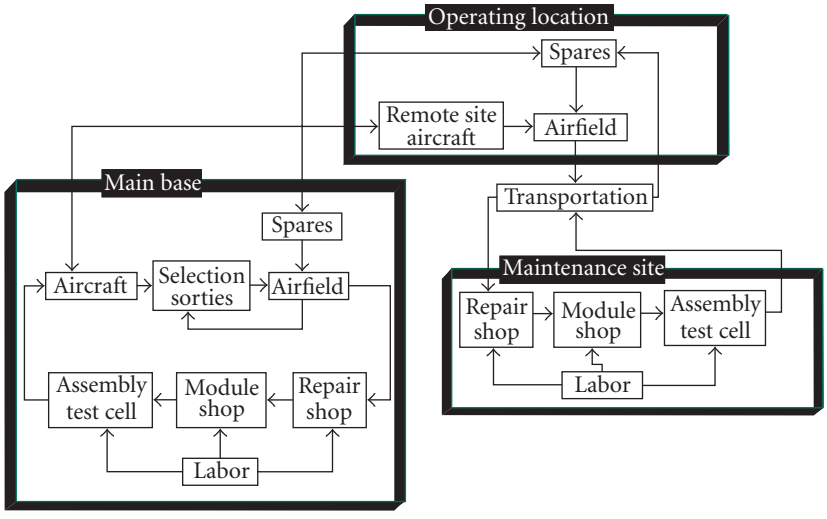


Figure 1.1. Operation and maintenance sequence.

this facility the engines are inspected, repaired, tested, and then returned to the flight line as serviceable spares. At each operating site there is a cache of serviceable spares to replace engines sent to IMF. However, there is only a limited inventory of such spares; there may be time where aircraft are grounded due to the engines availability. The ultimate goal is to increase the efficiency of the maintenance process while keeping the least number of spares as possible.

1.2. Model formation. The nature of this problem has lent itself to a closed loop networks of multiple servers/queues, some sequential and others parallel (over forty queues and servers). A queueing system is said to be closed if the servicing facility processes only a given group of permanent customers. When a customer needs service, it joins the queue and it is either served based on FIFO discipline or is given priority if it meets a certain criteria (e.g., a particular engine is required in the field faster than other type). The demand for service and duration of service depends on many variables and for this study we used historical data to compute the arrival and departure rates. The complexity of this problem led to a queueing model that could only be described with general arrival and service times or a G/G/c/n queueing system where n, the restriction on system capacity, varied depending on the process. G/G/c queue and its related families, M/G/c, G/G/1, are too complex to analyze mathematically and there are very few closed-formed results about such systems. However, several quite useful approximate and bounding results have been obtained. We used these approximations and bounds to create a robust simulation model for a large-scale engine maintenance system. These bounds and approximations were used in evaluating the robustness of our simulation model.

In the next section, we will describe some of the results associated with G/G/c. In the subsequent section, we will present the simulation model and some numerical results. We will end this paper with a few concluding remarks.

2. G/G/c system

The G/G/1 system and its theoretical results are used to derive what is presently known about the G/G/c system and thus will be discussed first. We consider a G/G/1 system consisting of a single server with independent and identically distributed interarrival times as well as service times and unlimited queueing capacity. Let X denote interarrival times and let $f_x(x)$, $1/\lambda$, and σ_x^2 denote the probability density function (pdf), the mean, and the variance of X, respectively. In addition, let S, $f_s(s)$, $1/\mu$, and σ_s^2 represent those corresponding for the service times. Although there are no closed form solutions for this model, there are some useful bounds developed in recent years for the quantities L, L_q, W, and W_q (see [4, 7]).

For G/G/1 systems with no restrictions on the interarrival or on the service time pdf's, several bounds have been developed (see [8, 9]). These bounds, in essence, state that for the average steady-state waiting time in queue, W_q, we have

$$\frac{\rho^2(1+C_s^2)-2\rho}{2\lambda(1-\rho)} < W_q \leq \frac{\lambda(\sigma_x^2+\sigma_s^2)}{2(1-\rho)}, \tag{2.1}$$

where $C_s = \sigma_s\mu$ is the coefficient of variation for the service times, and $\rho = \lambda/\mu$ is the utilization factor. For the stability of the system we must have $\rho < 1$. Note that the lower bound given above is not tight. This becomes obvious from the fact that, even at very high utilization rates, the bounds take negative values, unless $C_s > 1$. But for C_s to be greater than 1, it must be that the service time pdf must be "more random" than the negative exponential pdf which has its $C_s = 1$.

2.1. Desired class property. A tight simple lower bound is given in [9] for a class of G/G/1 queues, which includes most practical problems encountered in the real world. Thus, class requirement is that all queueing systems in it must have interarrival time pdf, $f_x(x)$, satisfying the following property:

$$E[X - t \mid X > t] \leq \frac{1}{\lambda} \quad \forall t \geq 0. \tag{2.2}$$

If it is known that any given interarrival gap lasted more than a time t, then the condition above requires that the expected length of the remaining time, $X - t$, in that gap be less than the unconditional expected length of the gap, $E[X](= 1/\lambda)$. This is of course true for the negative exponential variable, and in that case the condition becomes equality. When the condition holds, then we have

$$U - \frac{1+\rho}{2\lambda} \leq W_q \leq U, \quad U = \frac{\lambda(\sigma_x^2+\sigma_s^2)}{2(1-\rho)}. \tag{2.3}$$

The upper and lower bounds may now be derived using this and by applying Little's formula, $L = \lambda W$, $L_q = \lambda W_q$, and the fact that $W = 1/\mu + W_q$. The following is easily

obtained:

$$\lambda \cdot U - \frac{1+\rho}{2} \leq L_q \leq \lambda \cdot U. \tag{2.4}$$

This implies that the difference between the upper and the lower bounds is $(1+\rho)/2$, but $0 < \rho < 1$, so this difference is always between 0.5 and 1. Thus, we can find the average queue length to within an accuracy of between 0.5 and 1 (depending on the value of ρ). Note that most "well-behaved" arrival time distributions satisfy the condition, including uniform, triangular, or beta-type pdf's, which often are reasonably good approximations of many general interarrival time pdf's. Only a few common continuous random variables, such as those in the hyperexponential family, which are "more random" (informally speaking) than the negative exponential random variable, do not satisfy the condition.

2.2. Under heavy traffic. Another important result that is available for the G/G/1 system is known as the heavy-traffic approximation (for more information see [5]). It applies for values of ρ near 1 and thus provides estimates for waiting times when it is known that waiting times are large. When ρ is near 1, the distribution of steady-state waiting time in queue in a G/G/1 system is approximately *negative exponential* with mean value $W_q = U$. The average waiting time for G/G/1 queueing systems is dominated by a $(1-\rho)^{-1}$ term under steady-state conditions, as the utilization ratio tends to 1. Consequently, the type of behavior that is normally seen in a simple M/M/1 system is also present for entirely general arrival- and service-time distributions, G/G/1.

2.3. G/G/c bounds. The only general results on G/G/c system [2] that have been obtained to date are in the form of quite relaxed upper and lower bounds on average steady-state queueing characteristics. These bounds are often computed by, first, comparing a G/G/c system with a G/G/1 system that has the same "service behavior" as the G/G/c system. That is, the single server in G/G/1 works c times as fast as each of the servers in G/G/c and by applying the earlier results on G/G/1, given in the previous section. The most useful and applicable bounds on the average waiting time in queue which have been derived to date for G/G/c systems, based on those of G/G/1, is

$$W_q^1 - \frac{(c-1)\mu E[S^2]}{2c} \leq W_q \leq \frac{[\sigma_X^2 + (1/c)\sigma_S^2 + ((c-1)/c^2)(1/\mu^2)]\lambda}{2(1-\lambda/c\mu)}, \tag{2.5}$$

where for each of the c servers, μ, σ_S^2, and $E[S^2]$ are the rate, variance, and the second moment of service time, respectively. W_q^1 denotes the mean waiting time for a G/G/1 system with a service time denoted by a random variable $S^1 = S/c$ with service c times faster than that of each of the c servers in the G/G/c system, but with an identical arrival process. If W_q^1 is known or is computed using the results discussed above, we can substitute an exact expression. Note that for the general M/G/1 system we have the following well-known

results, which can be used in deriving the G/G/c approximation bounds:

$$
\begin{aligned}
P_{\circ} &= 1-\rho, \qquad L = \rho + \frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)},\\
W &= \frac{L}{\lambda} = \frac{1}{\mu} + \frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)},\\
W_q &= W - \frac{1}{\mu} = \frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)} = \frac{\lambda\left[(1/\mu^2)+\sigma_S^2\right]}{2(1-\rho)},\\
L_q &= \lambda W_q = \frac{\rho^2+\lambda^2\sigma_S^2}{2(1-\rho)}.
\end{aligned}
\tag{2.6}
$$

Thus, for example, for the M/G/c queueing system, one should use the exact expression for W_q^1 given above with $1/c\mu$ and σ_S^2/c^2, for the expected value and variance of the service times, respectively.

The corresponding heavy-traffic approximation for G/G/c systems has been derived [6]. This result implies that for $\lambda c\mu$ approaching 1 in a G/G/c system, the waiting time in queue under steady-state conditions assumes a distribution that is approximately *negative exponential* with mean value

$$
W_q = \frac{\left[\sigma_X^2 + (\sigma_S^2/c)\right]\lambda}{2(1-\lambda/c\mu)}. \tag{2.7}
$$

Note once more that expected waiting time is dominated by a $(1-\rho)$ term, as ρ approaches 1 ($\rho = \lambda/c\mu$ for multiserver systems). We used the above results for G/G/c to have a point of reference for the simulation and tested the results against these theoretical backdrops.

3. An overview of the simulation model

In terms of modeling, we are interested in the flow of entities (e.g., spares, personnel), the state of the system (e.g., engine not serviceable, spares inventory), and the processes (e.g., service time, sortie rates). The structure of the model is based on a set of hierarchical, functional blocks that generate and modify entities, processes, and attributes. These blocks represent main bases, airfields, and intermediate maintenance shops.

In general, the simulation is based on the following sequence of events: aircraft are flown from main bases or remote sites to meet certain flying requirements. After each mission, the aircraft and their engines are inspected at the airfield and in most cases they are fully operational within hours. However, when engines accumulate enough flying hours, or when unscheduled maintenance is required, engines are removed from the planes and sent to an intermediate maintenance facility. Flightline maintenance includes servicing, repairs, cycle recording, and tracking, which are coordinated with the engine management branch (EMB) and IMF. On the flightline, installed aircraft engines are serviced on a daily basis, which includes servicing the oil, inspecting the chip detectors, and entering the intakes and augmentor to inspect for foreign object damage (FOD) and

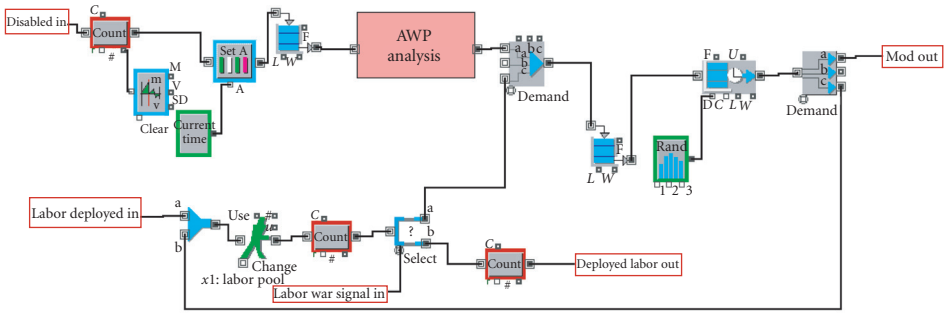


Figure 3.1. Intermediate maintenance shop.

external engine damage. In addition, engine cycles are recorded in the comprehensive engine management systems (CEMS) database. CEMS enables the EMB to monitor usage of engines and modules (when used) to determine the need for inspections and time change technical orders (TCTOs). The flightline also performs all engine removals and installations. After the flightline removes an engine for maintenance at the IMF, it sometimes performs sheet metal work on the engine bay and replaces some of the hydraulic lines and cables in the aircraft engine bay that have been damaged due to chafing, cracks, or heat.

The IMF is responsible for both scheduled and unscheduled off-equipment engine maintenance. Scheduled maintenance includes module time changes, TCTOs, and other inspections and repairs. Unscheduled maintenance consists primarily of performance-related problems that either cannot be corrected by the flightline or are beyond their capabilities per technical order. For unscheduled maintenance, the intermediate maintenance shop often performs a preliminary test cell run to troubleshoot the engine and identify other potential problems. The IMF is capable of replacing any module in a modular engine and also repairs some of the modules while sending others to the depot. It is also responsible for packing engines for transportation.

The IMF operates the engine test cell facility and functions. As part of this function, the IMF personnel transport engines, hook up cables, and fuel lines conduct pre- and post-run engine inspection, and disconnect cables and fuel lines. In many instances, the IMF also serves as a source of expertise to back up the flightline and provide quick response repair or cannibalizing key parts as needed. This organization is quite large (100–150 people for a fighter wing) and occupies an industrial space equipped with five or more work bays of 1500 square feet each, an overhead crane, supply storage, backshops for specialized repair activities, and a test cell. The test cell is typically located offsite in a "hush house" where a fully-assembled engine can be run at full power for testing purposes.

The general flow of IMF work is as follows with portion of the process depicted in Figure 3.1:

(i) receive engine from the flightline;

(ii) perform inspection and time change check;

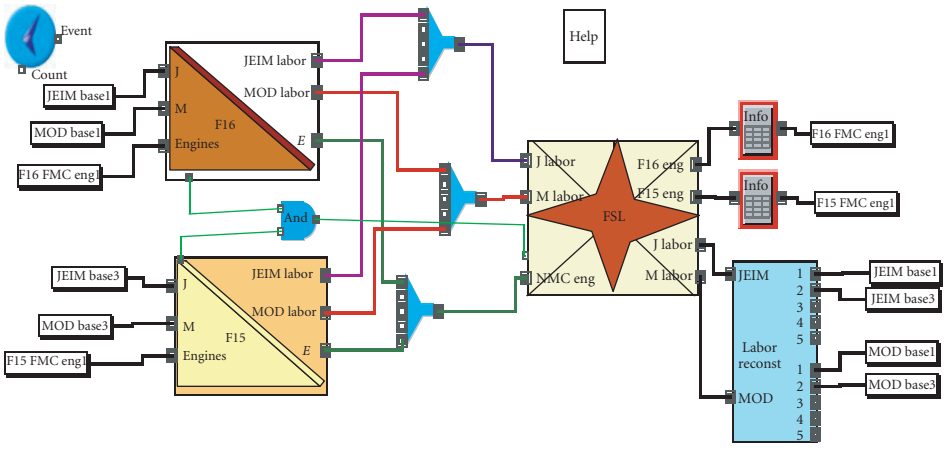


Figure 3.2. Top-level view of the simulation model.

(iii) perform database history check;
(iv) create a job in core automated maintenance system (CAMS);
(v) assign engine to a crew;
(vi) determine required repairs;
(vii) decide on complete or partial disassembly;
(viii) conduct other inspections;
(ix) conduct teardown;
(x) perform IMF repair and maintenance;
(xi) perform module work, if needed, at module shop;
(xii) assemble engine;
(xiii) send to test cell "(hush house);"
(xiv) conduct final inspection.

The first requirement for the model is the number and types of aircraft, and the number and the age of installed engines. The aircraft and engines are combined to form fully operational aircraft. They are sorted, based on the age of the engine, and are then queued for flying. After each sortie, the aircraft is sent to the airfield block where it is inspected and maintained. Each aircraft that passes the inspection is sent back to the pool of available aircraft. Some aircraft require minor repair, which is performed on the flight line. The number of engines pulled from the aircraft is a function of the age and the type of the engine. The detached engines are tagged according to the removal type (i.e., scheduled or unscheduled) and are sent to the IMF shop. Aircraft are then identified as not operational and are queued for the next available serviceable engine. These aircraft are either put back to service immediately, if there are serviceable spares available, or they await the arrival of engines from the maintenance shop. Figure 3.2 illustrates the top level view of the simulation model using block diagrams from Extend software (Extend is a registered

trademark of Imagine That, Inc.). This figure presents notional F-15 and F-16 jet fighters with a centralized maintenance facility.

At the maintenance facility, engines are queued in two parallel lines, the first is for the engines that require parts that are not available and the other is for engines that await maintenance. The modular engines that have been processed by the IMF shop are sent to the module shops. Engines that enter the module shop are separated into five modules. Engines that leave the module shop are sent to the assembly and test cell. In this section, engines are queued for assembly, the test cell, and the final inspection. After assembly and test cell, engines are sent to the spare engines pool to be installed on the aircraft to create fully operations aircraft. These aircraft leave this section to join the pool of other aircraft and the whole cycle starts again. Figure 1.1 illustrates this process for only one main and operating base. The model, however, has taken into account a problem with several such bases (for more information on the simulation model, see [1]).

3.1. Simulation setup and data analysis. We analyzed a number of possible support configurations for the IMF involving various combinations of centralized and decentralized locations. Centralized maintenance structures include forward support bases, while decentralized locations include home base support and maintenance at forward operating bases. Each structure was assessed under both a war and a peacetime scenario.

Here we describe in detail the specific IMF alternatives we evaluate in this analysis.

(i) *Decentralized-deployed.* In this alternative peacetime, maintenance is provided by IMFs located at each base. When part of a unit is deployed, part of that unit's IMF deploys to the appropriate forward bases as well.

(ii) *Decentralized-no deployment.* As with the previous alternative, each of the peacetime bases has its own IMF, but in this case the home IMF supports any deployed forces from its unit as well. The home base is sized so that it has the resources to support both peacetime and wartime flying.

(iii) *Decentralized-forward support location.* As with the previous two alternatives, each peacetime base has its own IMF, but when the units deploy, some of the IMF personnel (but not their equipment) deploy to a single overseas base in theater from which all deployed units are supported.

(iv) *US support location-forward support location.* In this alternative all units are supported in peacetime by a single centralized operation at home, which deploys personnel to an overseas base in theater when conflict occurs. In peacetime, the home IMF is staffed with the sum of the rail teams needed for deployment and those required to keep the nonengaged forces flying.

(v) *Home support location.* In this last alternative all units everywhere are supported by a single shop both during peacetime and in deployment.

During the simulations, we evaluated each of these alternatives using three broad metrics. The first is performance: does the alternative provide the required support for

operational flying? In peacetime this means being able to maintain the requisite flying for pilot training; in wartime it means being able to meet the required number of sorties day by day. The second metric is resources: what does the alternative require to provide adequate performance? For jet engines, one of the key resources is spare engines, which can provide a hedge against uncertainties. Other resources are personnel and transportation costs, and the evaluation provides an indication of the tradeoff between these two. The third metric is uncertainty: how well does the alternative respond to unforeseen events? For this metric, we evaluate how robust the Alternatives are to changes in the engine removal rate.

Many of the inputs to the model were provided by analysis of data drawn from the comprehensive engine management system (CEMS), reliability and maintainability management information system (REMIS), which rolls up data from the base-level core automated maintenance system (CAMS), as well as data in both electronic and paper form provided by the units we visited. The CEMS data provided information on total repair time for individual engines, engine not mission capable due to supply (ENMCS) times and transportation times for some engines were provided by some of the bases. REMIS provided a check on the CEMS data for overall engine repair and provided repair data for module work.

3.2. Simulation results. In this section, we will present some of the parameters used in our analysis and the results of the simulation runs. We will illustrate these parameters by running a scenario with 36 F-16s and 66 F-15s. The model is run for about two simulated years.

Engines are typically set on a rail and require a 5-person team per shift. The regular shift is about 8 hours and the shops operate at 2 shifts a day. During peak demand period, the shops may shift their operations to 24 hours a day, seven days a week with each shift as long as 12 hours. The capacity of the IMF is determined by the combination of rails and the personnel, a "rail team." Other shops have different architecture but all are bounded by number of staff and the equipment. Airfields and the transportation network are bounded by the capacity of the flight line and the number of transporters, respectively. There are three smaller main bases with three-rail team capacity and a large one with 7-rail teams capacity; in other words, 3 and 7 parallel servers, respectively. There is also a remote facility with 8 rail teams. The other parts of the shop (e.g., the module shop) are sized accordingly.

On average about 119 customers entered the system (with variance of 253 and standard deviation of 15). At the end of the simulation run, about 105 customers were served. The IMF shop at the remote site (with 8 servers) reported an average wait time of 5.538461538462 days. Although the arrival and the service times varied widely, as they depend heavily on the other parts of the system, the reported wait time seemed reasonable and was consistent with the theoretical bounds. Using the Poisson distribution, we get a wait time of 2.06 days and 4.13 using the exponential distribution. Table 3.1 illustrates the theoretical bounds for a single server process in the inspection shop. The simulation model reported an average of 0.808499576845 for the queue length and 10 days for the average wait.

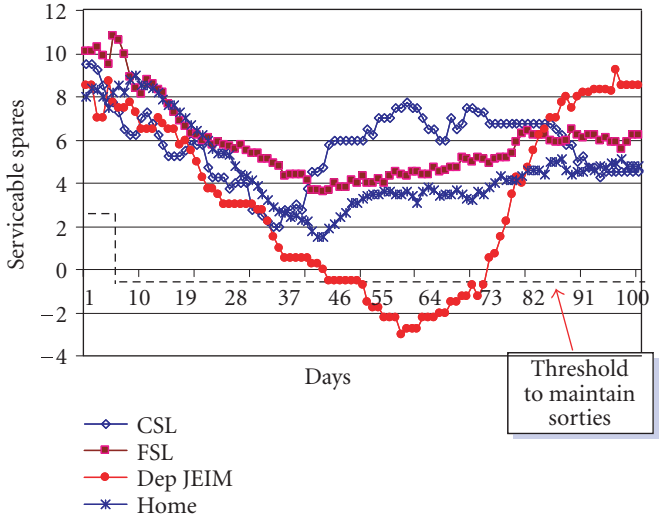


Figure 3.3. Deployed F-16 results.

Table 3.1. Sample results for a G/G/1 system.

Distribution	Utilization factor	Var (X)	Var (S)	Wait time		Length	
				LB	UB	LB	UB
Poisson	1.32	0.33	0.25	—	—	0.06130	—
Eponential	0.757576	0.1089	0.0625	0.78125	1.07125	3.125	3.24621
Uniform	—	0.02083	—	—	—	—	—
	0.08	3	0.003333	—	0.01050	—	0.00840
Normal	—	0.90090	—	—	—	—	—
	0.18018	1	5	2.58725	3.24225	2.51103	2.92094

Table 3.2 illustrates the arrival and departure rates for the sequence of servers in the maintenance process. Some customers bypass the first queue and enter the second queue with multiple servers. After the service, some customers, again, bypass the next server. In this section, there are five parallel servers and customers depending on their requirement must enter a particular server queue. Finally all customers enter the last server.

Table 3.3 illustrates the theoretical versus simulated bounds for the first queue, in the eight-server scenario discussed above.

Figure 3.3 presents the results from the deployment portion of the operation for the F-16 aircraft, comparing the centralized US support location (US), forward support location (FSL), deployment maintenance shops (Dep IMF), and home base locations (Home).

Table 3.2. Arrive and departure in the IMF.

Server (single)			Server (multiple)		Server (single) 5 parallel			Server (single)	
A	D	B	A	D	A	D	B	A	D
33	25	86	111	104	94	92	10	102	96

Table 3.3. Sample results for a G/G/8 system.

Distribution	Queue wait time		Length		Simulation results	
	LB	UB	LB	UB	W	L
Poisson arrivals	—	—	—	—	—	—
Exponential	3.21107	4.97622	1.22465	1.64215	5.04683	5.66852
Uniform	0.68181	1.3561	.02367	.15627	3.97198	3.20338
Normal	2.58725	3.24225	2.51103	2.92094	5.04280	5.89083

4. Concluding remarks

We presented a closed stochastic simulation network model and several approximation and bounding options available in a G/G/c system. The model was implemented under several simulation environments, including Extend v6 [3]. The analysis was conducted to verify the integrity of the simulation model used to developed alternative policy options conducted on behalf of the US Air Force and presented in [1]. We showed that the theoretical bounds could be used to approximate mean capacities at various queues. In this paper only the results for G/G/8 was presented in order to avoid lengthy numerical tabulation of the results. However, such consistency was observed amongst the other queues.

References

[1] M. Amouzegar, L. S. Galway, and A. Geller, *Supporting expeditionary aerospace forces: an analysis of jet engine intermediate maintenance options*, Tech. Rep. MR-1431-AF, RAND, California, 2001.

[2] S. L. Brumelle, *Some inequalities for parallel-server queues*, Operations Research **19** (1971), 402–413.

[3] B. Diamond, S. Lamperti, D. Krahl, and A. Nastasi, *Extend v6 User's Guide*, 2002.

[4] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 3rd ed., Wiley Series in Probability and Statistics: Texts and References Section, John Wiley & Sons, New York, 1998.

[5] J. F. C. Kingman, *On queues in heavy traffic*, Journal of the Royal Statistical Society. Series B. Methodological **24** (1962), 383–392.

[6] L. Köllerström, *Heavy traffic theory for queues with several servers: I*, Journal of Applied Probability **11** (1974), 544–552.

[7] R. Larson and A. Odoni, *Urban Operations Research*, Prentice-Hall, New Jersey, 1981.

[8] W. G. Marchal, *Some simpler bounds on the mean queuing time*, Operations Research **26** (1978), no. 6, 1083–1088.
[9] K. T. Marshall, *Some inequalities in queuing*, Operations Research **16** (1968), 651–665.

Mahyar Amouzegar: The RAND Corporation, Santa Monica, CA 90407-2138, USA;
College of Engineering, California State University, Long Beach, CA 90840, USA
E-mail address: mahyar@csulb.edu

Khosrow Moshirvaziri: Information Systems Department, California State University, Long Beach, CA 90840, USA
E-mail address: moshir@csulb.edu

AN ANALYTICAL CHARACTERIZATION FOR AN OPTIMAL CHANGE OF GAUSSIAN MEASURES

HENRY SCHELLHORN

Received 25 February 2006; Revised 9 June 2006; Accepted 9 June 2006

We consider two Gaussian measures. In the "initial" measure the state variable is Gaussian, with zero drift and time-varying volatility. In the "target measure" the state variable follows an Ornstein-Uhlenbeck process, with a free set of parameters, namely, the time-varying speed of mean reversion. We look for the speed of mean reversion that minimizes the variance of the Radon-Nikodym derivative of the target measure with respect to the initial measure under a constraint on the time integral of the variance of the state variable in the target measure. We show that the optimal speed of mean reversion follows a Riccati equation. This equation can be solved analytically when the volatility curve takes specific shapes. We discuss an application of this result to simulation, which we presented in an earlier article.

1. Introduction

We consider two Gaussian measures. In the "initial" measure the state variable is Gaussian, with zero drift, (we chose zero drift for ease of exposition, but the same development applies to a nonzero drift) and time-varying volatility. In the "target measure" the state variable follows an Ornstein-Uhlenbeck process, with a free set of parameters, namely, the time-varying speed of mean reversion. We look for the speed of mean reversion that minimizes the variance of the Radon-Nikodym derivative of the target measure with respect to the initial measure under a constraint on the time integral of the variance of the state variable in the target measure.

We studied this problem in an earlier article (see Schellhorn [10]), where we explained one application of this result to the field of Monte Carlo simulation. It is sometimes important to resimulate a system under a different measure than the initial measure. The immediate example is sensitivity analysis. Another example is in the field of finance, where practitioners are often interested in seeing the results of their simulations in two different

Hindawi Publishing Corporation
Journal of Applied Mathematics and Decision Sciences
Volume 2006, Article ID 95912, Pages 1–9
DOI 10.1155/JAMDS/2006/95912

measures, the "actual measure," and the "risk-neutral" measure. One of these measures has typically a free parameter, or sets of parameters. Suppose the goal is to calculate $E[z]$ under two different measures, and that the integrand $z(\omega)$—which is expensive to compute—was initially simulated in the initial measure. We argued that a computationally better resimulation estimator (compared to resimulating $z(\omega)$ in the target measure) was the sum of the initial $z(\omega)$ weighted by the Radon-Nikodym derivative $g(\omega)$ of the target measure with respect to the initial measure. However, the product $g(\omega)z(\omega)$ tends to have a larger variance than $z(\omega)$, and this fact may outweigh the performance gain of not resimulating z. Care must be therefore taken in selecting a target measure for simulation performance, and we suggested that a good performance measure was the variance of g. When the state variable x is assumed Gaussian in both measures (which is very often the case in practice for better analytical tractability), the only free parameter is the speed of mean reversion a of x in the target measure.

The problem above is completely characterized once one of several constraints on the autocovariance function of x in the target measure are introduced—we do not consider the usually less interesting case, where x is not first-moment stationary in the target measure. In Schellhorn [10] we considered in turn two possible constraints:

(i) a constraint on the terminal variance of x,

(ii) a constraint on the average variance of x,

and showed that, in both cases, the control satisfied (together with other variables) a system of four nonlinear ordinary differential equations. This system happened to be quite difficult to solve numerically. Nevertheless, the so-called "change of measure" resimulation technique proved out to be effective on various examples.

Another potential application of this problem is the theory of incomplete markets in mathematical finance. Several authors (see, e.g., Rouge and El Karoui [8], Delbaen et al.[2]) explore the duality between utility maximization and optimal choice of measure. If the utility function is exponential, the dual objective to minimize is the relative entropy of the target measure, that is, the first moment of $g \log g$. If the utility function is quadratic, the dual objective to minimize is the second moment of g (see Duffie and Richardson [3], Schweizer [11], Bellini and Frittelli [1]). A majority of authors seems to have pursued the first avenue, that is, minimizing entropy, because among others of its better tractability (Rheinlaender [7]). We conjecture that the result of this paper may help research in the second avenue, that is, quadratic utility functions.

In this article, we consider only a constraint on the average variance of x. Compared to our earlier article, we use a different representation of the second moment of g, which turns out to be easier to handle analytically. Using the maximum principle, we show that the optimal speed of mean reversion follows a Riccati equation. We show solutions of the problem in two cases, when volatility is constant, and when volatility is an exponential function of time. We suspect that other cases are also amenable to closed form formulae. Finally, we compare our exact results to the approximation given in Schellhorn [10].

2. Model and results

Notation 1. The complete filtered probability space $(\Omega, \mathcal{F}, P^I)$ supports a Brownian motion W^I. We use the superscripts ^{I} and ^{T} to refer to the probability measure, expectation

operator, variance (Var) operator, and Brownian motion in the initial/terminal measure. When not shown otherwise, the expectation and variance operators are taken at time zero.

The dynamics of the variable x of interest are

$$\begin{aligned} dx(t) &= \sigma(t)dW^I(t), \\ x(0) &= 0, \end{aligned} \tag{2.1}$$

where $\sigma > 0$ is a deterministic function of time. The terminal measure P^T supports one Brownian motion W^T, with

$$dW^T(t) = dW^I(t) + \frac{a(t)x(t)}{\sigma(t)}dt. \tag{2.2}$$

Once the speed of mean reversion $a(t)$ is specified, P^T becomes fully specified. We define the Radon-Nikodym derivative process:

$$g(t) \equiv E^I\left[\frac{dP^T}{dP^I}\,\middle|\,\mathcal{F}_t\right]. \tag{2.3}$$

By Girsanov theorem,

$$dg(t) = \frac{a(t)x(t)}{\sigma(t)}dW^I(t). \tag{2.4}$$

The optimization problem is to minimize the variance of g under a constraint on the average variance of the state variable in the terminal measure:

$$\min E^I[g^2(t)], \tag{2.5}$$

$$\mathrm{Var}^T\left[\int_0^\infty x^2(t)dt\right] \le A. \tag{2.6}$$

Theorem 2.1. *The speed of mean reversion that solves (2.5) and (2.6) is of the form*

$$a(t) = \sigma^2(t)y(t), \tag{2.7}$$

where y solves the Riccati equation

$$\begin{aligned} \frac{dy(t)}{dt} &= -\sigma^2(t)y^2(t) - \lambda, \\ y(T) &= 0, \end{aligned} \tag{2.8}$$

and $\lambda \ge 0$ is the Lagrange multiplier of relation (2.6).

We now look at the solution of the Riccati equation for particular volatility functions.

Case 1 (σ is constant). The solution to the Riccati equation is

$$a(t) = \sigma\sqrt{\lambda}\tan\left(\sigma\sqrt{\lambda}(T-t)\right). \tag{2.9}$$

As required by the transversality conditions, $a(T) = 0$. As expected the speed of mean reversion is increasing in λ and decreasing in t. We notice that when λ is small the speed of mean reversion is a linear decreasing function.

Case 2 ($\sigma^2(t) = \alpha\exp(-kt)$ for $\alpha > 0$). We write J_1 for the Bessel functions of the first kind. Let

$$C \equiv \frac{-(1/2)J_1\left((2/k)\sqrt{\lambda\alpha\exp(-kT)}\right) - (2/k)\sqrt{\lambda\alpha\exp(-kT)}J_1'\left((2/k)\sqrt{\lambda\alpha\exp(-kT)}\right)}{J_{-1}\left((2/k)\sqrt{\lambda\alpha\exp(-kT)}\right) + (2/k)\sqrt{\lambda\alpha\exp(-kT)}J_{-1}'\left((2/k)\sqrt{\lambda\alpha\exp(-kT)}\right)}. \tag{2.10}$$

Then

$$\begin{aligned} y(t) &= -\sigma^2(t)\frac{k\exp(kT)}{\alpha}\frac{u'\left(e^{k(T-t)}\right)}{u\left(e^{k(T-t)}\right)}, \\ u(s) &= s^{1/2}J_1\left(\frac{2}{k}\sqrt{\lambda\alpha\exp(-kT)s}\right) + CJ_{-1}\left(\frac{2}{k}\sqrt{\lambda\alpha\exp(-kT)s}\right). \end{aligned} \tag{2.11}$$

Lemma 2.2. *Let* $v(t) = E^T[x^2(t)]$. *Then*

$$E^I[g^2(T)] = \exp\left(\int_{t=0}^{T}\frac{a^2(t)}{\sigma^2(t)}v(t)dt\right). \tag{2.12}$$

Proof. Let $\mu = ax/\sigma$. Then

$$\begin{aligned} E^I[g^2(T)] &= E^T[g(T)] \\ &= E^T\left[\exp\left(-\int_0^T\mu(t)dW^I(t) - \frac{1}{2}\int_0^T\mu^2(t)dt\right)\right] \\ &= E^T\left[\exp\left(-\int_0^T\mu(t)\left(dW^T(t) - \mu(t)dt\right) - \frac{1}{2}\int_0^T\mu^2(t)dt\right)\right] \\ &= E^T\left[\exp\left(\int_0^T\mu^2(t)dt\right)\right]. \end{aligned} \tag{2.13}$$

We obtain then

$$\begin{aligned} E^I[g^2(T)] &= E^T\left[\exp\left[\int_0^T \frac{a^2(t)}{\sigma^2(t)}\left[W^T(v(t))\right]^2 dt\right]\right] \\ &= E^T\left[\exp\left[\int_0^{v(T)} \frac{dt}{dv}\bigg|_u \frac{a^2(v^{-1}(u))}{\sigma^2(v^{-1}(u))} W^2(u)du\right]\right] \\ &= E^T\left[\exp\left[\int_0^{v(T)} h(u)W^2(u)du\right]\right], \end{aligned} \tag{2.14}$$

where we have defined

$$h(u) = \frac{dt}{dv}\bigg|_u \frac{a^2(v^{-1}(u))}{\sigma^2(v^{-1}(u))}. \tag{2.15}$$

To calculate the Carleman-Fredholm determinant (see, e.g., Grasselli and Hurd [4] or Levendorskii [5]), we resort to a discrete approximation. We first define V as the smallest value larger than or equal to $v(T)$ so that $V/\Delta u$ is integer. We also define

$$\begin{aligned} H(u) &= \int_u^{v(T)} h(s)ds, \\ z &= \left[z(1) \quad \cdots \quad z\left(\frac{V}{\Delta u}\right)\right], \\ \Sigma^{-1} &= \begin{bmatrix} 1-2H(\Delta u)\Delta u & -4H(2\Delta u)\Delta u & \vdots & -4H(V)\Delta u \\ 0 & 1-2H(2\Delta u)\Delta u & \vdots & \cdots \\ \vdots & \vdots & \vdots & -4H(V)\Delta u \\ 0 & \vdots & 0 & 1-2H(V)\Delta u \end{bmatrix}. \end{aligned} \tag{2.16}$$

We calculate

$$\begin{aligned} E^I[g^2(T)] &= E^T\left[\exp\left[\int_{u=0}^{v(T)} h(u)W^2(u)\right]\right] \\ &= \lim_{\Delta u\to 0} E^T\left[\exp\left[\sum_{u=1}^{V/\Delta u} h(u\Delta u)\sum_{s=1}^{u}\sum_{t=1}^{u} z(s)z(t)(\Delta u)^2\right]\right] \\ &= \lim_{\Delta u\to 0} \frac{1}{\sqrt{(2\pi)^{V/2\Delta u}}} \int\cdots\int \exp\left(-\frac{1}{2}z\Sigma^{-1}z\right)dz \end{aligned}$$

$$
\begin{aligned}
&= \lim_{\Delta u\to 0} \frac{1}{\sqrt{(1-2H(\Delta u)\Delta)\cdots(1-2H(V)\Delta u)}} \\
&= \lim_{\Delta u\to 0} \exp\left(\int_0^{\nu(t)} H(u)du\right) \\
&= \exp\left(\int_{u=0}^{\nu(T)}\int_{s=u}^{\nu(t)} \frac{dt}{dv}\Big|_s \frac{a^2(\nu^{-1}(s))}{\sigma^2(\nu^{-1}(s))}ds\,du\right) \\
&= \exp\left(\int_{s=0}^{\nu(T)} \frac{dt}{dv}|_s \frac{a^2(\nu^{-1}(s))}{\sigma^2(\nu^{-1}(s))}\nu(\nu^{-1}(s))ds\right) \\
&= \exp\left(\int_{t=0}^{T} \frac{a^2(t)}{\sigma^2(t)}\nu(t)dt\right).
\end{aligned}
\tag{2.17}
$$

□

Proof of Theorem 2.1. The problem is

$$
\begin{aligned}
&\min_a \int_0^T \left(\frac{a^2(t)}{\sigma^2(t)}+\lambda\right)\nu(t)dt, \\
&\frac{d\nu(t)}{dt} = -2a(t)\nu(t)+\sigma^2(t), \\
&\nu(0)=0.
\end{aligned}
\tag{2.18}
$$

□

The Hamiltonian is

$$
H(\nu(t),a(t),t) = -\left(\frac{a^2(t)}{\sigma^2(t)}+\lambda\right)\nu(t)+z(t)\big(-2a(t)\nu(t)+\sigma^2(t)\big). \tag{2.19}
$$

The Pontryagin optimality conditions are

$$
\frac{\partial H}{\partial a} = -\frac{2a\nu}{\sigma^2}-2z\nu = 0, \tag{2.20}
$$

$$
\frac{dz(t)}{dt} = \frac{a^2(t)}{\sigma^2(t)}+\lambda+2a(t)z(t), \tag{2.21}
$$

$$
z(T)\nu(T)=0. \tag{2.22}
$$

We note that these optimality conditions are sufficient (see Mangasarian [6]). From (2.20) we obtain

$$
z(t) = -\frac{a(t)}{\sigma^2(t)}, \tag{2.23}
$$

which we reinsert in (2.21)

$$\frac{d}{dt}\left(\frac{a(t)}{\sigma^2(t)}\right) = -\frac{a^2(t)}{\sigma^2(t)} - \lambda. \tag{2.24}$$

We let $y = a/\sigma^2$ and obtain the result. The transversality condition (2.22) imposes $a(T) = 0$.

3. Example

In this section, we compare on two examples the optimal control given by the solution of the theorem, to the approximate optimal control given in Schellhorn [10]. We now expose briefly the approximation approach. In the latter article, we did not exploit the lemma, but used the following representation for our objective:

$$E^I\left[g^2(T)\right] = \exp\left[\int_0^T \sigma^2(t) f(t)dt\right], \tag{3.1}$$

where

$$\begin{aligned} \frac{df(t)}{dt} &= -\frac{a^2(t)}{\sigma^2(t)} + 4a(t)f(t) - 2\sigma^2(t)f^2(t), \\ f(T) &= 0. \end{aligned} \tag{3.2}$$

The representation (3.1)-(3.2) when inserted in the optimization problem (2.5), (2.6) results in optimal control problem involving two state variables: f and v. The optimality conditions of that problem (which were not even sufficient) turned out to be quite difficult to solve numerically. Instead, we suggested to reduce the state space to only one variable, in a line similar to Sannutti [9].

The approximated optimal control follows, then

$$a_{\text{approx}}(t) = -\frac{\sigma^2(t)z(t)v(t)}{\int_0^t \sigma^2(u)du}, \tag{3.3}$$

where the costate variable z follows:

$$\frac{dz}{dt} = \lambda + 2az, \tag{3.4}$$

under the terminal constraint $z(T) = 0$.

We report in Figures 3.1 and 3.2 our results for two different volatilities:

(i) $\sigma(t) = 0.2$ in Figure 3.1;

(ii) $\sigma(t) = 0.2(1 + 0.2\cos(t/4))$ in Figure 3.2.

In both cases, the *relative average variance of* x is the ratio between the cumulated variance $\text{Var}^T[\int_0^T x^2(t)dt]$ "with mean reversion" and the cumulated variance $\text{Var}^I[\int_0^T x^2(t)dt]$ "without mean reversion." Since the constraint (2.6) is clearly tight, the numerator of this ratio is equal to our constraint A. On the y-axis, we report the logarithm of the second moment of $g(T)$, calculated according to the representation of this article, that is, (2.12).

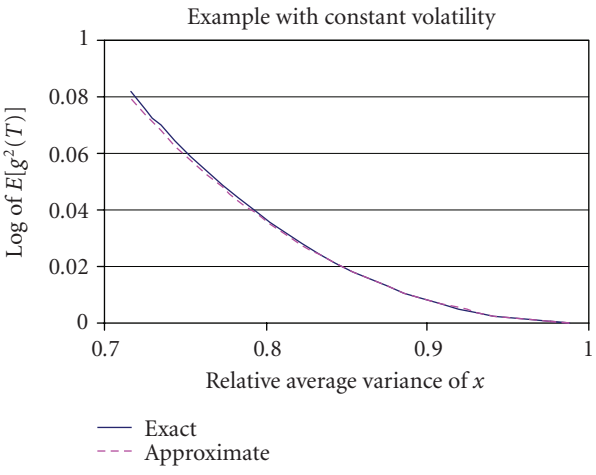


Figure 3.1. Logarithm of $E[g^2(T)]$ as a function of the ratio of A over the cumulated variance of x in the uncontrolled case ($a = 0$). The volatility is $\sigma(t) = 0.2$ and $T = 3$.

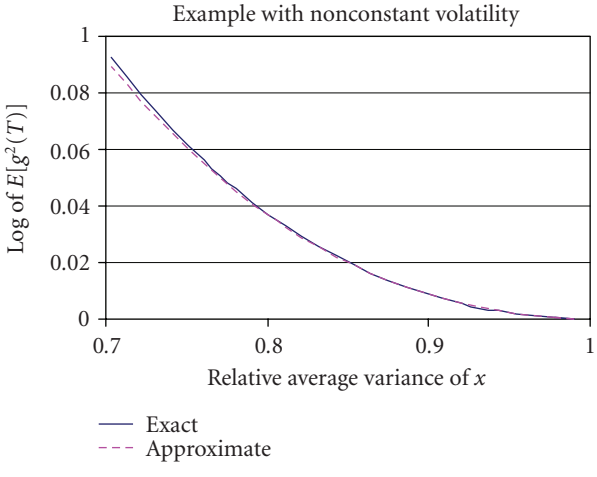


Figure 3.2. Logarithm of $E[g^2(T)]$ as a function of the ratio of A over the cumulated variance of x in the uncontrolled case ($a = 0$). The volatility is $\sigma(t) = 0.2(1 + 0.2\cos((t/4)))$ and $T = 3$.

It turns out that both methods, the exact method of this article, and the approximate one, yield remarkably similar results in these two examples.

4. Conclusions

This article provides an alternate characterization of the solution of an optimal control problem first introduced in the literature by us in Schellhorn [10]. The result presented in this article is stronger, since it is not the result of a reduction of the problem. The

examples we presented show however a remarkable coincidence in results between both methods. We emphasize that this needs not be the case.

Our methodology can be applied to Monte Carlo resimulation, that is, simulation in two different measures. We reported in our earlier article that the "change of measure" resimulation scheme, where we simulate the cash flows $z(\omega)$ only once (to calculate market value), and then adjust them by $g(\omega)$ to calculate the empirical distribution, was up to twice faster than a "traditional scheme," where two independent simulations were performed. The same speed improvement can be attained using the method presented here.

References

[1] F. Bellini and M. Frittelli, *On the existence of minimax martingale measures*, Mathematical Finance **12** (2002), no. 1, 1–21.

[2] F. Delbaen, P. Grandits, T. Rheinländer, D. Samperi, M. Schweizer, and C. Stricker, *Exponential hedging and entropic penalties*, Mathematical Finance **12** (2002), no. 2, 99–123.

[3] D. Duffie and H. R. Richardson, *Mean-variance hedging in continuous time*, The Annals of Applied Probability **1** (1991), no. 1, 1–15.

[4] M. R. Grasselli and T. R. Hurd, *Wiener chaos and the Cox-Ingersoll-Ross model*, Proceedings of The Royal Society of London. Series A. Mathematical, Physical and Engineering Sciences **461** (2005), no. 2054, 459–479.

[5] S. Levendorskii, *Pseudo-diffusions and quadratic term structure models*, Working paper, University of Texas, Austin, January 2004.

[6] O. L. Mangasarian, *Sufficient conditions for the optimal control of nonlinear systems*, SIAM Journal on Control and Optimization **4** (1966), no. 1, 139–152.

[7] T. Rheinlaender, private communication, 2004.

[8] R. Rouge and N. El Karoui, *Pricing via utility maximization and entropy*, Mathematical Finance **10** (2000), no. 2, 259–276.

[9] P. Sannutti, *Singular perturbation method in the theory of optimal control*, Report R-379, University of Illinois, Illinois, 1966.

[10] H. Schellhorn, *Optimal changes of Gaussian measures, with an application to finance*, Under second round of revision with Applied Mathematical Finance, 2006.

[11] M. Schweizer, *Approximation pricing and the variance-optimal martingale measure*, The Annals of Probability **24** (1996), no. 1, 206–236.

Henry Schellhorn: School of Mathematical Sciences, Claremont Graduate University, Claremont, CA 91711, USA
E-mail address: henry.schellhorn@cgu.edu