Aalto University
School of Science
Degree programme in Systems and Operations Research

Teemu Seeve

A Method for Estimating Selection and Systematic Biases in Project Portfolio Selection

Instructor

D.Sc. Eeva Vilkkumaa

MS-E2108 Independent Research Projects in Applied Mathematics
February 17, 2016

Espoo

Contents

1 Introduction

Organizations typically consider more projects proposals than there are resources for pursuing them. Consequently, organizations are faced with the problem of selecting the subset or *portfolio* of projects that best matches their preferences. Often the aim is to minimize the costs of the portfolio, such that this portfolio fulfills some specified value requirements. For example, municipalities may need to organize their legitimate duties, such as education and health care, with the lowest possible expenses.

At the time of selection, the projects' costs are typically uncertain. Hence, the selection decision is made based on the projects' estimated costs. However, generally the ex post costs of the selected projects differ from their ex ante estimates. In particular, studies on real data have shown that the selected projects often cost more than anticipated (Flyvberg et al. 2002;Siemiatycki 2009; Jørgensen 2013).

Many studies suggest that this occurs due to a downward *systematic bias* in the projects' cost estimates. For instance, Flyvberg et al. (2002) show that the average cost overrun of 258 large transportation infrastructure projects is 27.6%. They explain these overruns by strategic misrepresentation, i.e., lying, motivated by economical or political agendas of the project promoters. Siemiatycki (2009) also explains the cost overruns found in transportation infrastructure projects by systematic bias. He finds that governmental audits tend to explain the systematic bias in cost estimation by technical difficulties of delivering large, complicated projects.

Yet, even if there is no systematic bias in the projects' cost estimates, cost overruns are expected to occur due to *selection bias*. Selection bias occurs, because the costs of the projects that have been selected based on low estimated costs are likely to have been underestimated. As a result, the cost estimates of the selected projects are biased, even though the set of all the cost estimates would be unbiased. This bias was first noted in the financial literature by Brown in his note in 1974, and then described more formally by Harrison and March (1984). More recent studies have been made by, e.g., Vilkkumaa et al. (2014), Begg and Bratvold (2008), and Smith and Winkler (2006) and Jørgensen (2013).

Both causes for cost overruns are undesired. However, due to the different nature of these causes, the two biases should be mitigated in a different fashion. For example, Vilkkumaa et al. (2014) show how selection bias can be diminished with Bayesian modeling. On the other hand, Flyvberg et al.

(2002) suggest that the mitigation of systematic bias in cost estimation could be possible with regulations and penalties for consistent estimation errors. Thus, the mitigation of the biases in the estimates requires knowledge about the relative amounts of the two biases.

The relative magnitudes of the biases could be estimated by building a statistical model between the costs and the cost estimates, the parameters of which could be estimated from empirical data on the realized and estimated costs. This approach, however, is problematic because the realized costs are only observed for the implemented projects. Thus, the data of the estimated and realized costs is *incomplete*, and consequently ordinary maximum likelihood (ML) estimation of the model parameters is not possible.

ML parameter estimation from an incomplete data set is possible by utilizing the Expectation Maximization (EM) algorithm, introduced by Dempster et al. (1977). This algorithm consists of two steps: *expectation step* and *maximization step*. The expectation step computes the expected values of the missing data given the current estimates for the parameter values. In the maximization step, the ML-estimates of the model parameters are updated by using the complete data, which consists of the incomplete data and the expected values of the missing data. These two steps enable iterative computation of the ML-estimates of the parameters with a guaranteed convergence for regular exponential families (see, e.g., Wu (1983)).

In this study we develop a statistical model between the costs and cost estimates, which is based on the work of Vilkkumaa and Liesiö (2015). The parameters of this model can be used to estimate the relative magnitudes of selection and systematic biases. To estimate these parameters, we formulate the EM-algorithm for project portfolio selection when the portfolio is selected with the aim of minimizing costs. Then, we analyze by simulation how well the parameters estimated with the algorithm can be used to compute the relative magnitudes of the two biases. In particular, we demonstrate how the number of estimated projects and the relative amount of missing data affects the estimation of the biases.

The rest of the study is structured as follows. In section 2, we present the statistical model for the true and estimated costs and formulate the biases. In section 3, we show how to apply the EM-algorithm to the model. In section 4, we apply the algorithm to empirical data by Ohlsson et al. (1998, 1999) to obtain realistic parameter values for testing the model. We then use the estimated parameters in section 5 for validating algorithm with Monte Carlo simulation. Finally, section 6 concludes.

2 Model framework

Consider N project candidates, out of which the portfolio will be selected. The true costs of these projects are $c = [c_1, \ldots, c_N]$. These costs are assumed to be realizations of independent lognormally distributed random variables $C_i \sim \text{LogN}(\mu, \sigma^2)$, $i = 1, \ldots, N$. Lognormal distribution has a non-negative domain, which includes arbitrary high values, and is thus suitable for modeling costs.

Cost estimates $c^E = [c_1^E, \ldots, c_N^E]$ are modeled as realizations of conditionally independent random variables $(C_i^E | C_i = c_i) = \Delta c_i$, where $\Delta \sim \text{LogN}(\eta, \tau^2)$. This model is justified by, e.g., the shape of the data of cost escalations in Flyvberg et al. (2002), which resembled lognormal distributions. Moreover, a multiplicative error model for the cost estimates reflects the fact that errors in cost estimates are often proportional to the true costs of the projects (Keisler 2004).

Because the projects' true costs $c = [c_1, \ldots, c_N]$ are unknown ex ante, the optimal portfolio is selected based on the cost estimates $c^E = [c_1^E, \ldots, c_N^E]$. We denote the selected project portfolio with a binary vector $z = [z_1, \ldots, z_N]^T \in \{0, 1\}^{N \times 1}$, where $z_i = 1$ if and only if the project i is selected in the portfolio and $z_i = 0$ otherwise. Using this notation, the optimal project portfolio can be obtained as a solution to the problem

$$z\left(c^E\right) = \underset{z \in \mathcal{Z}}{\arg\min}\, c^E \cdot z, \tag{1}$$

where $\mathcal{Z} = \{z | g(z) \leq \mathbf{0}\}$ is the set of feasible portfolios. If, for instance, the value of the selected portfolio $v(z)$ is required to exceed some threshold v^*, then $\mathcal{Z} = \{z | v^* - v(z) \leq \mathbf{0}\}$.

Biases

The conditional expected values $\mathbb{E}\left[C_i^E | C_i = c_i\right]$ of the cost estimates $C^E = [C_1^E, \ldots, C_N^E]$ do not necessarily coincide with the true costs $C = [C_1, \ldots, C_N]$, in which case the cost estimates are biased. Formally, there is a systematic bias (SyB) in cost estimate C_i^E if and only if $\mathbb{E}\left[C_i^E | C_i = c_i\right] = \mathbb{E}[\Delta] c_i = e^{\eta + \frac{1}{2}\tau^2} c_i \neq c_i$, i.e., when $\eta + \frac{1}{2}\tau^2 \neq 0$.

Given η and τ^2, the cost estimates can be debiased by multiplying them by factor $e^{-\eta - \frac{1}{2}\tau^2}$. The debiased cost estimates form a vector $\tilde{c}^E = [\tilde{c}_1^E, \ldots, \tilde{c}_N^E]$, where $\tilde{c}_i^E$ are realizations of random variables $\tilde{C}_i^E = e^{-\eta - \frac{1}{2}\tau^2} C_i^E$. Because $\mathbb{E}\left[\tilde{C}_i^E \middle| C_i = c_i\right] = e^{-\eta - \frac{1}{2}\tau^2} e^{\eta + \frac{1}{2}\tau^2} c_i = c_i$ for all $i = 1, \ldots, N$, these estimates are indeed unbiased.

In general, using the debiased cost estimates $\tilde{c}^E$ can yield different optimal portfolios than the initial cost estimates c^E. Nevertheless, since the debiasing factor $e^{-\eta-\frac{1}{2}\tau^2}$ is constant for all cost estimates, replacing the possibly biased c^E with the debiased estimates in (1) yields the same optimal portfolio. Thus, we note that $z\left(\tilde{c}^E\right) = z\left(c^E\right)$.

In addition to systematic bias, the cost estimates are expected to differ from the true costs due to selection bias (SeB), which occurs when cost underestimation increases the chances of selecting a project. The expected relative cost overrun of a project portfolio - referred to as the *total bias* (TB) - thus consists of both systematic and selection biases. These biases are defined more formally below.

Definition 1 *Let C, C^E and $\tilde{C}^E$ be the true, estimated and debiased estimated costs, respectively. Let $z\left(C^E\right)$ be defined as in 1, but with random cost estimates C^E. The total bias (TB), selection bias (SeB) and systematic bias (SyB) are defined as*

$$\begin{aligned} TB &= \frac{C_{portfolio} - C^E_{portfolio}}{C^E_{portfolio}} = \frac{C \cdot z\left(C^E\right) - C^E \cdot z\left(C^E\right)}{C^E \cdot z\left(C^E\right)} \\ SeB &= \frac{C_{portfolio} - \tilde{C}^E_{portfolio}}{\tilde{C}^E_{portfolio}} = \frac{C \cdot z\left(C^E\right) - \tilde{C}^E \cdot z\left(C^E\right)}{\tilde{C}^E \cdot z\left(C^E\right)} \\ SyB &= TB - SeB. \end{aligned} \tag{2}$$

From the above formulas it can be seen, that positive values of the biases correspond cost underestimation, referred to as *downward* biases in the estimation. Correspondingly, the negative values and overestimation will be referred to as *upward* biases.

3 Parameter estimation

The statistical model presented in the previous section includes four parameters, $\theta = [\mu, \sigma^2, \eta, \tau^2]$. Would the data (c^E, c) be complete, the Maximum Likelihood (ML) estimates of these parameters could be computed in a conventional way. However, those projects that are not selected in the portfolio are never carried out, and consequently there is no data of their true costs. Thus, the data of the true and estimated costs is incomplete, and more elaborate methods to compute the ML-estimates for the parameters θ are needed.

We present an Expectation Maximization algorithm to compute the ML-estimates of the model parameters from incomplete data. The algorithm is initialized by computing the initial values for the model parameters from the incomplete data. In the Expectation step, the expected values of the missing data are computed, conditioned on the incomplete data and the current parameter values. In the Maximization step, the algorithm computes the conventional ML-estimates from completed data, which consist of the incomplete data and the expected values computed in the previous step. The parameter values are then updated for the following iteration, using the newly computed ML-estimates. These steps will be repeated until a specified tolerance level is met.

The EM-algorithm is especially useful when the distribution of the complete data comes from an exponential family, as the lognormal distribution does. For exponential families, deriving a closed formula for both the steps of the algorithm is possible, thus much alleviating computation.

Before moving on to the formal definition of the algorithm for our statistical model, some of the used notations should be clarified. We denote the indices of the $M \leq N$ selected projects by I^* and those of the $N - M$ non-selected projects by I°. Using these notations, the logarithms of the observed true costs of the selected projects are $\ln c^* := \{\ln c_i | i \in I^*\}$. Similarly, the logarithms of the non-observable costs of the non-selected projects are $\ln c^\circ := \{\ln c_i | i \in I^\circ\}$. For notational convenience, we denote the data as sets instead of vectors.

Initialization of the algorithm: *Set the initial values of the parameters* $\theta = \hat{\theta}_0 = \left[\hat{\mu}_0, \hat{\sigma}_0^2, \hat{\eta}_0, \hat{\tau}_0^2\right]$ *as*

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{N}\sum_{i=1}^{N} \ln c_i^E - \hat{\eta}_0, & \hat{\sigma}_0^2 &= \frac{1}{M-1}\sum_{i\in I^*}\left[\ln^2 c_i - \left(\frac{1}{M}\sum_{i\in I^*}\ln c_i\right)^2\right] \\ \hat{\eta}_0 &= \frac{1}{M}\sum_{i\in I^*}\left(\ln c_i^E - \ln c_i\right) & \hat{\tau}_0^2 &= \frac{1}{M-1}\sum_{i\in I^*}\left(\ln^2 c_i^E - 2\ln c_i^E \ln c_i + \ln^2 c_i - \hat{\eta}_0^2\right). \end{aligned} \tag{3}$$

Expectation step:
Given that $\theta = \hat{\theta}_k = \left[\hat{\mu}_k, \hat{\sigma}_k^2, \hat{\eta}_k, \hat{\tau}_k^2\right]$*, compute the conditional expected values for the missing data* $\ln c^\circ = \left\{\mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right] \middle| i \in I^\circ\right\}$ *and* $\ln^2 c^\circ = \left\{\mathbb{E}\left[\ln^2 C_i \middle| \ln C_i^E = \ln c_i^E\right] \middle| i \in I^\circ\right\}$*, with*

$$\begin{aligned} \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right] &= \frac{\hat{\tau}_k^2}{\hat{\tau}_k^2 + \hat{\sigma}_k^2}\hat{\mu}_k + \frac{\hat{\sigma}_k^2}{\hat{\tau}_k^2 + \hat{\sigma}_k^2}\left(\ln c_i^E - \hat{\eta}_k\right) \\ \mathbb{E}\left[\ln^2 C_i \middle| \ln C_i^E = \ln c_i^E\right] &= \frac{\hat{\tau}_k^2\hat{\sigma}_k^2}{\hat{\tau}_k^2 + \hat{\sigma}_k^2} + \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]^2. \end{aligned} \tag{4}$$

Maximization step:
Using the complete data $\left(\ln c^E, \ln c^*, \ln c^\circ, \ln^2 c^*, \ln^2 c^\circ\right)$*, compute the ML-estimates* $\hat{\theta}_{k+1} = \left[\hat{\mu}_{k+1}, \hat{\sigma}^2_{k+1}, \hat{\eta}_{k+1}, \hat{\tau}^2_{k+1}\right]$ *from*

$$\hat{\mu}_{k+1} = \frac{1}{N}\sum_{i=1}^{N} \ln c_i, \qquad \hat{\sigma}^2_{k+1} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\ln^2 c_i - \hat{\mu}^2_{k+1}\right) \tag{5}$$

$$\hat{\eta}_{k+1} = \frac{1}{N}\sum_{i=1}^{N} \ln c_i^E - \hat{\mu}_{k+1} \quad \hat{\tau}^2_{k+1} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\ln^2 c_i^E - 2\ln c_i^E \ln c_i + \ln^2 c_i - \hat{\eta}^2_{k+1}\right).$$

If $\left\|\hat{\theta}_{k+1} - \hat{\theta}_k\right\| < \delta$*, terminate the algorithm. If not, set* $k := k+1$ *and* $\theta = \hat{\theta}_{k+1}$ *and return to the Expectation step with the new* θ*.*

Detailed derivations of the steps of the algorithm are in the appendix.

4 Application to student project data

In this section we apply the EM-algorithm to empirical data. We use data from a software development course held at the department of Communication Systems in Lund University. The data was initially reported by Ohlsson et al. (1998) and Ohlsson and Wohlin (1999), who studied ways to improve the students' ex ante effort estimates.

Data characteristics

The data consists of estimated and realized efforts of student projects. Because every student project was carried out, the data is complete, i.e., the true efforts are observed for each project. This is useful, because (i) it enables estimating θ with the algorithm using different amounts of implemented projects M and (ii) the estimation results can be compared to conventional ML-estimates computed from the complete data.

The software development course would be held annually, and the data presented in Ohlsson et al. (1998) and Ohlsson and Wohlin (1999) consists of courses from three different years (1995, 1996 and 1998). During these years, the projects and groups implementing them would not vary significantly; nearly identical projects would be carried out by different groups that were made as equal in size and expertise as it is possible in an educational environment. The data on the different groups' true and estimated efforts is presented in Table 1.

Table 1: The data of student software development projects from Ohlsson et al. (1998) and Ohlsson and Wohlin (1999)

Data	Estimate (h)			Outcome (h)			Δ=Estimate/Outcome		
i/year	1995	1996	1998	1995	1996	1998	1995	1996	1998
1	1182	1346	704	963	977	734	1.23	1.38	0.959
2	886	1415	908	599	1385	665	1.48	1.02	1.37
3	755	1172	1200	745	958	895	1.01	1.22	1.34
4	690	1010	868	810	1021	792	0.852	0.989	1.10
5	965	1183	986	1030	1265	1217	0.937	0.935	0.810
6	1001	1200	956	1000	1630	1332	1.00	0.736	0.718
7	1004	1096	776	894	951	739	1.12	1.15	1.05
8			816			736			1.11
9			1443			1500			0.962
10			1364			1017			1.34
11			1232			1189			1.04
12			796			958			0.831

To check whether the data from three different years could be combined to compose one large sample, we tested that the variances and the means of the data did not vary annually. Because our statistical model assumes lognormal random variables and the used statistical tests require normal data, the tests were made for the logarithms of the data.

First, Bartlett's test was used to test the homogeneity of the variances. In this test, the null hypothesis of equal population variances across all the k populations $H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$ is tested against the alternative hypothesis of unequal variances between at least two populations $H_1 : \exists\, i, j$ s.t. $\sigma_i^2 \neq \sigma_j^2$. Table 2 shows that it is likely to obtain these kinds of samples when the null hypothesis stands, because the p-values are large and $\chi^2 < \chi^2_{critical}$ for both the effort outcome and relative estimation error Δ. Thus, Bartlett's test indicates that the variances of the logarithms of the samples from three years are equal at $\alpha = 0.05$ significance level.

Table 2: Results from the Bartlett's test for equal variances of the logarithms of the data in Table 1.

data/variable	χ^2	p-value	$\chi^2_{critical}$ $(\alpha = 0.05, df = 23)$
ln Outcome	0.892	0.640	35.2
$\ln \Delta$	0.106	0.949	35.2

Then, the equality of the means of the annual data were tested with One-way ANOVA, which tests differences among population means in terms of a single factor (year of the course). The null hypothesis of ANOVA is equal means of all k populations $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, and the alternate hypothesis is that at least two populations have different means $H_1 : \exists i, j \text{ s.t. } \mu_i \neq \mu_j$. The results of the one-way ANOVA are presented in Table 3. There is no statistically significant difference between the logarithms of the outcomes and Δs of different years, because $F < F_{critical}$ for both of the sample sets. Thus, because the variances and means of the data from different years have no statistically significant difference, we could combine the data of different years and treat them as one large sample.

Table 3: Results of the one-way ANOVA for the student data logarithms.

data/variable	F	p-value	$F_{critical}$ $(\alpha = 0.05, (df_1, df_2) = (2, 23))$
ln Outcome	2.86	0.078	3.42
$\ln \Delta$	0.092	0.912	3.42

Table 4: Results of Shapiro-Wilk test for the data logarithms.

data/variable	W	p-value	$W_{critical}$ $(\alpha = 0.05, n = 26)$
ln Outcome	0.974	0.738	0.920
$\ln \Delta$	0.974	0.740	0.920

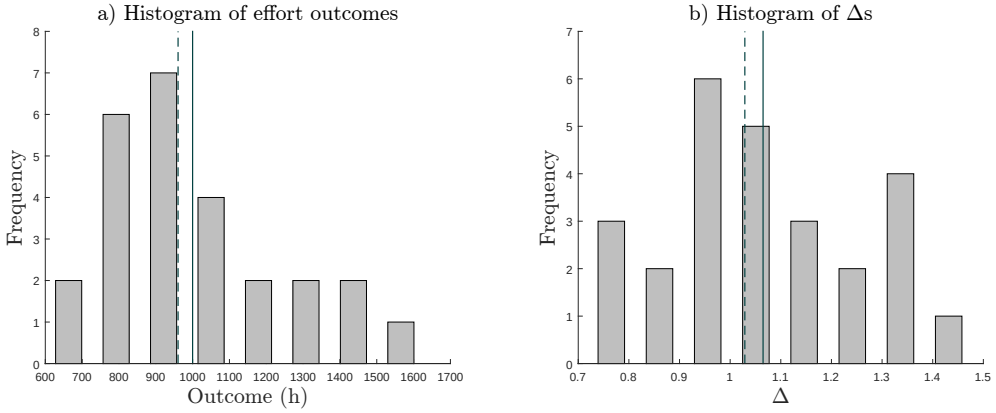

Figure 1: The histograms of the student project data. The artihmetic means of the samples are marked with solid and medians with dashed lines.

We also checked the lognormality of the data with with Shapiro-Wilk normality test, which tests the null hypothesis of a normally distributed population

(see Table 4). Because $W > W_{critical}$ and consequently $p > 0.05$ for both data sets, the null hypothesis is not rejected and the lognormality assumption stands. The lognormality of the data was also confirmed by the histograms of effort outcomes and estimation errors shown in Figure 1.

Application of the EM-algorithm

Having confirmed that the student data is lognormal and can be used as one large sample, we estimated the parameters θ with the algorithm from the presented data. To do this, we considered portfolio selection processes (as in (1)) with one feasibility constraint that limits the number of selected projects: $\mathcal{Z} = \left\{ z \middle| pN - \sum_{i=0}^{N} z_i = \mathbf{0} \right\}$, where $p = M/N \in (0, 1]$. This is useful, because

(i) selection bias is most distinctive when projects are selected solely based on their estimated costs (see, e.g., Jørgensen 2013), and

(ii) we are essentially interested in the effect of the amount of missing data on the estimation precision, and the constraint of fixing the amount of selected projects is the simplest one for this purpose.

In addition, this kind of project portfolio selection fits well to the student data where all projects can be considered to provide equal value, and thus it would be reasonable to select a portfolio of the student projects based the mere effort (cost) estimates.

The results of the parameter estimation are presented in Table 5. Based on this table, the values of $\hat{\mu}$ vary relatively little. However, the rest of the parameters fluctuate more; for example, $\hat{\eta}$ varies to ten times greater value when p is increased from 0.1 to 0.3. More implemented projects and larger p and M yield faster convergence of the algorithm and seemingly better estimates of the parameters as compared to the complete data ML-estimates computed with $p = 1$.

Table 5: Parameters θ estimated with EM-algorithm from the student data and the iterations required for convergence at tolerance level $\delta = 10^{-}7$.

(p, M)/parameter	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\eta}$	$\hat{\tau}^2$	iterations
(0.1, 3)	6.90	0.0351	0.0207	0.0111	780
(0.3, 8)	6.71	0.0185	0.207	0.0387	126
(0.5, 13)	6.85	0.0515	0.0734	0.0383	36
(0.7, 18)	6.84	0.0450	0.0824	0.0356	22
(0.9, 23)	6.87	0.0588	0.0562	0.0397	10
(1, 26)	6.88	0.0648	0.0451	0.0373	1

Interestingly, the effort estimates of the students were pessimistic; the effort estimates were on average larger than the true efforts, because $\hat{\eta}+\frac{1}{2}\hat{\tau}^2 > 0$ with all the estimated parameters. This can perhaps be explained by incentives to overestimate effort in a noncompetitive student environment. Unlike in competitive settings, the student projects will be carried out regardless of the estimated effort. Being pessimistic about the estimated effort might pay off, since overestimating effort would lead to excessive available time and consequently ease the implementation of the project, which could be of some of the students' interests.

To see how well the estimated parameter values fit the data for different values of p, lognormal distributions corresponding to $p = 0.1$, $p = 0.5$ and $p = 1$ were plotted in the same graph with the histograms of the complete data (in Figure 2). With both $p = 0.1$ and $p = 0.5$, the distributions of the effort outcomes are close to the distribution using the complete data ML-estimates of the parameters with $p = 1$. While the distribution of the relative estimation errors Δ is rather narrow when p is very small 0.1, when $p = 0.5$ also the distribution of Δ is really close to such distribution with $p = 1$. Thus, even with half of the true efforts missing, we were able to estimate distributions nearly identical to the best-case scenario when $p = 1$.

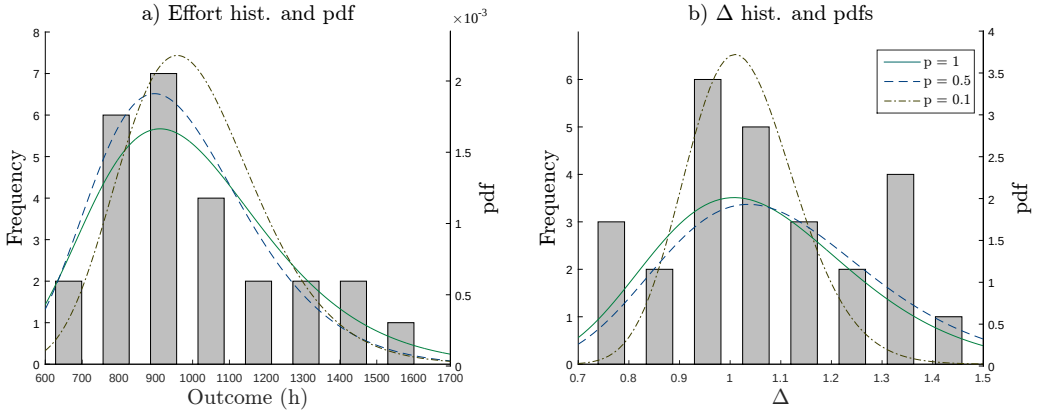

Figure 2: The histograms of the student project data and the estimated probability density functions of C_i and Δ.

5 Algorithm validation with simulation

Due to the small sample size and the fact that only one problem instance was considered, no statistically significant results on the performance of the EM-algorithm can be drawn from the results presented in the previous section. Furthermore, to compare the estimated and true values of the parameters θ, knowledge about the true values of the parameters is required.

We use Monte Carlo to systematically test the performance of the algorithm. Simulation allows generating data from a known distribution, thus enabling thorough analysis of the estimates by comparison with the true values. In keeping with the previous section, we simulate the selection of $p = M/N$ projects with the lowest estimated cost. The simulation proceeds as follows.

(i) Initialization

- Set the true values to the parameters $\theta = [\mu, \sigma^2, \eta, \tau^2]$.
- Select the amount N of project proposals and the relative share p of selected projects.

(ii) Data generation

- Generate N project costs c_i from $C_i \sim \text{LogN}(\mu, \sigma^2), i = 1, \ldots, N$.
- Generate N Δs from $\Delta \sim \text{LogN}(\eta, \tau^2)$ and compute the corresponding estimated costs $c_i^E = \Delta c_i, i = 1, \ldots, N$.

(iii) Portfolio selection

- Select $M = pN$ projects with the lowest costs to the portfolio.
- Divide the indices to those of the selected projects I^* and those of the non-selected projects I°.

(iv) Parameter estimation

- Execute the EM algorithm with the incomplete data $(\ln c^E, \ln c^*)$, and obtain the estimated model parameters $\hat{\theta} = [\hat{\mu}, \hat{\sigma}^2, \hat{\eta}, \hat{\tau}^2]$.
- Compute the different biases through (2).

(v) Store the results and return to step (ii).

Realistic parameter values for the validation of the algorithm were estimated from student data in the previous section. Because the earlier literature suggests that the systematic bias can generally be expected to be downward (Flyvberg et al. 2002; Siemiatycki 2009; Jørgensen 2013), a data set with a

downward systematic bias ($\eta + \frac{1}{2}\tau^2 < 0$) is included by changing the sign of parameter η. Table 6 presents such parameter values used in the simulation.

Table 6: The values of the parameters θ used for data generation in the simulation.

value set\parameter	μ	σ^2	η	τ^2
1	6.88	0.0648	0.0451	0.0373
2	6.88	0.0648	-0.0451	0.0373

Different values for the amount N of project proposals and the relative share p of selected projects are used. We analyze the effect of varying p from 0.1 to 1 and N from 20 to 1000.

For each p and N, project portfolio selection is simulated 10000 times. The expected values of the parameters θ estimated with the EM-algorithm are computed as arithmetic means of the estimated parameters $\hat{\theta}$ from the 10000 simulation rounds, and these means will be compared to the true values of the parameters. The expected values of the biases are studied in a similar fashion; the biases are computed with the true and the estimated parameter values, and the means of these biases are then compared.

Results with upward systematic bias

Figure 3 presents the mean estimates of the model parameters when $\eta + \frac{1}{2}\tau^2 > 0$. Increased relative amount of implemented projects p was observed to increase estimation precision. When $p \geq 0.5$, the estimates were fairly accurate, besides $\hat{\tau}^2$ with small N. The average estimates differed more from the true parameter values when p was decreased, especially with small N. When $p = 0.1$ and $N = 20$, we observed a systematic overestimation of the parameters μ and σ^2, and a systematic underestimation of the parameters η and τ^2. However, as the number of the estimated projects was larger ($N \geq 150$) the bias was changed to the overestimation of η and τ^2 and underestimation of μ and σ^2.

Table 7 illustrates the relative errors $\mathrm{Err}\left[\hat{x}\right] = 100\% \times \left(x - \mathrm{Avg}[\hat{x}]\right) / \mathrm{Avg}\left[\hat{x}\right],\ x \in \{\mu, \sigma^2, \eta, \tau^2\}$ of the means $\mathrm{Avg}\left[\cdot\right]$ of the estimates for a medium sample size of $N = 150$. It can be seen that the estimates of μ and σ^2 are fairly accurate, but the estimates of η and τ^2 are slightly worse.

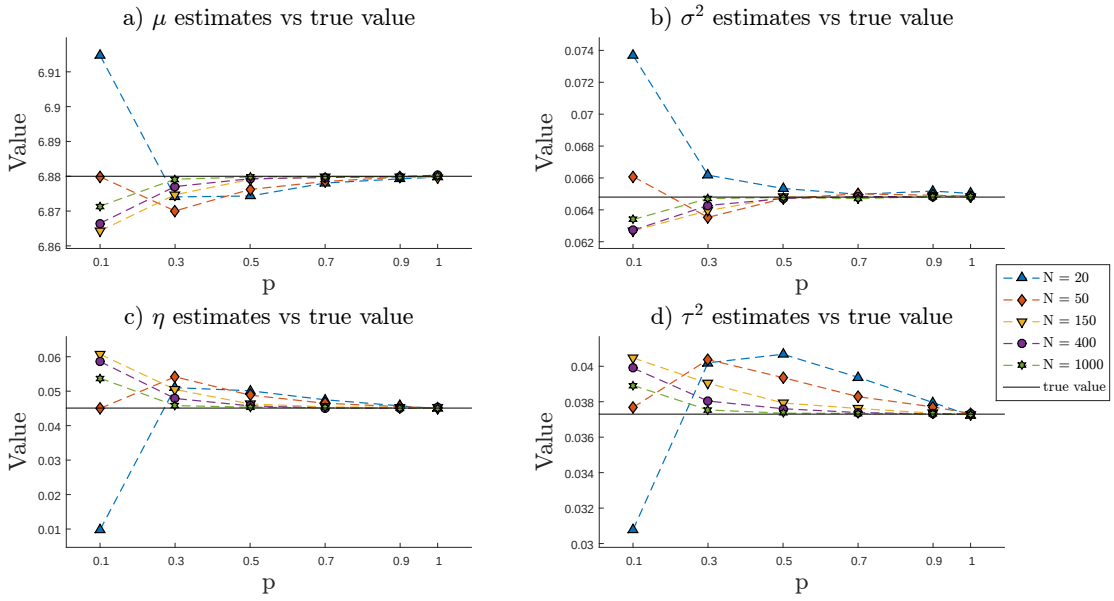

Figure 3: Average values of parameter estimates for different p and N, when $\eta + \frac{1}{2}\tau^2 > 0$. 10000 simulation rounds.

Table 7: The relative error of the means of each model parameter estimate, when $N = 150$ and p varies.

parameter$\backslash p$	0.1	0.3	0.5	0.7	0.9	1
Err $[\hat{\mu}]$ (%)	0.23	0.077	0.013	0.0021	0.0033	0.0050
Err $[\hat{\sigma}^2]$ (%)	3.4	1.3	0.0020	0.051	-0.11	-0.015
Err $[\hat{\eta}]$ (%)	-26	-11	-2.6	-0.75	-0.43	-0.21
Err $[\hat{\tau}^2]$ (%)	-7.8	-4.5	-1.7	-0.85	-0.15	0.026

Selection and systematic biases were computed with the estimated and true values of the parameters θ through (2) during each simulation round. The averages of these biases are presented as stacked columns in Figure 4, where the two biases are distinguished by color and the total bias is shown at the top of each bar and marked with a black bar.

A smaller share p of selected projects leads to an overestimation of selection bias. Because selection bias is computed as $SeB = C_{portfolio}/\tilde{C}^E_{portfolio} - 1$, where $\tilde{C}^E_{portfolio} = e^{-\hat{\eta}-\frac{1}{2}\hat{\tau}^2} C^E \cdot z\left(C^E\right)$, the overestimation of selection bias results from overestimation of $\hat{\eta}$ and $\hat{\tau}^2$, which can be observed in Figure 3 and Table 7. Nevertheless, the bias estimation results are rather satisfactory when $p \geq 0.5$,

since the biases computed with the estimated parameters' values are close to the biases computed with the true values. Thus, the EM-algorithm computes fairly accurate estimates of the magnitudes of selection and systematic biases, provided that p is sufficiently large (when $N = 150$).

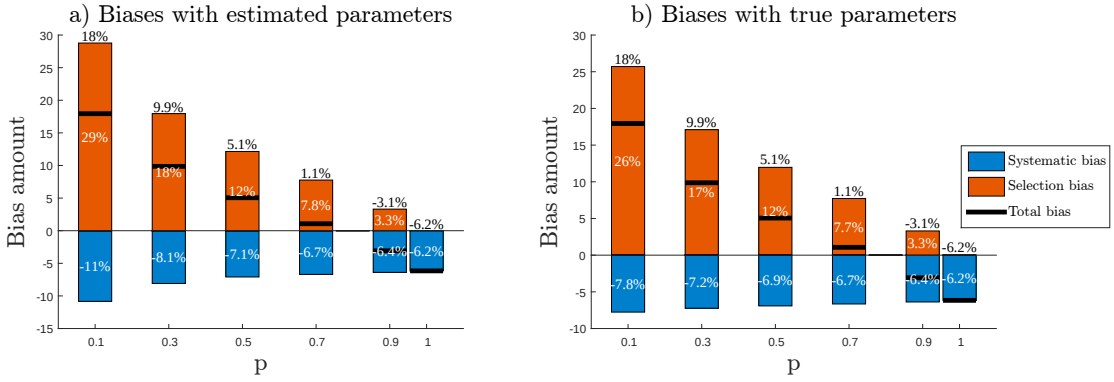

Figure 4: Average biases, when $\eta + \frac{1}{2}\tau^2 > 0$ and $N = 150$.

Results with downward systematic bias

A similar simulation was made with negative η to study the performance of the algorithm when there is a downward systematic bias in the cost estimates. Figure 5 presents the average parameter estimates of the EM-algorithm with such downward systematic bias. The estimates are nearly identical to the estimated parameters corresponding to Figure 3. Increasing p yields better estimates of θ, and decreasing p results in larger difference between the mean estimates and true values of θ, the difference being more drastic when N is also small. When $N = 150$, the relative errors of the parameters coincide with the values in table 7, and are not be presented here for that reason.

The average downward biases in the cost estimates were then computed using the estimated and true parameter values. As seen in Figure 6, the overestimation of η and τ^2 leads to minor overestimation of selection bias, and consequently systematic bias is slightly smaller when computed with the estimated parameters. The estimation precision again increases when p is increased.

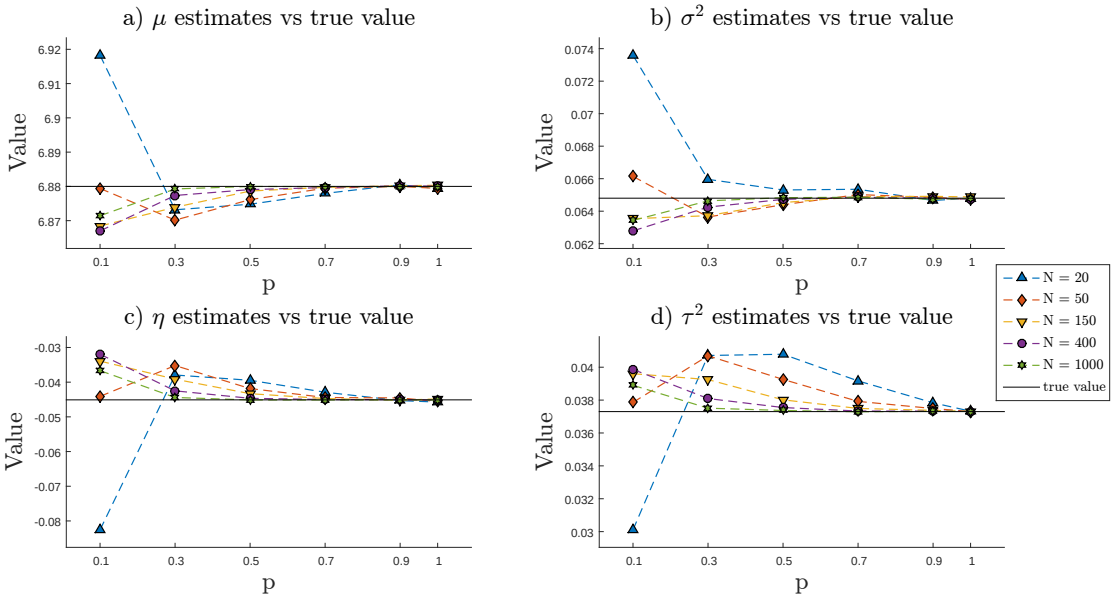

Figure 5: Average values of parameter estimates for different p and N, when $\eta + \frac{1}{2}\tau^2 < 0$. 10000 simulation rounds.

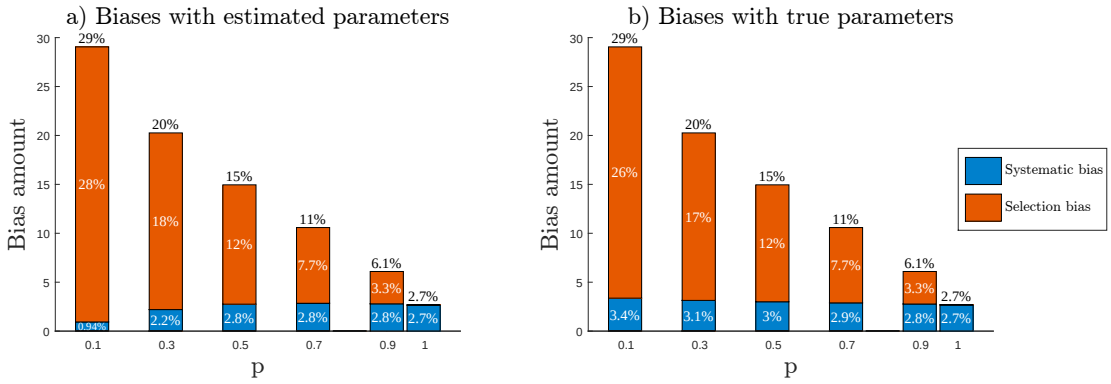

Figure 6: Average biases, when $\eta + \frac{1}{2}\tau^2 < 0$ and $N = 150$.

Table 8 presents the relative magnitudes of average systematic and selection biases, computed as $SyB_{rel} = 100\% \times \text{Avg}\,[SyB]\,/\,\text{Avg}\,[TB]$ and $SeB_{rel} = 100\% \times \text{Avg}\,[SeB]\,/\,\text{Avg}\,[TB]$, respectively. It can be seen that the relative magnitudes of the two biases can be computed relatively accurately using EM-algorithm, when p is large enough. While the errors in the relative amounts of the biases are somewhat large when $p = 0.1$, the rather accurate parameter estimates

with larger p provide good estimates for the biases. These estimates even coincide with the true values within the used working precision, when $p = 0.7$.

Table 8: Relative biases for different values of p, when $\eta + \frac{1}{2}\tau^2 < 0$ and $N = 150$.

bias	parameters θ/p	0.1	0.3	0.5	0.7	0.9	1
$SyB_{rel}(\%)$	estimated	3.2	11	18	27	46	100
	true	12	15	20	27	45	100
$SeB_{rel}(\%)$	estimated	97	89	82	73	54	0.36
	true	88	85	80	73	55	-0.19

Estimation precision dependence on N

The dependence of the parameter estimates on N are shown in Figure 7. Based on this Figure, increasing the number of projects proposals N results in better estimation precision on average. This increase becomes more distinctive as the relative amount of the selected projects p decreases. We note that the series of $p = 0.1$ crossing the true value when $N = 50$ is consistent with the results for constant p in Figures 5 and 3, where the biases of the parameter estimates changed their signs when N was increased at $p = 0.1$.

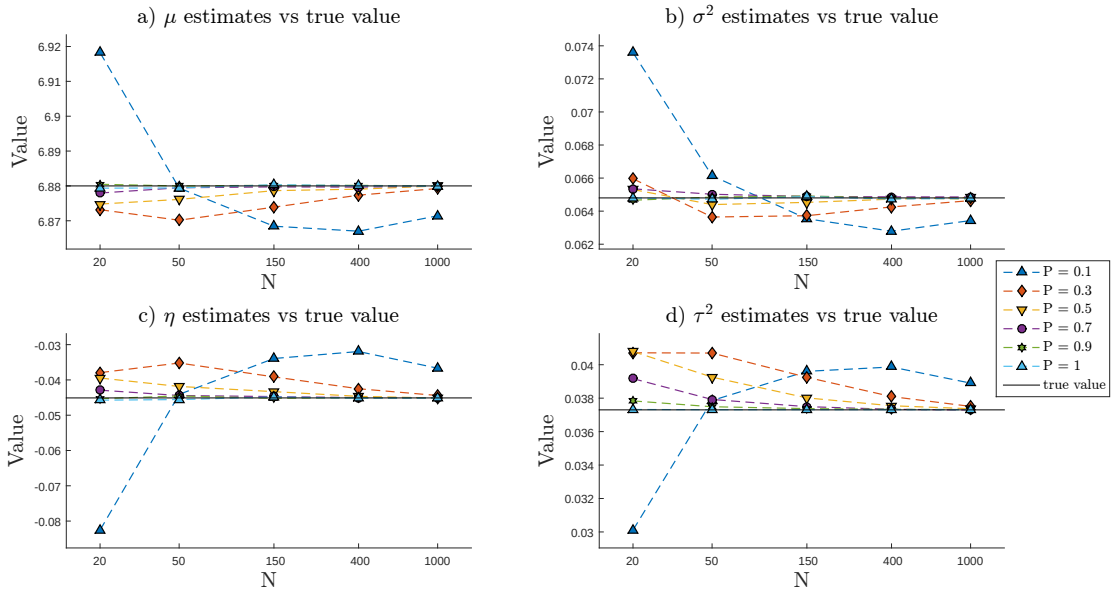

Figure 7: Average values of parameter estimates for different p and N, when $\eta + \frac{1}{2}\tau^2 < 0$. 10000 simulation rounds.

The dependence of the average estimated biases on N is illustrated in Figure 8. Table 9 then presents the relative magnitudes of these biases, computed similarly as in Table 8.

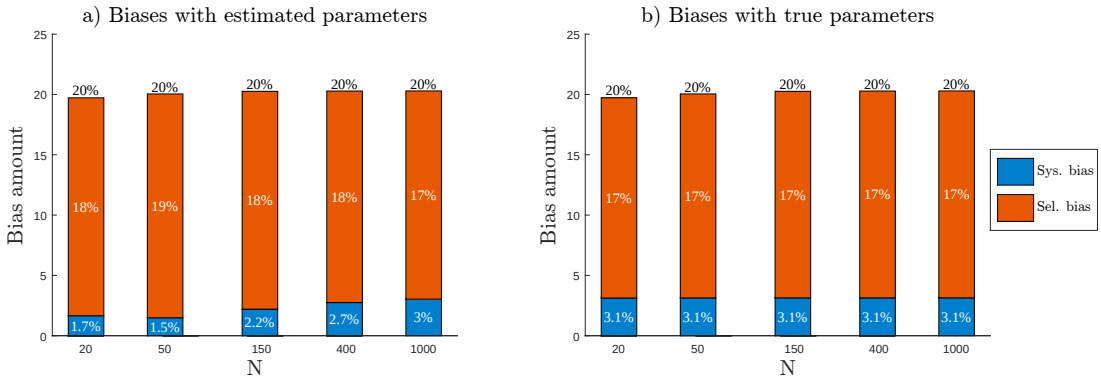

Figure 8: Average biases, when $\eta + \frac{1}{2}\tau^2 < 0$ and $p = 0.3$.

Table 9: Relative biases for different values of p, when $\eta + \frac{1}{2}\tau^2 < 0$ and $p = 0.3$.

bias	parameters θ/N	20	50	150	400	1000
$SyB_{rel}(\%)$	estimated	8.4	7.4	11	14	15
	true	16	16	15	15	15
$SeB_{rel}(\%)$	estimated	92	93	89	86	85
	true	84	84	85	85	85

Figure 8 and Table 9 show that when $p = 0.3$, a large amount of project proposals N results in rather accurate estimates of the two biases. Table 9 shows that the relative average systematic and selection biases computed with estimated and true parameter values are within one percent, when $N \geq 400$. $N = 150$ yields biases within 5% of the true values, and $N \leq 50$ leads to larger, over 5%, errors in the estimated relative biases.

Bias estimation precision dependence on both p and N

Because the estimation precision of the EM-algorithm depends on both the amount N of projects proposals and the relative share p of selected projects, we examined the estimation precision when both of these values are varied.

We computed the differences between the true and estimated relative average selection biases $SeB_{rel} = 100\% \times \text{Avg}\,[SeB] \,/\, \text{Avg}\,[TB]$. The differences were

computed as $SeB_{rel} - \widehat{SeB}_{rel}$, where SeB_{rel} and $\widehat{SeB}_{rel}$ stand for the relative biases computed with the true and estimated parameters, respectively. Thus, positive values mean that selection bias is underestimated on average and negative values correspond to the overestimation of selection bias. Computing these differences for the systematic bias is not necessary, because such differences are simply the opposite value of that of the selection bias: if the amount of selection bias is underestimated, then the amount of systematic bias is overestimated by exactly the same amount, and vice versa.

The errors in the estimated relative biases are illustrated in Figure 9 below. In the Figure, the stacked columns on the left are the biases computed with the estimated parameters and the columns on the right correspond to the true biases. The areas encircled with black dots mark the differences between the biases computed with the estimated and true parameters. The percentages within the columns correspond to the relative magnitudes SyB_{rel} and SeB_{rel}, and the difference $SeB_{rel} - \widehat{SeB}_{rel}$ is shown within the encircled area. The absolute percentages of the biases can be read from the y-axis.

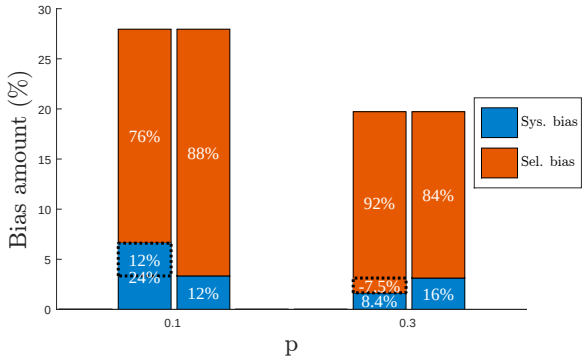

Figure 9: Differences of the relative biases, when $N = 20$.

Table 10 below presents the differences between the relative average biases for different values of p and N. Those cells that correspond to values of p and N for which the relative errors in the biases are less than 1% are colored with green. Yellow cells correspond to errors less than 5% but greater than 1%, and errors greater than 5% are marked with red cells.

Table 10: The differences between the relative average selection biases $SeB_{rel} - \widehat{SeB}_{rel}$, when $\eta + \frac{1}{2}\tau^2 < 0$.

N/p	0.1	0.3	0.5	0.7	0.9	1
20	12	-7.5	-6.0	-2.6	1.5	5.2
50	-5.5	-8.2	-3.4	-0.73	-0.48	3.0
150	-8.4	-4.6	-1.7	-0.39	0.32	-0.55
400	-7.9	-1.9	-0.44	-0.066	0.061	0.46
1000	-4.8	-0.53	-0.011	0.061	0.028	0.20

With a few exceptions, increasing either p or N yields better estimated relative average errors. Surprisingly, the conventional ML-estimation of the parameters θ when $p = 1$ seemingly results in worse estimated biases than EM-algorithm estimation with $p = 0.9$ and $p = 0.7$. One can also note the sign change of the difference in the relative biases when $p = 0.1$, and the relatively small error with this p and $N = 50$. This observation corresponds to the series of $p = 0.1$ crossing the true value when $N = 50$ in Figure 7.

Conclusion of the simulation results

The precision of the EM-algorithm in estimating the parameters of our project portfolio selection model was found to depend on (i) the relative share p of selected projects and (ii) the amount N of project proposals, as Figures 3, 5 and 7 illustrate. For example, Table 7 shows how the true values of the model parameters θ are all within 1% of the means of the estimated parameters when p is greater than 0.7, while the true values lie as far as 26% away from the mean estimates when p is 0.1 (when $N = 150$).

Using the estimated parameters, the computation of selection and systematic biases in project portfolio selection also depends on p and N, as is presented in Figures 4, 6 and 8. Table 10 shows that the average estimated biases are computed accurately, within 1% from the average true bias as percentages of the total bias, when both p and N are large. Decreasing either of these yields greater $1 - 5\%$ differences between the estimated and true biases. If both p and N are small, the errors in the estimated relative average biases are greater than 5%.

6 Conclusion

This study applies a method (presented in Vilkkumaa and Liesiö 2015) to compute the relative magnitudes of selection and systematic biases in decision settings, in which a project portfolio is selected based on uncertain estimates about the projects' costs. In this method, the parameters of the statistical model for the projects' true and estimated costs are estimated through an EM-algorithm using incomplete data. Using realistic model parameters estimated from a real case study, the performance of the presented method was analyzed with Monte Carlo simulation. This analysis was carried out by first comparing the average parameter estimates to the true values of the parameters. Then, the estimated average selection and systematic biases were compared to the true biases.

We found that it is possible to estimate the model parameters relatively accurately on average, when the share p of selected projects and the amount N of project proposals are large enough. Only the cases with both small p and N yielded rather poor average estimation precision. Following from the accurate parameter estimation, the average biases computed with the estimated parameters are close to those computed with the true parameter values, as long as p and N are sufficiently large. In particular, for each of the studied p, estimation precision of selection and systematic biases could be enhanced by increasing N. For example with a relatively small $p = 0.3$, the estimated biases were within 1% from the true biases, when $N = 1000$.

Because the average parameter estimates computed with the EM-algorithm differ from the true values when p and N were small, there remains room for improvement in the estimation precision. Future work concerning the improvement of the method presented in this study could head in two different directions. On the one hand, the presented method could be enhanced to improve estimation accuracy. On the other hand, methods could be developed to calibrate the estimated parameters or the computed biases of the project portfolio selection, for example, by building a model between the estimation errors and the parameters of the project portfolio selection, p and N.

Appendices

Appendix A provides a detailed proof of the computation of the expected values in the Expectation step. For the Maximization step, the log-likelihood function used in likelihood maximization is deduced in appendix B, and the maximization of this function is presented in appendix C.

A Conditional expected values of $\ln C_i$ proof

In order to compute the conditional expected values $\mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]$ and $\mathbb{E}\left[\ln^2 C_i \middle| \ln C_i^E = \ln c_i^E\right]$, we model $\left(\ln C_i, \ln C_i^E\right)^T$ as a random vector that follows a bivariate normal distribution. Hence

$$\begin{gathered}\left(\ln C_i, \ln C_i^E\right)^T \sim \mathrm{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \text{where} \\ \boldsymbol{\mu} = \begin{pmatrix} \mathbb{E}\left[\ln C_i\right] \\ \mathbb{E}\left[\ln C_i^E\right] \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{Var}\left[\ln C_i\right] & \mathrm{Cov}\left[\ln C_i, \ln C_i^E\right] \\ \mathrm{Cov}\left[\ln C_i, \ln C_i^E\right] & \mathrm{Var}\left[\ln C_i^E\right] \end{pmatrix}\end{gathered} \tag{6}$$

We begin with the descriptive statistics of $\ln C_i$. Since $C_i \sim \mathrm{LogN}\left(\mu, \sigma^2\right)$, it is easy to see that

$$\ln C_i \sim \mathrm{N}\left(\mu, \sigma^2\right) \Rightarrow \mathbb{E}\left[\ln C_i\right] = \mu \text{ and } \mathrm{Var}\left[\ln C_i\right] = \sigma^2. \tag{7}$$

The cost estimates C_i^E are assumed conditionally independent random variables $\left(C_i^E \middle| C_i = c_i\right) = \Delta c_i$. If we leave out the condition of given c_i, C_i^E is a product of two lognormally distributed random variables, $\Delta \sim \mathrm{LogN}(\eta, \tau^2)$ and $C_i \sim \mathrm{LogN}(\mu, \sigma^2)$. By taking the logarithm of this random variable, we get

$$\begin{aligned}\ln C_i^E = \ln \Delta C_i = \ln \Delta + \ln C_i = X + Y, \text{ where } X \sim \mathrm{N}\left(\eta, \tau^2\right) \text{ and } Y \sim \mathrm{N}\left(\mu, \sigma^2\right) \\ \Rightarrow \ln C_i^E \sim \mathrm{N}\left(\eta + \mu, \tau^2 + \sigma^2\right) \Rightarrow \mathbb{E}\left[\ln C_i^E\right] = \eta + \mu, \mathrm{Var}\left[\ln C_i^E\right] = \tau^2 + \sigma^2.\end{aligned} \tag{8}$$

The covariance of C_i and C_i^E can be derived as follows.

$$\begin{aligned}\mathrm{Cov}\left[\ln C_i, \ln C_i^E\right] =& \mathbb{E}\left[\ln C_i \ln C_i^E\right] - \mathbb{E}\left[\ln C_i\right] \mathbb{E}\left[\ln C_i^E\right] \\ =& \mathbb{E}\left[\ln C_i \left(\ln \Delta + \ln C_i\right)\right] - \mathbb{E}\left[\ln C_i\right] \mathbb{E}\left[\ln C_i^E\right] \\ =& \mathbb{E}\left[\ln C_i \ln \Delta\right] + \left(-\mathbb{E}\left[\ln C_i\right] \mathbb{E}\left[\ln \Delta\right] + \mathbb{E}\left[\ln C_i\right] \mathbb{E}\left[\ln \Delta\right]\right) \\ & + \mathbb{E}\left[\ln^2 C_i\right] + \left(-\mathbb{E}\left[\ln C_i\right]^2 + \mathbb{E}\left[\ln C_i\right]^2\right) - \mathbb{E}\left[\ln C_i\right] \mathbb{E}\left[\ln C_i^E\right] \\ =& \mathrm{Cov}\left[\ln C_i, \ln \Delta\right] + \mathbb{E}\left[\ln C_i\right] \mathbb{E}\left[\ln \Delta\right] + \mathrm{Var}\left[\ln C_i\right] + \mathbb{E}\left[\ln C_i\right]^2 \\ & - \mathbb{E}\left[\ln C_i\right] \mathbb{E}\left[\ln C_i^E\right] \\ =& 0 + \mu\eta + \sigma^2 + \mu^2 - \mu\left(\eta + \mu\right) = \sigma^2,\end{aligned} \tag{9}$$

where $\text{Cov}\left[\ln C_i, \ln \Delta\right] = 0$ follows from the independence of C_i and Δ. In general, for a binomial random vector $X = (X_1, X_2) \sim \text{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, the bivariate conditional expectation and the bivariate conditional variance are, respectively,

$$\begin{aligned} \mathbb{E}\left[X_1 | X_2 = x_2\right] &= \mathbb{E}\left[X_1\right] + \text{Corr}\left[X_1, X_2\right] \sqrt{\frac{\text{Var}\left[X_1\right]}{\text{Var}\left[X_2\right]}} \left(x_2 - \mathbb{E}\left[X_2\right]\right) \text{ and} \\ \text{Var}\left[X_1 | X_2 = x_2\right] &= \text{Var}\left[X_1\right] \left(1 - \text{Corr}\left[X_1, X_2\right]^2\right). \end{aligned} \tag{10}$$

Thus, we can compute the conditional expected value of $\ln C_i$ given $\ln c_i^E$ as follows.

$$\begin{aligned} \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right] &= \mathbb{E}\left[\ln C_i\right] + \text{Corr}\left[\ln C_i, \ln C_i^E\right] \sqrt{\frac{\text{Var}\left[\ln C_i\right]}{\text{Var}\left[\ln C_i^E\right]}} \left(\ln c_i^E - \mathbb{E}\left[\ln C_i^E\right]\right) \\ &= \mathbb{E}\left[\ln C_i\right] + \frac{\text{Cov}\left[\ln C_i, \ln C_i^E\right]}{\sqrt{\text{Var}\left[\ln C_i\right] \text{Var}\left[\ln C_i^E\right]}} \sqrt{\frac{\text{Var}\left[\ln C_i\right]}{\text{Var}\left[\ln C_i^E\right]}} \left(\ln c_i^E - \mathbb{E}\left[\ln C_i^E\right]\right) \\ &= \frac{\text{Var}\left[\ln C_i^E\right]}{\text{Var}\left[\ln C_i^E\right]} \mathbb{E}\left[\ln C_i\right] + \frac{\text{Cov}\left[\ln C_i, \ln C_i^E\right]}{\text{Var}\left[\ln C_i^E\right]} \left(\ln c_i^E - \mathbb{E}\left[\ln C_i^E\right]\right) \\ &= \frac{1}{\tau^2 + \sigma^2} \left[\left(\tau^2 + \sigma^2\right) \mu + \sigma^2 \left(\ln c_i^E - \eta - \mu\right)\right] \\ &= \frac{\tau^2}{\tau^2 + \sigma^2} \mu + \frac{\sigma^2}{\tau^2 + \sigma^2} \left(\ln c_i^E - \eta\right) \end{aligned} \tag{11}$$

Finally, we use the algebraic formula for the variance $\text{Var}\left[X\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2$ and the bivariate conditional variance to compute the conditional expected value of $\ln^2 C_i$ given $\ln c_i^E$.

$$\begin{aligned} \mathbb{E}\left[\ln^2 C_i \middle| \ln C_i^E = \ln c_i^E\right] &= \text{Var}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right] + \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]^2 \\ &= \text{Var}\left[\ln C_i\right] \left(1 - \text{Corr}\left[\ln C_i, \ln C_i^E\right]^2\right) + \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]^2 \\ &= \text{Var}\left[\ln C_i\right] \left(1 - \frac{\text{Cov}\left[\ln C_i, \ln C_i^E\right]^2}{\text{Var}\left[\ln C_i\right] \text{Var}\left[\ln C_i^E\right]}\right) + \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]^2 \\ &= \sigma^2 \left[1 - \frac{\left(\sigma^2\right)^2}{\sigma^2 \left(\tau^2 + \sigma^2\right)}\right] + \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]^2 \\ &= \frac{\sigma^2 \left(\tau^2 + \sigma^2\right) - \left(\sigma^2\right)^2}{\tau^2 + \sigma^2} + \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]^2 \\ &= \frac{\sigma^2 \tau^2}{\tau^2 + \sigma^2} + \mathbb{E}\left[\ln C_i \middle| \ln C_i^E = \ln c_i^E\right]^2 \end{aligned} \tag{12}$$

B Log-likelihood function proof

The likelihood of the parameters in a model of independent random variables is

$$\mathcal{L}(\theta|x) = \prod_{i=1}^{n} f_X(x_i|\theta), \tag{13}$$

where $f_X(x_i|\theta)$ is the probability density function of X at the data points x_i, given the parameter values θ. This function represents the *likelihood* of the parameters θ , given observed data x. Consequently, by maximizing this function with respect to θ we get the parameters that best fit to the data, the maximum likelihood estimates $\hat{\theta}$.

Because $\ln f(x)$ is maximized by the same x as $f(x)$, it is possible to work with the logarithm of (13), the log-likelihood function. This is often more convenient, since the logarithm of a product equals the sum of logarithms of the factors. Thus, our goal is to find the parameters θ, that maximize the log-likelihood function. Formally this can be stated as

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \ell(\theta|x), \text{ where} \\ \ell(\theta|x) &= \ln \mathcal{L}(\theta|x) = \sum_{i=1}^{n} \ln f_X(x_i|\theta). \end{aligned} \tag{14}$$

We begin by noting that the bivariate density of C_i^E and C_i can be written as the marginal density of C_i times the conditional density of C_i^E given c_i. That is

$$f_{C_i^E, C_i}\left(c_i^E, c_i|\mu, \sigma^2, \eta, \tau^2\right) = f_{C_i}\left(c_i|\mu, \sigma^2\right) f_{C_i^E}\left(c_i^E|C_i = c_i, \eta, \tau^2\right). \tag{15}$$

$C_i \sim \text{LogN}\left(\mu, \sigma^2\right)$ gives us

$$f_{C_i}(c_i|\mu, \sigma) = \ln \varphi\left(c_i|\mu, \sigma^2\right), \tag{16}$$

where $\ln \varphi\left(x|\mu, \sigma^2\right)$ stands for the probability density function of the lognormal distribution. Correspondingly, $\left(C_i^E|C_i = c_i\right) = \Delta c_i$, where $\Delta \sim \text{LogN}(\eta, \tau^2)$, yields

$$\begin{aligned} &\left(C_i^E|C_i = c_i\right) \sim \text{LogN}(\ln c_i + \eta, \tau^2) \\ \Rightarrow &f_{C_i^E}\left(c_i^E|C_i = c_i, \eta, \tau^2\right) = \ln \varphi\left(c_i^E|\ln c_i + \eta, \tau^2\right) \end{aligned} \tag{17}$$

We note here that the relationship between the lognormal and normal density function for scalar variables is

$$\ln \varphi\left(x|\mu, \sigma^2\right) = \frac{1}{x} \varphi\left(\ln x|\mu, \sigma^2\right), \tag{18}$$

where $\varphi\left(x|\mu,\sigma^2\right)$ stands for the probability density function of the normal distribution.

The likelihood, given complete data (c^E, c), is

$$\begin{aligned}\mathcal{L}\left(\theta|c^E,c\right) &= \prod_{i=1}^{N} f_{C_i^E,C_i}\left(c_i^E,c_i|\mu,\sigma^2,\eta,\tau^2\right)\\ &= \prod_{i=1}^{N} f_{C_i}\left(c_i|\mu,\sigma\right) f_{C_i^E}\left(c_i^E|C_i=c_i,\eta,\tau^2\right)\\ &= \prod_{i=1}^{N} \ln\varphi\left(c_i|\mu,\sigma^2\right)\ln\varphi\left(c_i^E|\ln c_i+\eta,\tau^2\right).\end{aligned} \tag{19}$$

Substituting (18) into (19) yields

$$\mathcal{L}\left(\theta|c^E,c\right) = \prod_{i=1}^{N}\left(c_i c_i^E\right)^{-1}\varphi\left(\ln c_i|\mu,\sigma^2\right)\varphi\left(\ln c_i^E|\ln c_i+\eta,\tau^2\right). \tag{20}$$

Next, we take the logarithm of (20) to get the log-likelihood function.

$$\ell\left(\theta|c^E,c\right) = \sum_{i=1}^{N} -\ln c_i c_i^E + \ln\left[\varphi\left(\ln c_i|\mu,\sigma^2\right)\varphi\left(\ln c_i^E|\ln c_i+\eta,\tau^2\right)\right] \tag{21}$$

We note that the probability density function of a normal random variable is of the form

$$\varphi\left(x|\mu,\sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{22}$$

First, we write the probability density functions as (22) and drop the first term in (21) that is constant with respect to θ (since eventually we are defining the θ that maximizes the log-likelihood).

$$\ell\left(\theta|c^E,c\right) = \sum_{i=1}^{N}\ln\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\ln c_i-\mu)^2}{2\sigma^2}}\frac{1}{\tau\sqrt{2\pi}}e^{-\frac{\left(\ln c_i^E-\ln c_i-\eta\right)^2}{2\tau^2}}\right] \tag{23}$$

Then, we apply the logarithm to get

$$\ell\left(\theta|c^E,c\right) = \sum_{i=1}^{N} -\ln 2\pi - \ln\left(\tau\sigma\right) - \frac{\left(\ln c_i-\mu\right)^2}{2\sigma^2} - \frac{\left(\ln c_i^E-\ln c_i-\eta\right)^2}{2\tau^2}. \tag{24}$$

Again, we can drop the first term that has no effect on the maximization. Then, we can sum over i in the first term since there is no dependence on i. Finally, we get the function in the form

$$\ell\left(\theta|c^E,c\right) = -N\ln\left(\tau\sigma\right) - \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\left(\ln c_i-\mu\right)^2}{\sigma^2} + \frac{\left(\ln c_i^E-\ln c_i-\eta\right)^2}{\tau^2}\right]. \tag{25}$$

C Complete data ML estimates proof

The log-likelihood function will be maximized (as in (14)) using Mathematica. However, the form (25) is not possible to be directly maximized with the software. Thus, the expression requires some modification.

The terms will be arranged so that the estimated parameters μ, σ^2, η and τ^2 are shifted out from the sums. This is possible since the parameters are not dependent on i. By extracting the denominator from the sum in (25), we get

$$\ell\left(\theta\middle|c_i^E, c_i\right) = -N\ln\left(\tau\sigma\right) - \frac{1}{2\tau^2\sigma^2}\sum_{i=1}^{N}\left[\tau^2\left(\ln c_i - \mu\right)^2 + \sigma^2\left(\ln c_i^E - \ln c_i - \eta\right)^2\right]. \quad \textbf{(26)}$$

Let us take a more detailed look at the second sum in (26).

$$\begin{aligned}
&\sum_{i=1}^{N}\tau^2\left(\ln c_i - \mu\right)^2 + \sigma^2\left(\ln c_i^E - \ln c_i - \eta\right)^2 \\
&= \sum_{i=1}^{N}\tau^2\left(\ln^2 c_i - 2\mu\ln c_i + \mu^2\right) + \sigma^2\left[\left(\ln c_i^E - \ln c_i\right)^2 - 2\eta\left(\ln c_i^E - \ln c_i\right) + \eta^2\right] \\
&= \sum_{i=1}^{N}\tau^2\left(\ln^2 c_i - 2\mu\ln c_i + \mu^2\right) + \sigma^2\left[\ln^2 c_i^E - 2\ln c_i^E\ln c_i + \ln^2 c_i\right. \\
&\qquad \left. - 2\eta\left(\ln c_i^E - \ln c_i\right) + \eta^2\right] \qquad \textbf{(27)} \\
&= \sum_{i=1}^{N}\sigma^2\ln^2 c_i^E + \left(\tau^2+\sigma^2\right)\ln^2 c_i - 2\sigma^2\ln c_i^E\ln c_i - 2\sigma^2\eta\ln c_i^E \\
&\qquad + 2\left(\sigma^2\eta - \tau^2\mu\right)\ln c_i + \left(\tau^2\mu^2 + \sigma^2\eta^2\right) \\
&= \sigma^2\left(\sum_{i=1}^{N}\ln^2 c_i^E\right) + \left(\tau^2+\sigma^2\right)\left(\sum_{i=1}^{N}\ln^2 c_i\right) - 2\sigma^2\left(\sum_{i=1}^{N}\ln c_i^E\ln c_i\right) \\
&\qquad - 2\eta\sigma^2\left(\sum_{i=1}^{N}\ln c_i^E\right) + 2\left(\sigma^2\eta - \tau^2\mu\right)\left(\sum_{i=1}^{N}\ln c_i\right) + \left(\tau^2\mu^2 + \eta^2\sigma^2\right)N
\end{aligned}$$

Now, we will note the sums of $\ln c_i$, $\ln c_i^E$, $\ln^2 c_i$, $\ln^2 c_i^E$ and $\ln c_i^E \ln c_i$ as

$$\begin{aligned}
A &= \sum_{i=1}^{N}\ln^2 c_i^E, & B &= \sum_{i=1}^{N}\ln^2 c_i, & C &= \sum_{i=1}^{N}\ln c_i^E\ln c_i \qquad \textbf{(28)} \\
D &= \sum_{i=1}^{N}\ln c_i^E, & E &= \sum_{i=1}^{N}\ln c_i,
\end{aligned}$$

These notations give the log-likelihood function in the form

$$\ell\left(\theta\middle|\ln c^E, \ln c\right) = -\left(\ln(\tau\sigma) + \frac{\tau^2\mu^2 + \eta^2\sigma^2}{2\tau^2\sigma^2}\right)N \\ - \frac{1}{2\tau^2\sigma^2}\left[\sigma^2 A + \left(\tau^2+\sigma^2\right)B - 2\sigma^2 C - 2\eta\sigma^2 D + 2\left(\sigma^2\eta - \tau^2\mu\right)E\right] \tag{29}$$

This form of log-likelihood can be maximized with a computational engine. We use the software Wolfram Mathematica to calculate the maximum. This gives us

$$\begin{aligned} &\hat{\theta} = \left[\hat{\mu}, \hat{\sigma}^2, \hat{\eta}, \hat{\tau}^2\right], \qquad \text{where} \\ &\hat{\mu} = \frac{E}{N}, && \hat{\sigma}^2 = \frac{B}{N} - \frac{E^2}{N^2} = \frac{B}{N} - \hat{\mu}^2, \\ &\hat{\eta} = \frac{D}{N} - \frac{E}{N} = \frac{D}{N} - \hat{\mu}, && \hat{\tau}^2 = \frac{A-2C+B}{N} - \frac{D^2-2DE+E^2}{N^2} = \frac{A-2C+B}{N} - \hat{\eta}^2 \end{aligned} \tag{30}$$

The original sums will be retrieved from the formulas in (28). We get the ML estimates of the complete data as

$$\begin{aligned} &\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \ln c_i, && \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} \ln^2 c_i - \hat{\mu}^2 \\ &\hat{\eta} = \frac{1}{N}\sum_{i=1}^{N} \ln c_i^E - \hat{\mu} && \hat{\tau}^2 = \frac{1}{N}\sum_{i=1}^{N} \left(\ln^2 c_i^E - 2\ln c_i^E \ln c_i + \ln^2 c_i\right) - \hat{\eta}^2 \end{aligned} \tag{31}$$

It is easy to see that the formulas of $\hat{\sigma}^2$ and $\hat{\tau}^2$ in (31) are those of the ML estimates of the variances of random variables $\ln c_i \sim \mathrm{N}\left(\mu, \sigma^2\right)$ and $\left(\ln c_i^E - \ln c_i\right) \sim \mathrm{N}\left(\eta, \tau^2\right)$. However, the ordinary ML estimates of variances give, in general, biased estimates of the true values (i.e., the mean of the estimates do not equal the true values of the parameters). These estimates can be debiased simply by multiplying them with a factor $N/(N-1)$. We will also move $\hat{\mu}$ and $\hat{\eta}$ inside the summing statements to simplify the formulas. Making these changes gives us

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{N-1}\sum_{i=1}^{N} \left(\ln^2 c_i - \hat{\mu}^2\right) \\ \hat{\tau}^2 &= \frac{1}{N-1}\sum_{i=1}^{N} \left(\ln^2 c_i^E - 2\ln c_i^E \ln c_i + \ln^2 c_i - \hat{\eta}^2\right). \end{aligned} \tag{32}$$

The formulas of $\hat{\mu}$ and $\hat{\eta}$ in (31) and $\hat{\sigma}^2$ and $\hat{\tau}^2$ in (32) will be used in the Maximization step of the EM algorithm to compute the estimates $\hat{\theta}_{k+1} = \left[\hat{\mu}_{k+1}, \hat{\sigma}^2_{k+1}, \hat{\eta}_{k+1}, \hat{\tau}^2_{k+1}\right]$ in each iteration.

From (31) and (32) we also gain the initial values of the parameters $\hat{\theta}_0 = [\hat{\mu}_0, \hat{\sigma}_0^2, \hat{\eta}_0, \hat{\tau}_0^2]$ rather easily. Because at the beginning of the algorithm there are no estimates for the missing data $\ln c^\circ$, it is reasonable to compute the first parameters with all the data that exists, $(\ln c^E, \ln c^*)$ (i.e., we get the initial values $\hat{\theta}_0$ by summing over the selected projects $i \in I^*$ instead of all the projects, from 1 to N). Furthermore, since $\hat{\eta}_{k+1}$ is computed with $\hat{\mu}_{k+1}$ in (31), it is conversely possible to compute $\hat{\mu}_{k+1}$ by using $\hat{\eta}_{k+1}$ as follows.

$$\hat{\eta}_{k+1} = \frac{1}{N}\sum_{i=1}^{N} \ln c_i^E - \hat{\mu} \Rightarrow \hat{\mu}_{k+1} = \frac{1}{N}\sum_{i=1}^{N} \ln c_i^E - \hat{\eta}_{k+1} \qquad \textbf{(33)}$$

This modification will permit including all the data of the cost estimates when computing the initial value $\hat{\mu}_0$. However, if the initial $\hat{\mu}$ is computed with N data points from $\ln c^E$, the formula of $\hat{\sigma}^2$ in (31) does not stand when computed with M data points of the completed projects' $\ln c^*$. Instead, the following formula can be used.

$$\hat{\sigma}_0^2 = \frac{1}{M-1}\sum_{i \in I^*} \left[\ln^2 c_i - \left(\frac{1}{M}\sum_{i \in I^*} \ln c_i\right)^2\right], \qquad \textbf{(34)}$$

where instead of $\hat{\mu}_0$, the explicit formula of it is applied to the cost logarithms of the completed projects $\ln c^*$.

References

Begg, S. H. and Bratvold, R. B. (2008). Systematic prediction errors in oil and gas project portfolio selection. *SPE Annual Technical Conference and Exhibition*.

Brown, K. C. (1974). A note on the apparent bias of net revenue estimates for capital investment projects. *The Journal of Finance*, 29(4):1215–1216.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Flyvberg, B., Holm, M. S., and Soren, B. (2002). Underestimating costs in public works projects: Error or lie? *Journal of the American Planning Association*, 63(3):279–295.

Harrison, J. R. and March, J. G. (1984). Decision making and postdecision surprises. *Administrative Science Quarterly*, 29(1):26–42.

Jørgensen, M. (2013). The influence of selection bias on effort overruns in software development projects. *Information and Software Technology*, 55(9):1640–1650.

Keisler, J. (2004). Value of information in portfolio decision analysis. *Decision Analysis*, 1(3):177–189.

Ohlsson, M. C. and Wohlin, C. (1999). An empirical study of effort estimation during project execution. In *Proceedings International Symposium on Software Metrics*, pages 91–98, Boca Raton, Florida, USA.

Ohlsson, M. C., Wohlin, C., and Regnell, B. (1998). A project effort estimation study. *Information and Software Technology*, 40(14):831–839.

Salo, A., Keisler, J., and Morton, A. (2011). *Portfolio Decision Analysis: Improved Methods for Resource Allocation*. Springer, New York.

Siemiatycki, M. (2009). Academics and auditors: Comparing perspectives on transportation project cost overruns. *Journal of Planning Education and Research*, 29(2):142–156.

Smith, J. E. and Winkler, R. L. (2006). The optimizers curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322.

Vilkkumaa, E. and Liesiö, J. (2015). Estimating and mitigating systematic and selection biases in decision analysis. *Manuscript*.

Vilkkumaa, E., Liesiö, J., and Salo, A. (2014). Optimal strategies for selecting project portfolios using uncertain value estimates. *European Journal of Operational Research*, 233(3):772–783.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1):95–103.