# PUBLIC R&D PROJECT PORTFOLIO SELECTION UNDER EXPENDITURE UNCERTAINTY AND POLICY CONSTRAINTS

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Production and Operations Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Wiley - Manuscript type:</td>
<td>Original Article</td>
</tr>
<tr>
<td>Keywords:</td>
<td>project portfolio selection, expenditure uncertainty, input modeling, Normal approximation, chance-constrained stochastic programming</td>
</tr>
</tbody>
</table>

## Abstract:

We consider a project selection problem faced by research councils. The decision maker seeks to construct an optimal subset of projects, which maximizes the expected total score of supported projects subject to the total budget and policy constraints. At the time of funding decisions, exact expenditures of projects are not known. The decision maker selects projects based on approved budgets, which could be higher than exact expenditures. Therefore, the realized total expenditure of the portfolio tends to be lower than the total budget, causing the notion of budgetary slack. In this study, we put a special emphasis on the input modeling of expenditures by considering practical information, and propose a mixture distribution. Then, we develop a chance-constrained model with policy constraints to enhance the budget utilization. Due to the intractability of the developed model, we have shown that Normal distribution can be used for approximation. However, the decision maker wonders the quality of Normal approximation. Therefore, we quantify the approximation error of our model via a theoretical bound and simulation. We show that Normal distribution gives a good approximation and the theoretical bound is relatively tight for large-scale problems. We find that our model can be solved to optimal (or near optimal) in a reasonable amount of time by a commercial solver such as IBM CPLEX. We also show that the budget utilization rate of 100% is hard to achieve but that of 96-97% is within reach. The proposed approach can increase the budget utilization by 8.0% and 15.2%, which is remarkable for public decision makers. Thereby, more R&D projects could be supported and a higher socio-economic impact can be achieved.
PUBLIC R&D PROJECT PORTFOLIO SELECTION UNDER EXPENDITURE UNCERTAINTY AND POLICY CONSTRAINTS

ABSTRACT. We consider a project selection problem faced by research councils. The decision maker seeks to construct an optimal subset of projects, which maximizes the expected total score of supported projects subject to the total budget and policy constraints. At the time of funding decisions, exact expenditures of projects are not known. The decision maker selects projects based on approved budgets, which could be higher than exact expenditures. Therefore, the realized total expenditure of the portfolio tends to be lower than the total budget, causing the notion of budgetary slack. In this study, we put a special emphasis on the input modeling of expenditures by considering practical information, and propose a mixture distribution. Then, we develop a chance-constrained model with policy constraints to enhance the budget utilization. Due to the intractability of the developed model, we have shown that Normal distribution can be used for approximation. However, the decision maker wonders the quality of Normal approximation. Therefore, we quantify the approximation error of our model via a theoretical bound and simulation. We show that Normal distribution gives a good approximation and the theoretical bound is relatively tight for large-scale problems. We find that our model can be solved to optimal (or near optimal) in a reasonable amount of time by a commercial solver such as IBM CPLEX. We also show that the budget utilization rate of 100% is hard to achieve but that of 96-97% is within reach. The proposed approach can increase the budget utilization by 8.0% and 15.2%, which is remarkable for public decision makers. Thereby, more R&D projects could be supported and a higher socio-economic impact can be achieved.

Keywords: project portfolio selection, expenditure uncertainty, input modeling, Normal approximation, chance constrained stochastic programming.

1. Introduction

Project selection is a complex decision problem applied in various organizations such as high technology companies, municipalities, and research councils (Kettunen and Salo [2017]). In this study, we consider a project selection problem faced by research councils. A research council is a public funding agency, which is accountable for distributing government funds through grant programs to promote generation of new knowledge. Research councils generally adopt a call-based system. A call is announced under a grant program and researchers apply with their solicited project proposals. After eligibility check of project proposals, peer reviewers assess eligible proposals in panels according to the merit review criteria.

Funding decisions of projects are made according to project scores (i.e. ex ante value estimates by peer reviewers), approved budgets of projects (i.e. ex ante expenditure estimates), policy constraints (i.e. constraints formulated according to decision maker’s perception of fairness in a project portfolio) and the total available budget. In particular, the decision maker seeks to construct an optimal subset of projects (i.e. portfolio), which maximizes...
the expected total score of supported projects subject to the total budget and other policy constraints. This approach is also referred to as project portfolio optimization and provides effective use of available resources, accurately modeling of risks and optimization of an overall objective (Hall et al. [2015]).

In this process, one of the critical decisions is setting the budget of each project. An applying project requests an initial resource requirement (i.e. the budget proposal) and research councils make the final budget decision (i.e. the approved budget) according to recommendations of peer reviewers. The budget setting of a project is a participative process. Principal investigator (PI), project team members, peer reviewers and the research council get involved in the budget setting. It is worth noting that PIs and project team members get compensation from project budgets. PIs are usually faculty at universities and team members are usually PhD students and/or post-doc researchers.

Antle and Eppen [1985] provide evidence that if one’s compensation is linked with a budget resource which is in some degree accessible to him/her, then s/he might make an attempt to inflate the budget through the participative process. They define this behavior as “organizational slack” (i.e. excess of the budget allocated or budgetary slack). It is estimated that around 80% of supervisors overstate their budgets to get a budgetary slack (Dunk and Nouri [1998]). Hu and Szmerékovszky [2017] highlight that subordinates have a tendency of building budgetary slack by amplifying costs.

At the time of funding decisions, exact expenditures of projects are not known. The decision maker selects projects based on approved budgets, which could be higher than exact expenditures. Therefore, the realized total expenditure of the portfolio tends to be lower than the total available budget, causing the notion of a budgetary slack in a project portfolio. Note that the total funding budget of a research council is allocated among many grant programs. For instance, The Scientific and Technological Research Council of Turkey (TUBITAK) has more than thirty grant programs. The residual (i.e. unspent money) of each program contributes to the overall budget inefficiency. For instance, when we look at appropriations and expenditures of TUBITAK funding, we can observe the apparent budgetary slack throughout years (see Figure 1).

Pollack and Zeckhauser [1996] provide evidence that under-spent budgets imply negative message (i.e. reduce appropriations) about budgetary requirements for upcoming periods. Thus, officials warn their departments about fully using allocated budgets. Our work is motivated by a similar situation faced by a research council. TUBITAK experienced a serious budgetary slack (i.e. unspent budget) rate of 41% in 2012 (see Figure 1). During negotiations of budget appropriations with the Ministry of Development (formerly known as the State Planning Organization) at the end of 2012, the budget appropriation of 2013 was reduced due to escalating unspent budget rates in 2010, 2011, and 2012 (see Figure 1). For a research council, utilizing its funding budget at maximum, i.e. supporting more projects as possible, is critical to maximize the scientific and technological output of the country. When we evaluate unspent ratios in terms of magnitude, we can see that even 5% in 2015 amounts 65 million Turkish Lira (around $24 million), which is a pretty high quantity. This is a macro level evidence on budget inefficiency that decision makers of research councils should take into account.
If each program’s total budget could be used efficiently, then the total budget of a research council will be utilized at its maximum. For that reason, in this study, we focus on the 1001 program (i.e., Scientific and Technological Research Projects Funding Program) of TUBITAK to get managerial insights about the expenditure behavior of projects. The 1001 program is one of the important academic grants of TUBITAK. Its aim is to support research for generating new information or solving technological problems in compliance with scientific basis. The 1001 program receives research projects proposed by public agencies, private companies, and universities.

We have looked at the expenditure behavior of projects in the 1001 program and observed that there are two main determinants causing a budgetary slack. Firstly, cancellations of an unignorable number of R&D projects before spending most of their budgets. R&D activities are inherently subject to uncertainties that can arise during the execution of an R&D project. PI and other project members may fail to comply with the approved project plan, which may result with the cancellation of the project. Rules and conditions for cancellation of projects are usually arranged in the R&D program regulations (see NSF [2005], NIH [2013]). Canceled projects are regarded as unsuccessful and the expected scientific and technological benefit of them aren’t realized. Secondly, a significant number of granted projects end up successfully while providing their expected benefits without spending their whole budget. Since planning of R&D activities is a complex process, research councils provide detailed guidelines for the correct estimation of project budgets. However, as we have already noted, researchers may overestimate budgets of their projects (Dunk and Nouri [1998], Hu and Szmerekovsky [2017]).
We believe that if decision makers of research councils consider the factors causing budgetary slack in a decision model, they could remarkably improve the budget utilization while satisfying policy constraints. Hence, we study the expenditure uncertainty in a project selection problem faced by a research council and propose a decision model that will improve the budget utilization. Increased public R&D budget utilization will help to support more R&D projects and hence achieve a higher socio-economic impact.

Although we study budgetary slack in a research council, we want to note that cost overruns (i.e. being overbudget) are also studied in the literature (Conboy [2010] and Tseng et al. [2005]). However, as literature suggests, cost overruns arise in information technology development or infrastructure/construction projects. We also want to note that our proposed approach can be adapted to model cost overruns. Technically, a distribution function range should be set according to data, then derivations need to be obtained and integrated into our model. Next, we give the related literature.

1.1. Related Literature. In the literature, many analytical models have been proposed for project selection problems under various domains. A full review of the literature is beyond the scope of this study. We refer to Kavadias and Chao [2007] for a review. We notice that most of the studies focus on project selection problems faced by industry. Research on project selection problems faced by research councils is scant.

Existing studies of the problem in industry deal with very small numbers of projects (see Solak et al. [2010] and Medaglia et al. [2007] for problems with only 5 or 10 projects). We don’t exclude the possibility that calls with very small number of project proposals may exist in some exceptional programs (i.e. programs for defense or space projects) of a research council. However, the number of projects applying to nationwide bottom-up R&D grant programs can reach up to the scale of thousands (Çağlar and Gürel [2019] and Arratia et al. [2016]). When we look at TUBITAK, the number of project applications in nationwide programs varies between 250 and 2000. A need for large-scale project selection models including chance constraints of expenditures is mentioned by Gabriel et al. [2006]. Therefore, in this study, we develop a chance-constrained based model for moderate and large-scale project selection problems under expenditure uncertainty and policy constraints.

In the project portfolio selection literature, several researchers consider project cancellations. Solak et al. [2010] develop a multi-stage model in an industrial setting. They study optimal allocation of budgets to small number of projects by evaluating their annual returns. They also deal with dynamic cancellation decisions of projects by periodically comparing their returns with expenditures. Çağlar and Gürel [2017] study mixed integer and dynamic programming approaches by assuming number of cancellations are given in a public project portfolio selection problem. They also propose a model in which only cancellations are modeled through a simple Bernoulli distribution by assuming a fixed expenditure rate for canceled projects. They ignore policy constraints to limit the computational complexity. In this study, we put a special emphasis on the characterization of expenditure distribution of all projects (i.e. canceled and successfully completed projects), which is known as the input modeling in the simulation literature. Then, we propose a chance-constrained model with
policy constraints. We also conduct a simulation study to evaluate the empirical probability of our model.

In terms of budgetary slack in a project selection problem, to the best of our knowledge, there is only one study by Hu and Szmerekovsky [2017]. They propose a news-vendor based approach and solve an example problem with 20 projects in Microsoft Excel. They assume that per unit under budget cost for each project is known (i.e. for every dollar under budget a penalty will be incurred). However, this assumption is very challenging to quantify in the practice for a research council. In our study, we solve a large-scale non-linear model, which is very hard to solve in Microsoft Excel.

In the stochastic optimization literature, chance-constrained stochastic program (CCSP) is a method to deal with randomness of input data in a constraint. It is introduced by Charnes et al. [1958] and Charnes and Cooper [1959]. If random input data do not follow the multivariate normal distribution, CCSPs are very hard to solve numerically due to issues of checking feasibility and convexity. We refer to Luedtke et al. [2010] for a recent review on CCSP.

There are two main approximation approaches to deal with CCSPs. One approach is to transform the CCSP into a deterministic counterpart by analytical approximations using probability inequalities (Nemirovski and Shapiro [2006]). This approach can suffer from inconsistency of the approximation due to parameter tuning issues. Second one is sampling from the distribution of random data to generate samples and then formulating a binary integer program (a.k.a. sample average approximation, Pagnoncelli et al. [2009]). Sample average approximation needs large samples to obtain good solutions. Therefore, it becomes computationally intractable even for moderate-size problems (i.e. any problem with more than 200 projects/items requires more than 100,000 samples) due to solving a binary integer program. Both approaches are not easy to implement due to the sample size and problem specific tuning challenges.

Due to challenges of proposed models in the literature, in this study, we apply Normal approximation to make our model tractable. Our computational experiments show that large-scale problems can be solved to optimality (or near optimality) in 3 hours. However, the decision maker wonders the approximation error. Therefore, we assess the empirical probability of the chance constraint in our model via simulation and a theoretical bound.

1.2. Contributions. Our study contributes to project portfolio selection literature by elaborating on the input modeling of expenditures along with a CCSP under policy constraints. In the literature on CCSPs, it is usually assumed that the distribution for a random input is known and researchers focus on solution approaches in small-scale problems. In our large-scale practical problem, the expenditure distribution is not readily available. Therefore, we put a special emphasis on the input modeling of expenditures by considering practical information and propose a mixture distribution. Then, we incorporate the mixture distribution into a CCSP with policy constraints. Due to the computational intractability of the developed model, we show that Normal approximation can be applied. However, the decision maker wonders the quality of Normal approximation. Therefore, we show the approximation error of our model via a theoretical bound and simulation. To the best of our knowledge,
this is the first study that investigates the convergence quality of Normal Approximation in a CCSP when random data follow a mixture distribution. We show that Normal distribution gives a good approximation and the theoretical bound is relatively tight for large-scale problems. We find that our model can be solved to optimal (or near optimal) in a reasonable amount of time by a commercial solver such as IBM CPLEX. We also numerically show that the utilization rate of 100% is hard to achieve but that of 96-97% is within reach. The proposed approach can increase the budget utilization by 8.0% and 15.2%, which is remarkable for public decision makers.

In the next section, we state the problem and develop a CCSP. In Section 3, the computational study is presented. Conclusions are discussed in Section 4.

2. Problem Statement and Proposed Model

We address a project portfolio selection problem faced by a research council. Research councils announce a call for proposals. Researchers apply with their solicited research proposals for funding. Proposal are classified according to their area of research. Then, proposals are examined for the eligibility check by academic departments, which manage project evaluation process. Eligible proposals are evaluated by peer reviewers in panels. Scientific and technological values of projects are scored by peer reviewers according to a set of criteria. Budget of a project is not regarded as a selection criteria. However, it is considered in panels for a descent project budget. Project selection decisions are made according to the panel scores, project budgets, total budget, and policy constraints. PIs of funded projects sign a grant contract with the research council. Grant contract commits that approved budget will be gradually transferred to the project if project team complies with terms and conditions of the grant policy.

After funding decisions, granted projects start to conduct their planned research agenda. However, during research activities, some of R&D projects can be canceled by the research council due to the violation of terms and conditions. If a cancellation occurs, most of the budget expenditure is usually get back. In some cases, specific expenditures can be excluded. Those details are provided in grant policy documents. Besides, expected scientific and technological values of canceled projects are not achieved. On the other hand, some of the granted projects finish successfully but they don’t spend their whole budget. As we mentioned, researchers tend to amplify costs and there is a risk of overstating budget. Expected value of those projects are realized with a budgetary slack.

2.1. Input Modeling of Expenditure Uncertainty. Let $b_i$ denote the approved budget of project $i$. We examine expenditures in a historical data of a call. According to our observation, there could be three expenditure cases as follows:

1. cancellation of project $i$ and underspending of its budget, with probability $p_i$
2. successful completion of project $i$ and underspending of its budget, with probability $q$
3. successful completion of project $i$ and fully used its budget, with probability $1 - p_i - q$

We observed the expenditure ratio of the first case could be modeled by a beta distribution in interval $(0, \tau_1)$ and the expenditure ratio of the second case could be modeled
by a beta distribution in interval \((\tau_2, 1)\). Thus, both canceled and successful projects with underutilized budget cause the expenditure uncertainty. Let random variable \(b_i\) denote real expenditure of project \(i\). \(b_i\) will be within the interval \((0, b_i]\). Thus, we define a ratio \(R_i = \frac{b_i}{\hat{b}_i}\) to model budget ratio of a project \(i\). Mixture distributions can be employed when statistical sample data can be categorized into separate subsamples. Therefore, we formulate \(R_i\) as a mixture distribution according to above three cases. The probability density function (p.d.f) of the random variable \(R_i\) is modeled as follows:

\[
f_{R_i}(r) = \begin{cases} 
    p_i \cdot f_{W_1}(r; \alpha_1, \beta_1, 0, \tau_1) & \text{if } r \in (0, \tau_1) \\
    q \cdot f_{W_2}(r; \alpha_2, \beta_2, \tau_2, 1) & \text{if } r \in (\tau_2, 1) \\
    1 - p_i - q & \text{if } r = 1 
\end{cases}
\]

where \(f_{W_1}(r; \alpha_1, \beta_1, 0, \tau_1)\) is the p.d.f of a truncated beta random variable \(W_1\) in interval \((0, \tau_1)\) with parameters \(\alpha_1, \beta_1\) and \(f_{W_2}(r; \alpha_2, \beta_2, \tau_2, 1)\) is the p.d.f of a truncated beta random variable \(W_2\) in interval \((\tau_2, 1)\) with parameters \(\alpha_2, \beta_2\). Hence, \(R_i\) is the mixture distribution of two truncated beta random variables (i.e. \(W_1\) and \(W_2\)) and a degenerate distribution (inflated point at 1). Using Definitions 1 and 2 in Appendix A.1, we derive the cumulative distribution function (c.d.f) of \(R_i\) as follows:

\[
P(R_i \leq k) = F_{R_i}(k) = p_i \int_{0}^{\min(\tau_1, k)} \frac{\Gamma(\alpha_1 + \beta_1)r^{\alpha_1-1}(\tau_1 - r)^{\beta_1-1}}{\Gamma(\alpha_1)\Gamma(\beta_1)^{\alpha_1+\beta_1-1}} \, dr \\
+ q \int_{\tau_2}^{\min(1, k)} \frac{\Gamma(\alpha_2 + \beta_2)(r - \tau_2)^{\alpha_2-1}(1 - r)^{\beta_2-1}}{\Gamma(\alpha_2)\Gamma(\beta_2)(1 - \tau_2)^{\alpha_2+\beta_2-1}} \, dr + (1 - p_i - q)\mathbb{I}_1(k) \quad \text{for } 0 < k \leq 1
\]

where \(\mathbb{I}_1(k)\) is the indicator function and takes value 1 if \(k = 1\) and 0 otherwise.

In the statistics literature, Pereira et al. [2012] model the unemployment insurance benefit ratio by a truncated inflated beta distribution, which is a mixture of truncated beta distribution in some bounded interval \((a, b)\) and inflated in some points. We formulate \(R_i\) as a mixture of two beta distributions and a inflation point, which makes \(R_i\) more challenging to deal with.

The values of \(p_i\) and \(q\) can be estimated by expert judgments and/or past data. For example, there is usually a criterion that evaluates project management, team and research possibilities. Therefore, there can be a close relationship between cancellation probability and that kind of criterion, and this information can be gathered from past canceled projects’ data. Besides, proposals are gathered from different PIs, academic units and universities. There can also be a relation between cancellation risk and PIs, academic units, universities. Moreover, peer reviewers can comment on cancellation risk of each project and those comments can be integrated with past data to estimate cancellation probabilities. Probability information of successful projects with underutilized budget also can be obtained from past data of successfully finished projects.

2.2. CCSP with policy constraints. Let \(s_i\) denote the score of project \(i\). An aggregate scoring approach is considered since it is currently applied in many research councils such as TUBITAK. Namely, for criteria set \(C\) and peer reviewers set \(P\), \(s_i = \sum_{c \in C} \sum_{p \in P} s_{icp}\).
where $s_{icp}$ is score of project $i$ rated by peer reviewer $p$ for criterion $c$. We drop indices $c$ and $p$ for the ease of exposition. The expected value of a canceled project is not fulfilled. Therefore, the realized score of project $i$ becomes zero with the cancellation probability $p_i$. Let random variable $\hat{s}_i$ denote the realized score of project $i$. Then, the expectation of $\hat{s}_i$ is $E(\hat{s}_i) = s_i(1 - p_i)$. Next, we formulate the CCSP with policy constraints as follows.

\[
\text{maximize} \quad \sum_{i \in N} s_i (1 - p_i) x_i \\
\text{subject to} \quad P \left( \sum_{i \in N} b_i R_i x_i \leq B \right) \geq \theta \tag{3}
\]

\[
\sum_{i \in N_j} x_i \geq a_j \sum_{i \in N} x_i \quad \forall j \in M \tag{4}
\]

\[
x_i \in \{0, 1\} \quad \forall i \in N \tag{5}
\]

where $N$ indicates the set of all projects. $x_i$ is a decision variable representing 1, if project $i$ is supported and 0, otherwise. $P(\cdot)$ denotes the probability measure. $R_i$ is the random ratio defined in section 2.1. $B$ is the total available budget. $N_j$ is the set of projects in academic department $j$. $M$ is the set of academic departments. $a_j$ is the minimum acceptance rate for the academic department $j$. The objective function is to maximize the expected total score of supported projects. The chance constraint in (3) provides that the sum of random budget expenditure of supported projects does not exceed the total available budget with a probability level of $\theta$. The constraint set in (4) defines a minimum acceptance rate for projects applying to the academic department $j$. These are policy constraints specified by the decision maker, which provides fairness among academic research areas. Note that the decision maker may request similar policy constraints for different research types (i.e. basic research or applied research projects), for geographical regions of projects, and for institutions (i.e. university, public or industry) of projects, etc. Those constraints can be added to the model easily.

In order to solve the CCSP with an exact approach, we need to transform the chance constraint in (3) to its deterministic counterpart. Thus, we have to derive the quantile function (i.e. inverse cumulative distribution function) of the probabilistic part in (3). We derive the c.d.f of a single random variable $R_i$ in (2). Note that its c.d.f. has no closed-form expression. Therefore, the derivation of its quantile function is challenging. Besides, to obtain c.d.f. and inverse c.d.f of the probabilistic part in (3), the convolution (i.e. the joint distribution) of many $R_i$s should be derived, which is computationally intractable.

Due to those challenges, we will apply Normal approximation. In order to apply it, we have to show that the standardized sum of independent non-identically distributed (i.n.i.d.) random variables should obey the central limit theorem (CLT) (Theorem 1 in AppendixA.2). Proposition 1 shows that general family of truncated distributions obeys CLT.

**Proposition 1.** Let $\{T_i\}_{i=1}^n$ be an i.n.i.d. truncated random variables defined on the bounded interval $[l, u]$ such that $0 \leq l < u$. Then, CLT theorem applies to sequence $\{T_i\}_{i=1}^n$ as $n$ goes to infinity.
Proof. See Appendix A.3. □

In the following corollary, we show that the standardized version of the term \( \sum_{i \in N} b_i R_i x_i \) in the constraint (3) converges to the standard Normal distribution \( (N(0,1)) \).

**Corollary 1.** Let \( b_1 R_1, b_2 R_2, \ldots, b_i R_i \in N \) be i.i.d. random variables as defined in (3). Then \( H_n^x = \sum_{i \in N} \frac{|b_i R_i - \mathbb{E}(b_i R_i)| x_i}{\sqrt{\sum_{i \in N} \text{Var}(b_i R_i) x_i^2}} \) converges to \( N(0,1) \) if the solution of the CCSP includes statistically significant number of supported projects (i.e. \( x_i = 1 \)).

Proof. See Appendix A.4. □

We can transform the CCSP to its deterministic equivalent part by Normal approximation. However, as shown in Corollary 1, we need the mean and variance of random variable \( b_i \) to approximate the true distribution of \( H_n^x \). By using properties of expectation and variance, we obtain: \( \mathbb{E}(b_i) = b_i \mathbb{E}(R_i) \) and \( \text{Var}(b_i) = b_i^2 \text{Var}(R_i) \). \( b_i \) is known, we need to have \( \mathbb{E}(R_i) \) and \( \text{Var}(R_i) \) to apply Normal approximation. Since \( R_i \) is a mixture distribution, we first define moments of a mixture distribution in Definition 3 in Appendix A.1 and then we derive first three moments of a truncated beta distribution in Proposition 3 in Appendix A.5. We are ready to derive mean, variance and \( n^{th} \) moment of \( R_i \) in the following proposition.

**Proposition 2.** The mean, variance and \( n^{th} \) moment of the random variable \( R_i \) are given as follows:

\[
\mathbb{E}(R_i) = p_i \frac{\alpha_1 \tau_1}{\alpha_1 + \beta_1} + q \frac{\alpha_2 + \beta_2 \tau_2}{\alpha_2 + \beta_2} + (1 - p_i - q) \tag{6}
\]

\[
\text{Var}(R_i) = p_i \left[ \left( \frac{\alpha_1 \tau_1}{\alpha_1 + \beta_1} \right)^2 + \frac{\tau_1^2 \alpha_1 \beta_1}{(\alpha_1 + \beta_1)^2 (\alpha_1 + \beta_1 + 1)} \right] + q \left[ \left( \frac{\alpha_2 + \beta_2 \tau_2}{\alpha_2 + \beta_2} \right)^2 + \frac{(1 - \tau_2)^2 \alpha_2 \beta_2}{(\alpha_2 + \beta_2)^2 (\alpha_2 + \beta_2 + 1)} \right] + (1 - p_i - q)
\]

\[
- \left[ p_i \frac{\alpha_1 \tau_1}{\alpha_1 + \beta_1} + q \frac{\alpha_2 + \beta_2 \tau_2}{\alpha_2 + \beta_2} + (1 - p_i - q) \right]^2 \tag{7}
\]

\[
\mathbb{E}(R^n_i) = p_i \left( \tau_1^n \frac{\Gamma(\alpha_1 + \beta_1) \Gamma(\alpha_1 + n)}{\Gamma(\alpha_1 + \beta_1 + n)} \right) + (1 - p_i - q)
\]

\[
+ q \left( \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \tau_2^k (1 - \tau_2)^{n-k} \frac{\Gamma(\alpha_2 + \beta_2) \Gamma(\alpha_2 + n - k)}{\Gamma(\alpha_2) \Gamma(\alpha_2 + \beta_2 + n - k)} \right) \tag{8}
\]

Proof. See Appendix A.6. □

2.3. **Deterministic Equivalent Formulation by Normal Approximation.** We get the mean and variance of \( R_i \) in Proposition 2. In Corollary 1, we show that the total random budget spending obeys the CLT. Now, we formulate the deterministic equivalent of the CCSP by using Normal approximation and second order conic inequalities.
The constraint set (3) can be expressed as follows:

\[ P \left( \sum_{i \in N} b_i R_i x_i \leq B \right) \Rightarrow P \left( \frac{\sum_{i \in N} [b_i R_i - b_i \mathbb{E}(R_i)] x_i}{\sqrt{\sum_{i \in N} b_i^2 \text{Var}(R_i) x_i^2}} \leq \frac{B - \sum_{i \in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i \in N} b_i^2 \text{Var}(R_i) x_i^2}} \right) \quad (9) \]

In Corollary 1, we show that \( H^* \) converges to the standard Normal distribution. Therefore, we can obtain the following probabilistic constraint:

\[ P \left( Z \leq \frac{B - \sum_{i \in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i \in N} b_i^2 \text{Var}(R_i) x_i^2}} \right) \geq \theta \Rightarrow \Phi \left( \frac{B - \sum_{i \in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i \in N} b_i^2 \text{Var}(R_i) x_i^2}} \right) \geq \theta \]

\[ \Rightarrow \frac{B - \sum_{i \in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i \in N} b_i^2 \text{Var}(R_i) x_i^2}} \geq \Phi^{-1}(\theta) \quad (10) \]

where \( Z \) is the standard normal random variable and \( \Phi(\cdot) \) and \( \Phi^{-1}(\cdot) \) are its c.d.f and quantile function, respectively. We reorganize (10) as follows:

\[ \sum_{i \in N} b_i \mathbb{E}(R_i) x_i + \Phi^{-1}(\theta) \sqrt{\sum_{i \in N} b_i^2 \text{Var}(R_i) x_i^2} \leq B \quad (11) \]

We assume high probability level (i.e. \( \theta \geq 0.90 \)), then \( \Phi^{-1}(\theta) > 0 \), which makes the constraint set (11) convex and it can be reformulated by second-order conic inequalities. The resulting deterministic equivalent reformulation of the CCSP model is a mixed-integer second-order cone program:

\[
\text{(SOCP) } \max \sum_{i \in N} s_i (1 - p_i) x_i \\
\text{s.t } \eta = \frac{B}{\Phi^{-1}(\theta)} - \frac{\sum_{i \in N} b_i \mathbb{E}(R_i) x_i}{\Phi^{-1}(\theta)} \\
\sum_{i \in N} b_i^2 \text{Var}(R_i) x_i^2 \leq \eta^2 \\
\eta \geq 0 \\
\text{Constraint sets (4) and (5)}
\]

The objective function to maximize is the expected total portfolio score. The conic reformulation of the constraint (11) is derived in constraint sets (12) and (13). \( \eta \) in equation (12) is an auxiliary variable for the linear portion of the constraint (11). The constraint (13) is a second-order cone generated by the constraint (11). \( \mathbb{E}(R_i) \) is derived in equation (6) and \( \text{Var}(R_i) \) is derived in equation (7).

2.4. Quality of Normal Approximation. The convergence quality of the true unknown distribution to Normal distribution is a significant concern for decision makers since project portfolio selection decisions rely on public financial resources. In the CCSP, \( \theta \) provides the probability level and conversely \( 1 - \theta \) specify the risk level. Decision makers prefer high
level of probability level (i.e. low risk level) in CCSPs. However, this probability level is not a precise value for the true unknown distribution and in fact it represents the exact probability level of the standard Normal distribution.

When we solve the CCSP with probability level of \( \theta \), what can we say about the real probability level of the solution? Is it greater than \( \theta \) or less than \( \theta \)? Is there any bound for the worst case probability level? In order to get insights about those questions, we apply Berry-Esseen theorem along with a simulation study. In the computer science literature, Aljuaid and Yanikomeroglu [2010] argue that Normal distribution may not provide a good approximation and they employ the Berry-Esseen theorem to quantify the error of Normal approximation in the aggregate interference power of large wireless networks. To the best of our knowledge, it is worth noting that the quality of Normal approximation in operations management problems has not been studied.

In Theorem 2 in Appendix A.2, maximum error of Normal approximation is given by Esseen [1956] in terms of probability levels. Berry–Esseen theorem states that for any realization on the probability space, the maximum difference between the true unknown distribution and the standard Normal distribution in terms of probability levels has a bound. Hence, this theorem assist to quantify the worst case error when the Normal approximation is employed in CCSPs. For our case, we need to derive moments of random variable \( R_i \) to apply Berry–Esseen theorem. In the following Corollary 2, we derive Berry-Esseen bound for any solution obtained by the SOCP.

**Corollary 2.** Let \( b_1 \), \( b_2 \), \( b_3 \), \ldots, \( b_n \) be i.i.d. random variables as defined in equation \((6)\) and \((7)\). Let \( G_n \) be the true c.d.f of \( H_n = \sum_{i \in N} [b_i R_i - b_i E(R_i)] x_i \) for a specific solution vector \( x \). Then Kolmogorov distance between \( G_n \) and \( \Phi^x \) (Normal approximation for the solution vector \( x \)) satisfies:

\[
D_{Kol}^x \leq C \left( \sum_{i=1}^{n} |b_i^3 E(R_i^3) - 3b_i^2 E(R_i^2) E(R_i) + 2b_i^3 |E(R_i)|^3| x_i \right)^{3/2} \left( \sum_{i=1}^{n} b_i^2 Var(R_i) x_i \right)
\]

where \( D_{Kol}^x = \sup_{z \in \mathbb{R}} |G_n(z) - \Phi^x(z)| \). Note that \( E(R_i) \), and \( E(R_i^2) \) are determined in equations \((6)\) and \((7)\). \( E(R_i^3) \) is also derived in proof by using equation in \((8)\).

**Proof.** See Appendix A.7. \( \Box \)

We can calculate a bound value for the maximum error of normal approximation by using the best estimate of \( C \). The best estimate of \( C \) is 0.56 obtained recently by Shevtsova [2010]. Berry-Esseen theorem gives the worst case absolute difference for the probability level. Therefore, we also perform a simulation study to evaluate the empirical probability level of our CCSP. In the next section, we give our computational study.

### 3. Computational Results

In this section, we report results of computational experiments for the proposed model. We solve model SOCP by using IBM ILOG CPLEX 12.6.2 via Concert Technology and C++.
Simulation study is conducted in Python by using Pandas and Numpy libraries. The source code is available at https://github.com/MusaCaglar. All experiments are conducted on a computer with processor Intel Core i5 2.2 GHz, 8.00 GB memory (RAM), 64-bit operating system, and Windows 10 Home. We set the solution time limit to three hours (i.e. 10800 CPU seconds).

In order to obtain a good approximation that captures the underlying stochastic expenditures, we first examined the historical data of a call. Then, we interacted with an expert in the budget department. Based on the Kolmogorov-Smirnov and Anderson Darling goodness of fit tests and the expert opinion, we selected truncated beta distributions to model expenditures (Figure 2). We used EasyFit software for fitting. Range parameters $\tau_1$ and $\tau_2$ of random variables $W_1$ and $W_2$ are set to 0.4 and 0.7, respectively.

We define problem size as the number of projects applying for the grant. We use four problem sizes (250, 500, 1000, 2000) to represent different practical cases in a research council. For each problem size, we conduct a $2^4$ full factorial design to assess impacts of different problem parameters. The four factors and their levels are presented in Table 1 by considering practical cases. The cancellation probability of project $i$ ($p_i$) is generated according to uniform distribution (U) between 0.01-0.1 (low level) and 0.01-0.2 (high level). The common probability ($q$) of successful completion of projects with under-spent budgets is set to 0.4 and 0.5. The budget fraction ($bf$) is the ratio of available budget over sum

**Table 1. Factor Values for each problem size**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Description</th>
<th>Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p_i$</td>
<td>Cancellation probability</td>
<td>U(0.01-0.1) U(0.01-0.2)</td>
</tr>
<tr>
<td>$q$</td>
<td>Probability of underutilized budget for successful projects</td>
<td>0.4 0.5</td>
</tr>
<tr>
<td>$bf$</td>
<td>Ratio of available budget over sum of all project budgets</td>
<td>0.1 0.2</td>
</tr>
<tr>
<td>$\theta$</td>
<td>Probability of the chance constraint</td>
<td>0.90 0.95</td>
</tr>
</tbody>
</table>
of all project budgets, which is set to 0.1 and 0.2. Note that the total available budget is determined as \( B = (\sum_{i \in N} b_i) \times b_f \). The probability of the chance constraint \((\theta)\) is set to 0.90 and 0.95.

In research councils, there are academic departments that manage project evaluation process. We assume that there are seven academic departments in our problem instances (i.e. TUBITAK case). We examined a historical data of projects in the 1001 program and set to percent of project applications and budgets (in 10,000 monetary units) for each research area in Table 2. Note that the budget of a project in a research area is generated according to different uniform distributions between given lower and upper bounds. We set the minimum acceptance rate for each academic department to 0.1 (i.e. \( a_j = 0.1 \forall j \in M \)). An aggregate project score \((s_i)\) is uniformly distributed in interval [10-25].

\[
\begin{array}{|c|c|c|c|}
\hline
\text{Academic Departments} & \text{Ratio of Number of Applications} & U_{lb} & U_{ub} & \text{Average} \\
\hline
\text{Environment, Atmosphere, Earth and Marine Sciences (D1)} & 0.1 & 21 & 36 & 28.5 \\
\text{Electrical, Electronics and Information (D2)} & 0.1 & 9 & 36 & 22.5 \\
\text{Engineering Sciences (D3)} & 0.15 & 7 & 36 & 21.5 \\
\text{Health Sciences (D4)} & 0.1 & 23 & 36 & 29.5 \\
\text{Social Sciences and Humanities (D5)} & 0.15 & 7 & 22 & 14.5 \\
\text{Basic Sciences (D6)} & 0.25 & 11 & 36 & 23.5 \\
\text{Agriculture, Forestry and Veterinary (D7)} & 0.15 & 14 & 36 & 25 \\
\hline
\end{array}
\]

We solve five random instances for each factor combination and problem size. Hence, we have \(4 \times 2 \times 5 = 320\) runs. For each solution, we generate 10,000 simulations for spending pattern to observe the empirical probability of our chance constraint and the budget utilization rate.

Factor effects for medium and large size problems are given in Tables 3 and 4, respectively. We analyze the percentage of instances solved to optimum (opt (%)), the optimality gap (gap (%), if any), CPU time performance (in seconds), the empirical probability level (EmpCL), the theoretical bound (i.e. Berry-Esseen bound (BEB)), and the empirical budget utilization rate (EmpBUR). Each row presents averages of \(2^4 \times 5 = 80\) instances. Note that we measure the mean and standard deviation of CPU time for the instances solved to optimum in the given time limit. We exclude CPU time of not optimally solved instances to prevent dominance effect of 10800 seconds on CPU time of optimally solved instances.

The problem size factor has significant effects on the percentage of instances solved to optimum, BEB, and the empirical budget utilization rate. As the problem size increases, the percentage of instances solved to optimum clearly decreases as expected (Figure 3). However, the optimality gap of the instances not solved to optimum is less than 0.25% (near to optimal), which is remarkable and practical for a large-scale nonlinear model. BEB becomes relatively tight for large-scale instances. The reason for this behavior is that as the number of supported projects increases, the true unknown distribution theoretically converges to the Normal distribution (i.e. the result of Corollary 1). As the problem size gets larger, the utilization rate approaches to 98% (Figure 3). This is due to the aggregate random expenditure effect of increased number of supported projects. We may expect
Table 3. Factor Effects for Medium Size Problems

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>opt gap (%)</th>
<th>CPU (seconds)</th>
<th>EmpCL</th>
<th>BEB</th>
<th>EmpBUR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean std. dev.</td>
<td>mean std. dev.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>problem size</td>
<td>250</td>
<td>100  - 419.9 896.4 0.9759 0.2970 0.9390</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>74  0.25 322.2 882.2 0.9774 0.2010 0.9549</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>86  0.33 541.5 939.5 0.9763 0.2915 0.9400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>88  0.17 217.6 810.4 0.9770 0.2065 0.9539</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U(0.01-0.1)</td>
<td>75  0.26 443.4 1018.0 0.9791 0.2987 0.9528</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U(0.01-0.2)</td>
<td>99  0.06 329.0 778.6 0.9743 0.1993 0.9411</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U(0.01-0.1)</td>
<td>64  0.09 779.3 1978.2 0.9841 0.2147 0.9638</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U(0.01-0.2)</td>
<td>60  0.05 818.8 1924.1 0.9809 0.1427 0.9545</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>91  0.19 362.1 907.0 0.9638 0.2471 0.9521</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>83  0.29 396.4 874.1 0.9895 0.2509 0.9417</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Factor Effects for Large Size Problems

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>opt gap (%)</th>
<th>CPU (seconds)</th>
<th>EmpCL</th>
<th>BEB</th>
<th>EmpBUR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mean std. dev.</td>
<td>mean std. dev.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>problem size</td>
<td>1000</td>
<td>63  0.11 909.8 2399.6 0.9825 0.1400 0.9654</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>60  0.04 684.9 1341.3 0.9892 0.0991 0.9731</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>60  0.10 953.0 2249.7 0.9811 0.2100 0.9542</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>64  0.04 653.0 1609.7 0.9839 0.1474 0.9641</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U(0.01-0.1)</td>
<td>64  0.09 779.3 1978.2 0.9841 0.2147 0.9638</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>U(0.01-0.2)</td>
<td>60  0.05 818.8 1924.1 0.9809 0.1427 0.9545</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>63  0.07 993.8 2257.8 0.9828 0.1785 0.9593</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>61  0.08 399.1 1555.7 0.9822 0.1788 0.9591</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>64  0.07 639.3 1620.3 0.9730 0.1775 0.963</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>60  0.07 967.6 2239.6 0.9920 0.1799 0.9555</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

that as the problem size further increases, the utilization rate asymptotically approaches to 100%. This observation also shows the importance of the proposed model for large-scale project portfolio selection problems under expenditure uncertainty. In other words, if the expenditure uncertainty is modeled by a CCSP along with an appropriate distribution, the budget utilization clearly improves when the number of supported projects in the portfolio gets larger.
Figure 3. Hardness of the instances and budget utilization rate

The budget fraction (bf) factor (the ratio of available budget over sum of all project budgets) has clear effects on CPU time and the Berry-Esseen bound (BEB). As the budget fraction decreases, the average CPU time increases. The reason for this behavior is due to the decrease in the available budget. If the available budget decreases, the competition among similar projects increases and finding feasible project combinations becomes much harder from a combinatorial optimization perspective. As the budget fraction increases, BEB decreases due to the increase in the number of supported projects (i.e. the true unknown distribution theoretically converges to Normal distribution).

The cancellation probability of each project (pi) has a significant effect on BEB. The reason for this behavior is that as pi values increases, the variance of the true unknown distribution increases and BEB decreases due to equation (15). The probability of underutilized budget (q) has no significant effect at all.

The probability of the chance constraint (θ) has clear effects on the empirical probability level as expected. When we solve instances with θ = 0.9, the average empirical probability level is close to 0.97. When we solve instances with θ = 0.95, the average empirical probability level is close to 0.99. High empirical probability levels are promising from a decision maker’s perspective. As expected, there is some differences between θ and empirical probability levels due to Normal Approximation.

To get more insights about the approximation error in terms of probability levels in special cases, in Figure 4, the theoretical bound (i.e. BEB) and the probability difference (i.e. EmpCL-θ) are presented for different factors. We find that problem size, bf, and pi are significant factors for the theoretical bound. Therefore, we include those factors to compare the theoretical bound and the probability difference. We clearly observe that Normal distribution gives a good approximation. Note that empirical probability differences are less than or equal to 0.08 for θ = 0.9 and they are less than or equal to 0.04 for θ = 0.95. The theoretical bound is relatively tight for large-scale problems. As expected, the theoretical bound and the probability difference also come very close as the bf and pi increase. It is interesting to note that the theoretical bound and the empirical probability are
the same in sub-figure 4b for problem size of 2000. The reason for this behavior is that if the number of supported projects and cancellation probabilities increase, the theoretical bound gives good information about empirical probability level. On the other side, the theoretical bound is loose for moderate-size problems (i.e. problem size of 250 and 500 projects). The theoretical bound is less than 0.1 for large-scale problems when $bf = 0.2, p_i \in U(0.01 - 0.2)$. The key takeaway is that when we solve our chance constrained model with the Normal approximation, the optimal solutions of our model eminently satisfy the desired probability level of the chance constraint (albeit with mild conservatism).

Overall, we observe that the proposed model solves practical size instances in a reasonable amount of CPU time. Therefore, public managers can apply the proposed approach to deal with uncertain expenditures in project portfolio selection problems to improve the budget utilization.

In the next subsection, the value of the proposed model is discussed.

3.1. What does modeling of uncertainty offer to decision makers? In this section, we quantify the value of the proposed model (PM). If aforementioned PM is not applied,
a model (DM) with the deterministic budget constraint will be solved. Hence, for problem size 2000, we solve our 5 instances as a DM with $s_i$ and $b_i$ parameters. Then, we calculate the expected score, the expected number of successfully completed projects, and the budget utilization (for $\theta=0.95$) of solutions for DM by using probability information of PM. After, we compare these values with those of PM for different factors by using following formula and report improvements (%) in Table 5.

\[
\text{Improvement (\%)} = 100 \times \frac{\text{Expected Value of PM} - \text{Expected Value of DM}}{\text{Expected Value of DM}} \tag{16}
\]

Since we have 4 factors and 5 instances, we have calculate $2^4 \times 5 = 80$ improvement values. Each row in Table 5 reports information of 40 comparisons. $E(\text{score})$ represent the expected total score of supported projects. $E(\text{nscp})$ represents the expected number of successfully completed projects. $E(\text{nscp})$ can be obtained as follows. By using the cancellation probability of each project ($p_i$), we can obtain the expected total number of cancellations. Let $I_i$ be a Bernoulli random variable with the success probability $p_i$ and define $\Gamma = \sum_{i \in N} I_i x_i$ that denotes number of cancellations. $\Gamma$ follows Poisson-Binomial distribution (Hong [2013]). Then, the expected number of canceled projects for a solution vector $x$ is: $E(\Gamma) = \sum_{i \in N} E(I_i) x_i = \sum_{i \in N} p_i x_i$. Therefore, $E(\text{nscp}) = ns - E(\Gamma)$, where $ns$ is the number of supported projects.

**Table 5.** Value of Modeling Uncertainty (Improvements (%))

<table>
<thead>
<tr>
<th>Factor</th>
<th>Levels</th>
<th>$E(\text{score})$</th>
<th>$E(\text{nscp})$</th>
<th>Budget Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Min</td>
<td>Avg</td>
<td>Max</td>
</tr>
<tr>
<td>$b_f$</td>
<td>0.1</td>
<td>6.5</td>
<td>9.3</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>6.7</td>
<td>9.2</td>
<td>11.9</td>
</tr>
<tr>
<td>$p_i$</td>
<td>U(0.01-0.1)</td>
<td>6.5</td>
<td>7.6</td>
<td>8.7</td>
</tr>
<tr>
<td></td>
<td>U(0.01-0.2)</td>
<td>9.4</td>
<td>10.9</td>
<td>12.5</td>
</tr>
<tr>
<td>$q$</td>
<td>0.4</td>
<td>6.5</td>
<td>8.6</td>
<td>11.1</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>7.8</td>
<td>9.9</td>
<td>12.5</td>
</tr>
<tr>
<td>$\theta$</td>
<td>0.9</td>
<td>7.0</td>
<td>9.4</td>
<td>12.5</td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>6.5</td>
<td>9.0</td>
<td>11.8</td>
</tr>
</tbody>
</table>

Factors $p_i$ and $q$ have a notable effect on improvement values. The value of PM increases as those probability parameters increase since unused budgets of many projects can constitute large amounts that make additional room for the budget constraint. In overall, adopting the PM yields significant increases on the expected score, the expected number of successfully completed projects, and the budget utilization. For instance, it provides a minimum 6.5%, and a maximum 12.5% improvement in the expected total score. Besides, improvement of the expected number of successfully completed projects is minimum 6.5%, and maximum
14.3%. Moreover, the budget utilization improvement is minimum 8.0%, and maximum 15.2%.

4. Conclusions

In this study, we address a project portfolio selection problem under expenditure uncertainty and policy constraints. We are motivated by a research council to develop our model. We have made several contributions to the literature on project portfolio selection.

First, the input modeling in chance-constrained optimization models has got little attention in the literature. We pay special importance to the input modeling as the practical information suggests. We model the expenditure uncertainty with a mixture distribution. To the best of our knowledge, no similar study has been conducted in the literature on project portfolio selection.

Second, after discussing the challenges of proposed approximation methods in the literature on CCSP, we have shown that Normal distribution can be used to approximate the proposed model. We formulate the deterministic equivalent model via Normal approximation and second-order cone programming. We find that our large-scale non-linear model under policy constraints can be solved to optimum (or near optimum) by commercial solvers such as CPLEX in a reasonable amount of time. We believe that this is a significant finding from a practical standpoint.

Third, the key concern for applying Normal distribution is the quality of the approximation. We give managerial insights by applying the theoretical bound (i.e. Berry Esseen theorem) and computing the empirical probability of the chance constraint via simulation. We find that Normal distribution gives a good approximation and the theoretical bound is relatively tight for large-scale problems. Our main takeaway is that when we solve our chance constrained model with the Normal approximation, the optimal solutions of our model eminently satisfy the desired probability level of the chance constraint (albeit with mild conservatism). Surprisingly, to the best of our knowledge, there is no study that deals with the theoretical and empirical approximation error of Normal distribution for chance-constrained problems in the operations management literature.

We also conduct an analysis to show the value of proposed model by comparing it with a deterministic model. The proposed approach can increase the budget utilization between 8.0% and 15.2%, which is remarkable for public decision makers. Thereby, more R&D projects could be supported and a higher socio-economic impact can be achieved.

Our research can be applied by research councils to improve the budget utilization to support more R&D projects. In the practice, there are extensive data sets of completed projects for different grant programs. Research councils can elaborately analyze those data sets for the input modeling of expenditures. Alternatively, different types of mixture distributions can be adopted according to the expenditure data. Normal approximation can be used due to the computational tractability. However, the empirical experiments (i.e. simulation) and theoretical bound computations (i.e. Berry Esseen theorem) should be carried out to gain managerial insights about the approximation error.

Finally, in the proposed model, we define policy constraints for the fairness among different research areas. Research councils can formulate additional policy constraints according to
the grant program and those policy constrains can be added to the model. As a result, there can be various extensions of the proposed model and those extensions can be tailored according the grant program policy. Extended models can also be solved by commercial solvers such as CPLEX in a reasonable amount time.

References


A.1. Definitions.

**Definition 1.** Given a finite set of \( (J) \) distribution functions. Let \( f_j(x) \) and \( F_j(x) \) be the p.d.f. and c.d.f. of the distribution \( j \) respectively. Let \( w_j \) be the probability of selecting distribution \( j \) such that \( w_j > 0 \) and \( \sum_j w_j = 1 \). Then p.d.f and c.d.f of the mixture random variable \( X \), \( f(x) \) and \( F(x) \) can be expressed as follows:

\[
f(x) = \sum_{j \in J} w_j f_j(x) \tag{17}
\]
\[
F(x) = \sum_{j \in J} w_j F_j(x) \tag{18}
\]

**Definition 2.** Let \( W \) be a truncated (bounded) beta distribution in open interval \( (a,b) \) with shape parameters \( \alpha \) and \( \beta \), its p.d.f and c.d.f are given as follows (see Johnson et al. [1995], Chapter 25):

\[
f_W(w; a, b, \alpha, \beta) = \frac{\Gamma(\alpha + \beta)(w - a)^{\alpha - 1}(b - w)^{\beta - 1}}{\Gamma(\alpha)\Gamma(\beta)(b - a)\alpha + \beta - 1} \quad \text{for } a < W < b \tag{19}
\]
\[
P(W \leq w) = F_W(w) = \int_a^w \frac{\Gamma(\alpha + \beta)(w - a)^{\alpha - 1}(b - w)^{\beta - 1}}{\Gamma(\alpha)\Gamma(\beta)(b - a)\alpha + \beta - 1} dw \tag{20}
\]

where the Gamma function is defined as: \( \Gamma(u) = \int_0^\infty s^{u-1}e^{-s}ds \).

**Definition 3.** Let \( X \) be a mixture random variable as defined in Definition 1. \( n^{th} \) moment of \( X \) is formulated as follows:

\[
\mathbb{E}[X^n] = \int_{-\infty}^{+\infty} x^n f(x)dx = \int_{-\infty}^{+\infty} x^n \sum_{j \in J} w_j f_j(x)dx
\]
\[
= \sum_{j \in J} w_j \int_{-\infty}^{+\infty} x^n f_j(x)dx = \sum_{j \in J} w_j m_j^n \tag{21}
\]

where \( m_j^n \) is the \( n^{th} \) moment of distribution \( j \).

A.2. Theorems.

**Theorem 1.** Lyapunov’s Central Limit Theorem (CLT) for i.i.d. random variables (Baurer [1996] and Shapiro et al. [2009]): Let \( X_1, X_2, ..., X_n \) be i.i.d. random variables with finite expectation \( \mathbb{E}(X_i) \), positive variance \( \text{Var}(X_i) \) and finite moments \( \mathbb{E}(X_i^{2+\delta}) \) for \( \delta > 0 \). Then, if for some \( \delta > 0 \), the following condition
\[
\lim_{n \to \infty} \sum_{i=1}^{n} \frac{\mathbb{E} \left[ |X_i - \mathbb{E}(X_i)|^{2+\delta} \right]}{\left[ \sqrt{\sum_{i=1}^{n} \text{Var}(X_i)} \right]^{2+\delta}} = 0
\]  

(22)

is satisfied, then normalized summand \( S_n = \frac{\sum_{i=1}^{n} \mathbb{E} \left[ |X_i - \mathbb{E}(X_i)| \right]}{\sqrt{\sum_{i=1}^{n} \text{Var}(X_i)}} \) converges to standard normal distribution \( (N(0,1)) \) as \( n \) goes to infinity.

**Theorem 2.** Berry–Esseen theorem for the maximum error of normal approximation: Let \( Y_1, Y_2, \ldots, Y_n \) be i.i.d. random variables with \( \mathbb{E}(Y_i) = 0 \), positive second moment \( \mathbb{E}(Y_i^2) \) and a finite third moment \( \mathbb{E}(Y_i^3) < \infty \). Let \( Q_n = \frac{\sum_{i=1}^{n} Y_i}{\sqrt{\sum_{i=1}^{n} \mathbb{E}(Y_i^2)}} \), \( G_n \) is the c.d.f of \( Q_n \), \( \Phi \) is the c.d.f. of the standard normal distribution. Then, for the Kolmogorov distance is defined by

\[
D_{K\phi} = \sup_{z \in \mathbb{R}} |G_n(z) - \Phi(z)|.
\]  

(23)

there exists a constant \( C \), such that \( D_{K\phi} \leq C \psi \) where

\[
\psi = \left( \sum_{i=1}^{n} \mathbb{E}(|Y_i^3|) \right)^{-3/2} \left( \sum_{i=1}^{n} \mathbb{E}(Y_i^2) \right).
\]  

(24)

**Remark 1.** Esseen [1956] theoretically showed that the constant \( C \) satisfies

\[
7.59 \geq C \geq \frac{\sqrt{10} + 3}{6\sqrt{2\pi}} \approx 0.4097
\]

However, the best estimate on upper bound substantially improved by researchers over past decades. Shevtsova [2010] show that the best estimate of upper bound on \( C \) is 0.56.

**Remark 2.** Berry–Esseen’s theorem depends on only the first three moments to give the upper bound on the maximum error of normal approximation.

### A.3. Proof of Proposition 1.

**Proof.** Since the sequence is truncated, we can write \( \mathbb{E}(T_i) < u^k < \infty \) for \( i = 1, \ldots, n \). We can also write, \( |T_i - \mathbb{E}(T_i)| \leq u - l \) is true for every \( i = 1, \ldots, n \). Then for each \( \delta > 0 \), we can determine an upper bound for the term in (22) as follows:

\[
\frac{\sum_{i=1}^{n} \mathbb{E} \left[ |T_i - \mathbb{E}(T_i)|^{2+\delta} \right]}{\left[ \sqrt{\sum_{i=1}^{n} \text{Var}(T_i)} \right]^{2+\delta}} \leq \frac{(u-l)^\delta \sum_{i=1}^{n} \mathbb{E} \left[ |T_i - \mathbb{E}(T_i)| \right]}{\sum_{i=1}^{n} \text{Var}(T_i)} \left[ \sqrt{\sum_{i=1}^{n} \text{Var}(T_i)} \right]^{\delta}
\]  

(25)

We also know that \( \mathbb{E} \left[ |T_i - \mathbb{E}(T_i)|^2 \right] = \mathbb{E} \left[ T_i^2 - 2T_i\mathbb{E}(T_i) + (\mathbb{E}(T_i))^2 \right] = \mathbb{E}(T_i^2) - (\mathbb{E}(T_i))^2 \), which is the definition of \( \text{Var}(T_i) \). Then we can simplify the upper bound in (25) as follows:
\[
\frac{\sum_{i=1}^{n} \mathbb{E} \left[ (T_i - \mathbb{E}(T_i))^2 \right]}{\sqrt{\sum_{i=1}^{n} \text{Var}(T_i)}} \leq \left( \frac{u - l}{\sqrt{\sum_{i=1}^{n} \text{Var}(T_i)}} \right)^{\delta} \tag{26}
\]

Then, for every \( \delta > 0 \), the right hand side of (26) converges to zero as \( n \) goes to infinity. Since the upper bound converges to zero, then the term in the left hand side of (26) converges to zero. Therefore, Lyapunov CLT theorem is satisfied and \( \sum_{i=1}^{n} |T_i - \mathbb{E}(T_i)| \) converges to \( N(0, 1) \) as \( n \) goes to infinity, which completes proof. \( \square \)


*Proof*. \( \hat{b}_i = b_i R_i \) and we know that \( R_i \) is bounded in the interval \((0, 1] \) and we also know that usually R&D programs have a specified budget upper bound (let call it \( b_{\text{max}} \)) for applying projects. We can conclude that \( b_i \) is a truncated random variable in the interval \((0, b_{\text{max}}] \) for every \( i \). Hence, any subset of sequence \( b_1 R_1, b_2 R_2, ..., b_i R_i \in N \) that has statistically significant number of elements satisfies Lyapunov CLT theorem. \( \square \)

A.5. Proposition 3 and its proof.

**Proposition 3.** The mean, variance and \( n^{th} \) moment of the truncated beta random variable \( W \) defined in open interval \((a, b) \) are given as follows:

\[
\mathbb{E}(W) = \frac{ab + \beta a}{\alpha + \beta} \tag{27}
\]

\[
\text{Var}(W) = \frac{(b - a)^2 \alpha \beta}{(\alpha + \beta)^2 (\alpha + \beta + 1)} \tag{28}
\]

\[
\mathbb{E}(W^n) = \begin{cases} 
\sum_{k=0}^{n-k} \left( \frac{n! \Gamma(\alpha + \beta) \Gamma(a + n + k) \Gamma(a + \beta + n - k) - (b - a)^{n-k} \Gamma(\alpha + \beta) \Gamma(a + \beta + n - k) \Gamma(a) \Gamma(a + \beta + n)}{\Gamma(\alpha + \beta + n) \Gamma(a + \beta + n)} \right) & \text{if } 0 < a < b \\
\frac{b^n \Gamma(\alpha + \beta) \Gamma(a + n)}{\Gamma(\alpha + \beta + n)} & \text{if } a = 0 \text{ and } b > 0 
\end{cases} \tag{29}
\]

*Proof*. Let \( W \) be a truncated beta distribution as defined in Definition 2. Then define a transformed random variable \( T \) such that \( T = \frac{W - a}{b - a} \) where \( T \) is the standard beta distribution in open interval \((0, 1) \). We can write that \( W = a + (b - a)T \). We know the mean, variance and \( n^{th} \) moment of standard random variable \( T \) (see Johnson et al. [1995], Chapter 25). Therefore, we can obtain the mean, variance and \( n^{th} \) moment of truncated beta random variable \( W \) by using expectation or variance operator as follows:

\[
\mathbb{E}(W) = \mathbb{E}(a + (b - a)T) = a + (b - a)\mathbb{E}(T) = a + \frac{(b - a)\alpha}{\alpha + \beta} = \frac{ab + \beta a}{\alpha + \beta}
\]

\[
\text{Var}(W) = \text{Var}(a + (b - a)T) = \text{Var}((b - a)T) = (b - a)^2 \text{Var}(T)
\]

Proof. By using equation (21), we can write $\mathbb{E}(R_i) = \sum_{j=1}^{j=3} w_j m_j$ where $m_j^1$ is the mean (first moment) of the truncated distribution in open interval $(0, \tau_1)$, $m_j^2$ is the mean of the truncated distribution in open interval $(\tau_2, 1)$, $m_j^3$ is the mean of the degenerate distribution at point 1, by using equation in (27), we can obtain first moments (means) so that $m_j^1 = \frac{\alpha_1^{\tau_1}}{\alpha_1 + \beta_1}$ and $m_j^2 = \frac{\alpha_2 + \beta_2}{\alpha_2 + \beta_2}$, since mean of the constant value is itself, then $m_j^1 = 1$. Apparently, $w_1 = p_i$, $w_2 = q$, and $w_3 = (1-p_i-q)$. Therefore, we obtain: $\mathbb{E}(R_i) = p_i \frac{\alpha_1^{\tau_1}}{\alpha_1 + \beta_1} + q \frac{\alpha_2 + \beta_2}{\alpha_2 + \beta_2} + (1-p_i - q)$. We know $Var(R_i) = \mathbb{E}(R_i^2) - [\mathbb{E}(R_i)]^2$. By using equation (21), we can write $\mathbb{E}(R_i^2) = \sum_{j=1}^{j=3} w_j m_j^2$ where $m_j^2$ is the second moment of the truncated distribution in open interval $(0, \tau_1)$, $m_j^2$ is the second moment of the truncated distribution in open interval $(\tau_2, 1)$, $m_j^3$ is the second moment of the degenerate distribution at point 1. Let $W_1$ is the truncated beta random variable in interval $(0, \tau_1)$ and $W_2$ is the truncated beta random variable in interval $(\tau_2, 1)$. We know that $m_k^2 = \mathbb{E}(W_k^2) = [\mathbb{E}(W_k)]^2 + Var(W_k)$ for $k = 1, 2$ and $m_3^2 = 1$. Then we can write:

$$Var(R_i) = \mathbb{E}(R_i^2) - [\mathbb{E}(R_i)]^2 = p_i \mathbb{E}(W_1^2) + q \mathbb{E}(W_2^2) + (1-p_i-q) - [\mathbb{E}(R_i)]^2$$

$$= p_i ( [\mathbb{E}(W_1)]^2 + Var(W_1)) + q ([\mathbb{E}(W_2)]^2 + Var(W_2)) + (1-p_i-q) - [\mathbb{E}(R_i)]^2$$

(30)
Hence, by using equations in (27), (28) and (6), we derive:

\[
Var(R_i) = p_i \left[ \frac{\alpha_1 \tau_1}{\alpha_1 + \beta_1} + \frac{\tau_2 \alpha_1 \beta_1}{(\alpha_1 + \beta_1)^2(\alpha_1 + \beta_1 + 1)} \right] + q \left[ \frac{\alpha_2 + \beta_2 \tau_2}{\alpha_2 + \beta_2} + \frac{(1 - \tau_2)^2 \alpha_2 \beta_2}{(\alpha_2 + \beta_2)^2(\alpha_2 + \beta_2 + 1)} \right] + (1 - p_i - q)
\]

By directly applying equations in (21) and (29) together, we derive:

\[
E(R_i^n) = p_i \left[ \frac{\tau_1 \Gamma(\alpha_1 + \beta_1) \Gamma(\alpha_1 + n)}{(\tau_1 + \Gamma(\alpha_1 + \beta_1 + n))} \right] + (1 - p_i - q)
\]

\[
\quad + q \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \frac{\tau_2 (1 - \tau_2)^{n-k} \Gamma(\alpha_2 + \beta_2) \Gamma(\alpha_2 + n - k)}{\Gamma(\alpha_2) \Gamma(\alpha_2 + \beta_2 + n - k)}
\]

\[
\Box
\]


Proof. Define \(Y_i = \hat{b}_i - \mathbb{E}(\hat{b}_i)\). Then,

\[
\mathbb{E}(Y_i) = 0
\]

\[
\mathbb{E}(Y_i^2) = \mathbb{E}\left[(\hat{b}_i - \mathbb{E}(\hat{b}_i))^2\right] = \mathbb{E}\left[\hat{b}_i^2 - 2\hat{b}_i \mathbb{E}(\hat{b}_i) + (\mathbb{E}(\hat{b}_i))^2\right] = \mathbb{E}(\hat{b}_i^2) - (\mathbb{E}(\hat{b}_i))^2 = \text{Var}(\hat{b}_i) + \mathbb{E}(\text{Var}(R_i))
\]

\[
E(|Y_i^3|) = \mathbb{E}\left(|\hat{b}_i - \mathbb{E}(\hat{b}_i)|^3\right) = \mathbb{E}\left[|\hat{b}_i^3 - 3\hat{b}_i^2 \mathbb{E}(\hat{b}_i) + 3\hat{b}_i(\mathbb{E}(\hat{b}_i))^2 - (\mathbb{E}(\hat{b}_i))^3|\right]
\]

\[
\quad = |\mathbb{E}(\hat{b}_i^3) - 3\mathbb{E}(\hat{b}_i^2) \mathbb{E}(\hat{b}_i) + 2(\mathbb{E}(\hat{b}_i))^3| = |\mathbb{E}(\hat{b}_i^3) - 3\mathbb{E}(R_i^3) + 3\mathbb{E}(R_i^2) \mathbb{E}(R_i) + 2\mathbb{E}(R_i^3)|
\]

By using equation (8) and the property of gamma function such that \(\Gamma(t + 1) = t\Gamma(t)\), we derive:

\[
E(R_i^3) = p_i \left[ \frac{\tau_1 \Gamma(\alpha_1 + \beta_1) \Gamma(\alpha_1 + 3)}{(\tau_1 + \Gamma(\alpha_1 + \beta_1 + 3))} \right] + (1 - p_i - q)
\]

\[
\quad + q \sum_{k=0}^{3} \frac{3!}{k!(3-k)!} \frac{\tau_2 (1 - \tau_2)^{3-k} \Gamma(\alpha_2 + \beta_2) \Gamma(\alpha_2 + 3 - k)}{\Gamma(\alpha_2) \Gamma(\alpha_2 + \beta_2 + 3 - k)}
\]

\[
\quad = p_i \left[ \frac{\tau_1 \Gamma(\alpha_1 + \beta_1) \Gamma(\alpha_1 + 3)}{(\alpha_1 + \beta_1 + 1)(\alpha_1 + \beta_1 + 2)(\alpha_1 + \beta_1 + 3)} \right] + (1 - p_i - q)
\]

\[
\quad + q(1 - \tau_2)^3 \frac{(\alpha_2 + 2)(\alpha_2 + 1) \alpha_2}{(\alpha_2 + \beta_2 + 2)(\alpha_2 + \beta_2 + 1)(\alpha_2 + \beta_2)}
\]
By using equations (23), (24) and Remark 1, we can obtain inequality (15). This completes the proof. □