PUBLIC R&D PROJECT PORTFOLIO SELECTION UNDER EXPENDITURE UNCERTAINTY

ABSTRACT. We consider a project portfolio selection problem faced by research councils in project and call-based R&D grant programs. In such programs, typically, each applicant project receives a score value during specially-designed peer review processes. Each project also has a certain budget, estimated by its principle investigator. The problem is to select an optimal (maximum total score) subset of applicant projects under a budget constraint for the call. At the time of funding decisions, exact expenditures of projects are not known. The research councils typically don't provide more money than they funded a project to start with, so the realized total expenditure of a portfolio usually tends to be lower than the total budget, which causes budgetary slack. In this paper, we attempt to model this phenomenon in a project portfolio selection problem and show that budget utilization of a call can be increased to support more projects and hence achieve higher scientific impact. We model a project's expenditure using a mixture distribution that represents project success, underspending and cancellation situations. We develop a chance-constrained model with policy constraints. Due to the intractability of the developed model, we have shown that Normal distribution can be used for approximation. We also quantify the approximation error of our model via a theoretical bound and simulation. The proposed approach could rigorously increase the budget utilization up to 15.2%, which is remarkable for public decision makers.

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

1. Introduction

The inherent uncertainty in R&D projects makes it hard to know their exact expenditures in advance and project portfolio selection decisions are usually based on estimated project budgets. Typically, a public R&D grant program does not allow project budget overrun in any circumstance. Therefore, underspending of allocated budget is prevalent among public R&D projects. As a result, budgetary slack occurs and we face inefficient use of grant program budgets. In this paper, we attempt to model project expenditure uncertainty in a public R&D project portfolio selection problem. Our target is to improve program budget utilization which leads to supporting more projects and achieving higher scientific and technological output.

Research councils (or public funding agencies) distribute government funds to R&D projects usually through grant programs and via call-based systems. 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 some merit review criteria. In a grant program, funding decisions of projects are made according to project scores (i.e. ex ante value estimates by peer reviewers [1]), budgets of projects (i.e. ex ante expenditure estimates), policy constraints (i.e. constrains formulated according to the 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 portfolio of projects, 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 ([2], [3]).

If budget overrun is not allowed, which is often the case in public R&D grants, then the applicants usually tend to include some buffer in their budget, so that they can cover unexpected costs such as an increase in the price of a required equipment in their research. It is also worth noting that principle investigators (PI) and project team members usually get compensation from project budgets. Antle and Eppen [4] 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 [5]). Hu and Szmerekovsky [6] highlight that subordinates have a tendency of building budgetary slack by amplifying costs. As a result, it is highly likely that the realized total expenditure of the projects in a portfolio is lower than the total allocated budget, causing budgetary slack.

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. We give the total appropriations (i.e. approved budgets) and expenditures of TUBITAK funding programs between 2010-2017 in Figure 1. TUBITAK funding programs experienced a serious budgetary slack (i.e. unspent budget) rate of 41% in 2012. It is worth noting that the approval rates for submitted projects at TUBITAK are between 10%-20% depending on the program and there is a high level of competition among project applications. When we evaluate unspent ratios in terms of magnitude, we can see that even 5% in 2015 amounts 65 million Turkish Liras (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 a program's total budget could be used efficiently, then the total budget of a research council will be utilized at its maximum inline with the socio-economic goals of the council.

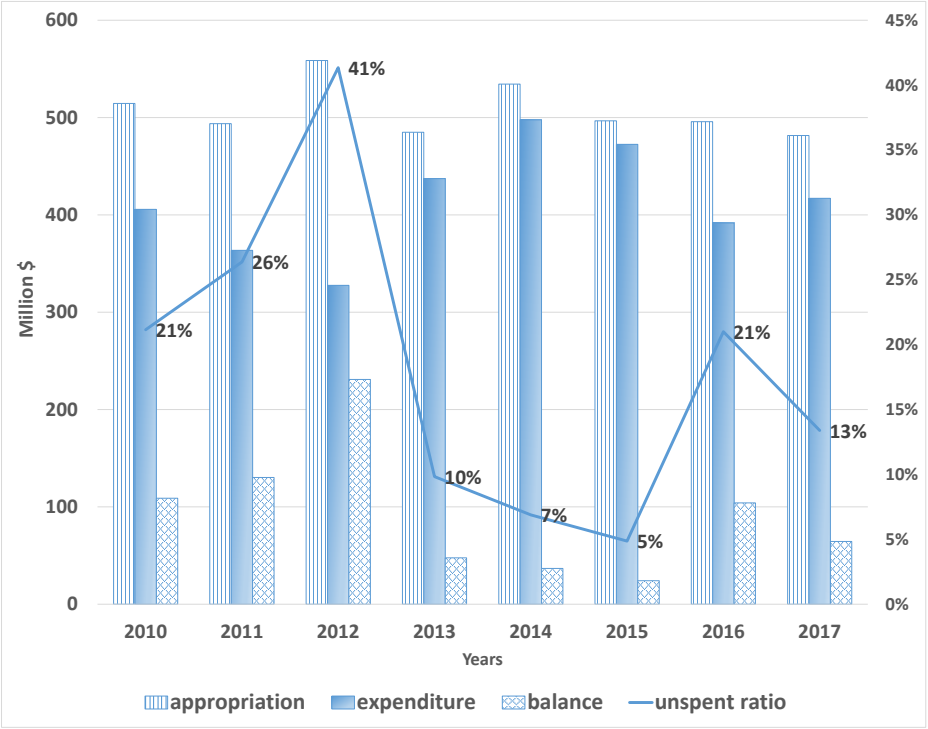


FIGURE 1. TUBITAK Funding Budget
Notes. Authors' own compilation from Activity Reports of TUBITAK (TUBITAK [7]).

There are two main determinants causing a budgetary slack in a research funding program. First, a significant number of granted projects end up successfully providing their expected benefits without spending their whole budget, as discussed before. Second, a certain number of projects are canceled before spending most of their budgets. Rules and conditions for cancellation of projects are usually arranged in the R&D program regulations (see NSF [8], NIH [9]). For instance, PI and other project members may fail to comply with the approved project plan, which may result with the cancellation of a project. If a project is canceled, unspent amount of

the released fund is returned to the research council. Canceled projects are regarded as unsuccessful and the expected scientific and technological benefit of them aren't realized.

In this study, we focus on modeling project expenditure uncertainty in a project portfolio selection problem. For a project, three cases can occur. The project may be completed successfully with whole budget spent or completed successfully with budget underspending or canceled with budget underspending. For the last two cases, proportion of the budget spent can be modeled using truncated beta distributions. Standard or truncated beta distribution is a reasonable choice for modeling ratios or proportions [10]. Hence, a project's expenditure can be modeled as a mixture distribution which considers budget spending of a project for both successful completion and cancellation situations. Then, we model the budget constraint of portfolio selection problem as a chance constraint, since total budget spending is the sum of expenditure random variables of selected projects. We show that the total expenditure of selected projects can be approximated by a Normal distribution which enables solving the problem using off-the-shelf optimization software. We show that Normal distribution gives a good approximation and the theoretical bound is relatively tight for large-scale problems. We also numerically show that the budget utilization rate of 100% is hard to achieve but that of 94-97% is within reach. The proposed approach could increase the budget utilization by 8.0% and 15.2%, which is remarkable for public decision makers.

1.1. **Related Literature.** We briefly review the relevant literature of project selection. There are mainly two approaches to select promising projects. First, selection decisions are made individually (i.e. one project at a time). Projects are evaluated by initial filtering mechanisms, which usually depend on financial and risk measures

of projects. Remaining projects that pass initial filtering are assessed in detail by quantitative and/or qualitative selection criteria. This approach is referred to as sequential project selection (Henriksen and Traynor [1] and Meade and Presley [11]). Some researchers have criticized this sequential approach because it does not necessarily achieve optimal project portfolios (Chien [12], Hall et al. [2], and Marcondes et al. [13]).

Second, selection decisions are made simultaneously through project portfolio optimization. In this approach, an objective function defines overall portfolio performance including project score values. The total available budget and policy restrictions are formulated through constraints of an optimization model. If project scores and budgets are deterministic, the optimization model boils down to a variant of well-known knapsack model (Kellerer et al. [14] and Dickinson et al. [15])

On the other hand, project portfolio optimization becomes complicated when there is an uncertainty in problem parameters (Medaglia et al. [16],Koç et al. [17], Solak et al. [18], and Çağlar and Gürel [19]). We refer the interested reader to Chien [12] and Kavadias and Chao [20], which provide reviews of the project portfolio selection literature. 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 relatively scant.

Existing studies of the problem in industry deal with small numbers of projects (see Solak et al. [18] and Medaglia et al. [16] 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

(Arratia et al. [21] and Çağlar and Gürel [22]). For instance, the number of project applications in nationwide programs of TUBITAK varies between 250 and 2000. A need for large-scale project selection models including chance constraints of expenditures is mentioned by Gabriel et al. [23]. Hall [24] emphasizes that one of the future research directions on project portfolio optimization could focus on the effect of uncertain parameters of projects. Marcondes et al. [13] have recently underlined that practical constraint formulations (i.e. budget and other policy restrictions) should be integrated into portfolio selection models. Motivated by those highlights of the literature, in this study, we develop a model for moderate and large-scale project selection problems under project expenditure uncertainty and policy constraints.

In the project portfolio optimization literature, several researchers consider project cancellations. Solak et al. [18] develop a multi-stage model in an industrial setting. They study optimal allocation of budgets to a set 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 [19] study cancellations in a public project portfolio selection problem. They develop mixed integer and dynamic programming approaches to maximize total score of selected projects assuming that a number of projects may be canceled. In this study, in a project portfolio selection problem, we consider underspending of budgets by successfully completed projects in addition to budgetary slack created by canceled projects.

In the next section, we give the problem definition, define probability distribution for project expenditure and give a mathematical model for project portfolio selection. In Section 3, we present the results of a numerical study and discuss insights obtained. Finally, we give some concluding remarks in Section 4.

2. Problem Formulation

We address a project portfolio selection problem typically faced by a research council in a project and call based grant program. Let N be the set of solicited project proposals for a call. Research councils usually classify these projects according to their area of research. M is the set of scientific areas in the program. We denote the set of projects in area j by N_j. Peer reviewers in panels evaluate the scientific and technological values of projects and assign a score (s_i) to each project i according to a set of criteria. We consider an aggregate scoring approach here, since it is currently applied in many research councils such as TUBITAK. Namely, each reviewer assigns a score for each criteria under consideration, but at the end scores from all reviewers for all criteria aggregates to a single score value (s_i). Each project has a budget b_i, estimated by the applicant. The budget of a project covers all of the planned expenditure during its duration. The duration of a project can be any value that is less than or equal to the maximum project duration of a program. The project budget is released for the use of PI in several installments. There are reporting time points at which projects are reviewed for compliance with the terms and conditions of the program. Decisions regarding to project cancellations and budget releases for a project are made at those time points. Let random variable $\hat{b}_i$ denote the real expenditure of project i if the project is selected and implemented. We assume that $\hat{b}_i$'s are independent. $\hat{b}_i$ will be within the interval $(0, b_i]$ as we assume that budget overrun is not allowed. Thus, we define a ratio $R_i = \frac{\hat{b}_i}{b_i}$ to model budget utilization ratio of a project i. We denote the total available budget for the call as B.

As discussed before, a granted project can be canceled by its PI or by the research council. In our model, we introduce a cancellation probability p_i for each project i. The value of p_i can be estimated by expert judgments and/or past data. For example,

in panel reviews there is usually a criterion that evaluates the feasibility of a project. This evaluation focuses on possibilities and limitations in terms of human resources, project management, etc. 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.

In project portfolio selection decisions, another concern is the fairness among the project proposals from different scientific areas, which we call policy constraints. One approach to achieve fairness for a scientific area is to ensure that a certain proportion of selected projects belongs to that area. So, we define a parameter a_j, which is the minimum proportion of all accepted projects that will belong to area j.

The problem is to select a portfolio with maximum total expected score under a budget constraint and policy constraints. For each project, we define a 0-1 decision variable x_i which equals 1 if project i is selected and 0, otherwise. We formulate the problem as below:

$$\text{(PPS)} \ \max \quad \sum_{i \in N} s_i(1-p_i)x_i \tag{1}$$

$$\text{s.t.} \ P\left(\sum_{i \in N} b_i R_i x_i \leq B\right) \geq \theta \tag{2}$$

$$\sum_{i \in N_j} x_i \geq a_j \sum_{i \in N} x_i \quad \forall j \in M \tag{3}$$

$$x_i \in \{0,1\} \qquad \forall i \in N \tag{4}$$

The objective function (1) to maximize is the total expected score of selected projects. With probability (1-p_i), the project will be completed successfully and its scientific value will be realized, so $s_i(1-p_i)$ gives the expected score of project i. (2) is a chance constraint and provides that the sum of random budget expenditure

of supported projects does not exceed the available budget (B) with a probability level of θ, where θ is a predefined value by the decision maker. In this constraint, $b_i R_i$ gives the realized expenditure for project i and $P(\cdot)$ denotes the probability measure. Constraint set (3) defines a minimum acceptance rate for projects applying from area j. 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.

2.1. **Modeling R_i.** As mentioned in Section 1, we consider three possible cases that can occur for an accepted project i:

(1) cancellation of project i and returning the remainder of its budget to research council, with probability p_i
(2) successful completion of project i and underspending of its budget, with probability q
(3) successful completion of project i that fully used its budget, with probability $1 - p_i - q$

Probability that a project will be completed successfully with underspent budget (q) can be obtained from past data of successfully finished projects. The expenditure ratio (W_1) of a canceled project (case (1)) can be modeled by a beta distribution in interval $(0, \tau_1)$, where $\tau_1 < 1$. Beta distribution is known for its flexibility in modeling various distributional shapes and is frequently used to model uncertain fractional quantities. Similarly, expenditure ratio (W_2) of a successful project can be modeled by a beta distribution in interval $(\tau_2, 1)$, where $\tau_2 > 0$.

We formulate R_i as a mixture distribution according to the aforementioned three cases for project expenditure. For instance, if we assume that $\tau_1 < \tau_2$, i.e. a canceled

project's budget utilization is always less than a successful one, then the probability density function (p.d.f) of the random variable R_i is as follows:

$$f_{R_i}(r) = \begin{cases} p_i \cdot f_{W_1}(r; \alpha_1, \beta_1, 0, \tau_1) & \text{if } r \in (0, \tau_1) \\ q \cdot f_{W_2}(r; \alpha_2, \beta_2, \tau_2, 1) & \text{if } r \in (\tau_2, 1) \\ 1 - p_i - q & \text{if } r = 1 \end{cases} \tag{5}$$

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 α_1, β_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 α_2, β_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). We derive the cumulative distribution function (c.d.f.) of R_i with p.d.f (5) as below:

$$\begin{aligned} P(R_i \leq k) = F_{R_i}(k) = p_i \int_0^{min(\tau_1, k)} \frac{\Gamma(\alpha_1 + \beta_1) r^{\alpha_1 - 1} (\tau_1 - r)^{\beta_1 - 1}}{\Gamma(\alpha_1)\Gamma(\beta_1)\tau_1^{\alpha_1 + \beta_1 - 1}} dr + \\ q \int_{\tau_2}^{min(1,k)} \frac{\Gamma(\alpha_2 + \beta_2)(r - \tau_2)^{\alpha_2 - 1}(1 - r)^{\beta_2 - 1}}{\Gamma(\alpha_2)\Gamma(\beta_2)(1 - \tau_2)^{\alpha_2 + \beta_2 - 1}} dr + \\ (1 - p_i - q)\mathbb{I}_1(k) \quad \text{for } 0 < k \leq 1 \end{aligned} \tag{6}$$

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

In order to solve the PPS with an exact approach, we need to transform the chance constraint in (2) to its deterministic counterpart. Thus, we have to derive the quantile function (i.e. inverse cumulative distribution function) of the probabilistic part $(\sum_{i \in N} b_i R_i x_i)$ in (2) for the selected projects, i.e. for i s.t. $x_i = 1$ in a solution. Even, the c.d.f of a single random variable R_i in (6) has no closed-form expression. Therefore, the derivation of the quantile function of $(\sum_{i \in N} b_i R_i x_i)$ is challenging.

Property 1 states that Normal approximation can be applied to the standardized sum of expenditures

$$H_n^x = \frac{\sum_{i\in N}[b_i R_i - \mathbb{E}(b_i R_i)]x_i}{\sqrt{\sum_{i\in N} \mathrm{Var}(b_i R_i)x_i^2}}.$$

Property 1. Let $b_i R_i \forall i \in N$ be independent non-identically distributed random variables. Then, $H_n^x = \frac{\sum_{i\in N}[b_i R_i - \mathbb{E}(b_i R_i)]x_i}{\sqrt{\sum_{i\in N} \mathrm{Var}(b_i R_i)x_i^2}}$ obeys the central limit theorem and its distribution approximates the standart normal distribution, if a solution to the PPS includes a significant number of selected projects (i.e. $x_i = 1$).

Proof. $\hat{b}_i = b_i R_i$ is a bounded random variable, i.e. $0 < \hat{b}_i \leq b_{\max}$ where $b_{\max}$ is the upper limit on the estimated budget b_i. In each grant program there is usually a $b_{\max}$ value imposed by the research council. Given a feasible solution to PPS, let I be the set of indices i such that $x_i = 1 \ \ \forall i \in I$ and $x_i = 0 \ \ \forall i \in N \setminus I$, then

$$H_n^x = \frac{\sum_{i\in I}[\hat{b}_i - \mathbb{E}(\hat{b}_i)]}{\sqrt{\sum_{i\in I} \mathrm{Var}(\hat{b}_i)}}$$

All $\hat{b}_i$ are uniformly bounded ($0 \leq \hat{b}_i \leq b_{\max}$) and they are independent random variables. Furthermore, $\sqrt{\sum_{i\in I} \mathrm{Var}(\hat{b}_i)} \to \infty$ as $|I| \to \infty$ since $\mathrm{Var}(\hat{b}_i)$ does not approach to zero as $i \to \infty$. Then, due to Lindeberg theorem, H_n^x approximates standart normal distribution [25, p.254]. □

We can transform the PPS to its deterministic equivalent part by Normal approximation. However, as shown in Property 1, we need the mean and variance of random variable $\hat{b}_i$ to approximate the true distribution of H_n^x. By using the properties of expectation and variance, we obtain: $\mathbb{E}(\hat{b}_i) = b_i \mathbb{E}(R_i)$ and $\mathrm{Var}(\hat{b}_i) = b_i^2 \mathrm{Var}(R_i)$. We need $\mathbb{E}(R_i)$ and $\mathrm{Var}(R_i)$ to apply Normal approximation. We give the mean, variance of R_i in Property 4 in Appendix A.3.

2.2. **Deterministic Equivalent Formulation for PPS.** In Property 1, we show that the total random budget spending obeys the central limit theorem. Now, we formulate the deterministic equivalent of the PPS by using Normal approximation and second order conic inequalities.

Constraint set (2) can be expressed as follows:

$$P\left(\sum_{i\in N} b_i R_i x_i \leq B\right) \Rightarrow$$
$$P\left(\frac{\sum_{i\in N}[b_i R_i - b_i \mathbb{E}(R_i)]x_i}{\sqrt{\sum_{i\in N} b_i^2 \mathrm{Var}(R_i) x_i^2}} \leq \frac{B - \sum_{i\in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i\in N} b_i^2 \mathrm{Var}(R_i) x_i^2}}\right) \quad (7)$$

Therefore, we can obtain the following probabilistic constraint:

$$P\left(Z \leq \frac{B - \sum_{i\in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i\in N} b_i^2 \mathrm{Var}(R_i) x_i^2}}\right) \geq \theta \Rightarrow$$
$$\Phi\left(\frac{B - \sum_{i\in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i\in N} b_i^2 \mathrm{Var}(R_i) x_i^2}}\right) \geq \theta \Rightarrow$$
$$\frac{B - \sum_{i\in N} b_i \mathbb{E}(R_i) x_i}{\sqrt{\sum_{i\in N} b_i^2 \mathrm{Var}(R_i) x_i^2}} \geq \Phi^{-1}(\theta) \quad (8)$$

where Z is the standard normal random variable and $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are its c.d.f and quantile function, respectively. As already known, (8) can be represented via second order conic inequalities. We reorganize (8) as follows:

$$\sum_{i\in N} b_i \mathbb{E}(R_i) x_i + \Phi^{-1}(\theta) \sqrt{\sum_{i\in N} b_i^2 \mathrm{Var}(R_i) x_i^2} \leq B \quad (9)$$

We assume a high probability level (i.e. $\theta \geq 0.90$), then $\Phi^{-1}(\theta) > 0$, which makes the constraint set (9) convex and it can be reformulated by second-order conic inequalities. The resulting deterministic equivalent reformulation of the PPS model is

a mixed-integer second-order cone program:

$$\begin{aligned}
\text{(MISOCP)}\ \max & \sum_{i\in N} s_i(1-p_i)x_i \\
\text{s.t}\ & \eta = \frac{B}{\Phi^{-1}(\theta)} - \frac{\sum_{i\in N} b_i \mathbb{E}(R_i)x_i}{\Phi^{-1}(\theta)} && (10)\\
& \sum_{i\in N} b_i^2 \mathrm{Var}(R_i)x_i^2 \leq \eta^2 && (11)\\
& \eta \geq 0 && (12)\\
& \text{Constraint sets (3) and (4)}
\end{aligned}$$

The conic reformulation of the constraint (9) is derived in constraint sets (10) and (11). η in equation (10) is an auxiliary variable for the linear portion of constraint (9). Constraint (11) is a second-order cone that represents constraint (9).

2.3. **Quality of Normal Approximation.** Research councils usually have strict budget constraints on R&D grant programs as they rely on public financial resources. Therefore, while modeling project expenditure uncertainty in a budget constraint in PPS, the convergence quality of the true expenditure distribution to Normal distribution is a significant concern for decision makers. In PPS model, θ provides the probability level and conversely $1-\theta$ specify the risk level. Decision makers prefer low risk levels. 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 PPS with risk level of $1-\theta$, what can we say about the real risk of the solution? Is it greater than $1-\theta$ or less than $1-\theta$? 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 which will be presented in

Section 3. In the computer science literature, Aljuaid and Yanikomeroglu [26] 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.

Esseen [27] give maximum error of Normal approximation in terms of probability levels (Theorem 1 in Appendix A.4). 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 assists us to quantify the worst case error when the Normal approximation is employed in PPS's. For our case, we need to derive moments of random variable R_i to apply Berry–Esseen theorem. In Property 2, we derive Berry-Esseen bound for any solution obtained by the MISOCP.

Property 2. Let $b_i R_i \forall i \in N$ be i.ni.d. random variables as defined in equation (2) and G_n^x be the true c.d.f of $H_n^x = \frac{\sum_{i\in N}[b_i R_i - b_i \mathbb{E}(R_i)]x_i}{\sqrt{\sum_{i\in N} b_i^2 \mathrm{Var}(R_i) x_i^2}}$ for a specific solution vector x. Then Kolmogorov distance between G_n^x and Φ^x (Normal approximation for the solution vector x) satisfies:

$$D_{\mathrm{Kol}}^x \leq C \left(\sum_{i=1}^{i=n} \left| b_i^3 \mathbb{E}(R_i^3) - 3 b_i^3 \mathbb{E}(R_i^2)\mathbb{E}(R_i) + 2 b_i^3 [\mathbb{E}(R_i)]^3 \right| x_i \right) \cdot \left(\sum_{i=1}^{i=n} b_i^2 \mathrm{Var}(R_i) x_i \right)^{-3/2} \tag{13}$$

where $D_{\mathrm{Kol}}^x = \sup_{z\in\mathbb{R}} |G_n^x(z) - \Phi^x(z)|$ and C is a constant.

Proof. See Appendix A.1. □

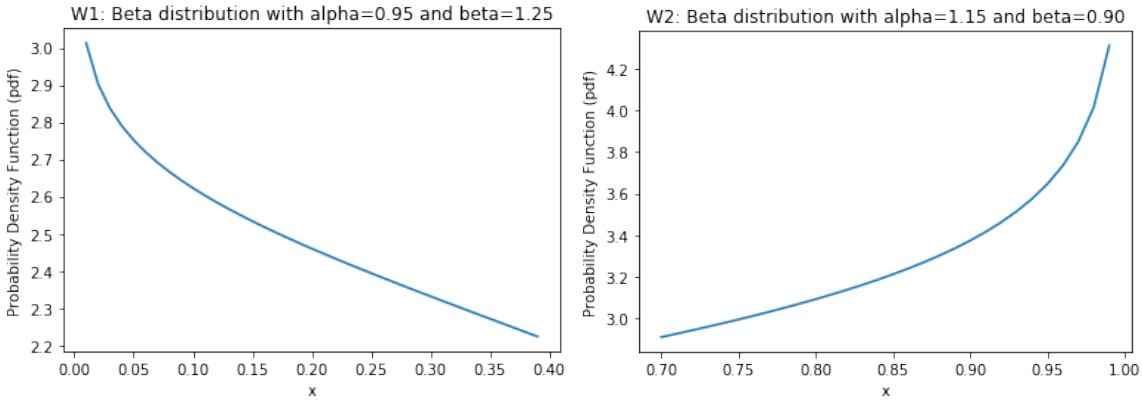
We can calculate a bound value for the maximum error of normal approximation by using the best estimate of C. Shevtsova [28] recently showed that the best estimate of C is 0.56. Berry-Esseen theorem gives the worst case absolute difference for the probability level. Therefore, we also perform a simulation study to evaluate the empirical probability level of our PPS. In the next section, we give numerical results.

3. Numerical Results

In this section, we report the results of numerical experiments for the proposed project portfolio selection approach. In our experiments, we solved randomly generated problem instances to see how modeling expenditure uncertainty in project portfolio selection problems improves budget utilization, how computational performance of MISOCP is affected by different experimental factors and how Normal approximation performs in representing the chance constraint in PPS.

As said before, we have used truncated beta distributions to model uncertainty in expenditure ratios W_1 and W_2. We obtained paramater values for the distributions of W_1 and W_2 by examining the historical data of a call, which is not publicly available. We have also interacted with an expert in the area. Kolmogorov-Smirnov and Anderson Darling goodness of fit tests on the data showed that truncated beta distributions can be used to model expenditures. We used EasyFit software for fitting. Range parameters τ_1 and τ_2 of random variables W_1 and W_2 are set to 0.4 and 0.7, respectively. We obtained the distributions given in Figure 2.

We define problem size as the number of projects applying to a call. We use four problem sizes ($|N|$ =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 given in Table 1 by considering practical cases. The first two factors are p_i and q, the



(A) Canceled Projects (Beta(0.95, 1.25, 0, 0.4)) (B) Successful Projects (Beta(1.15, 0.9, 0.7, 1))

FIGURE 2. Distributions of Expenditure Ratio

TABLE 1. Factor Values for each problem size

	Levels	
Factor	Low	High
p_i	U(0.01-0.1)	U(0.01-0.2)
q	0.4	0.5
bf	0.1	0.2
θ	0.90	0.95

cancellation probability of a project and budgetary slack probability for successful projects. For each project p_i is generated according to uniform distribution (U) between given bounds. The budget fraction (bf) is the ratio of available budget (B) over sum of all project budgets, which is set to 0.1 and 0.2. In our experiments, we set total available budget as $B = (\sum_{i \in N} b_i) \times bf$. The last factor is θ which controls the budget overrun risk for a selected project portfolio.

We assume that there are seven scientific areas in our problem instances. We examined historical data of projects in one of the programs in TUBITAK and calculated the share of each area in project applications (Table 2). Note that the budget of a project in a scientific area is randomly generated according to different uniform distributions between given lower (U_{lb}) and upper bounds (U_{ub}). We set the minimum

acceptance rate for each area 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].

TABLE 2. Number of project applications and budgets for different scientific areas

Scientific Area (j)	Ratio of Applications	Area Budget U_{lb}	U_{ub}
Environment, Atmosphere, Earth and Marine Sciences	0.1	210,000	360,000
Electrical, Electronics and Informatics	0.1	90,000	360,000
Engineering Sciences	0.15	70,000	360,000
Health Sciences	0.1	230,000	360,000
Social Sciences and Humanities	0.15	70,000	220,000
Basic Sciences	0.25	110,000	360,000
Agriculture, Forestry and Veterinary	0.15	140,000	360,000

We solve five random instances for each factor combination and problem size. Hence, we have $4x2^4x5=320$ runs. For each solution obtained by solving MISOCP, we conducted a simulation with 10,000 replications for budget expenditures. We aimed to observe the probability of budget overrun and the budget utilization rates under selected budget expenditure distributions.

We solve MISOCP models 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. 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).

First, we present how modeling budget expenditure uncertainty helps research councils to support more projects and improve budget utilization.

3.1. **What does modeling of uncertainty offer to decision makers?** In this section, we quantify the value of modeling uncertainty in a public R&D project portfolio selection decisions. The baseline scenario is solving the PPS with a deterministic

objective function ($\sum_{i\in N} s_i x_i$) and under a deterministic budget constraint, i.e. replacing constraint (2) with the deterministic budget constraint $\sum_{i\in N} b_i x_i \leq B$. We solved randomly generated problem instances with $|N| = 2000$ using MISOCP and deterministic PPS. For the solutions (i.e. selected project portfolios) achieved by both models, we calculated expected total score ($E(score)$), expected number of successfully completed projects ($E(nscp)$), and the budget utilization with probability θ. $E(nscp)$ 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. Γ follows Poisson-Binomial distribution (Hong [29]). Then, the expected number of canceled projects for a solution vector x is: $\mathbb{E}(\Gamma) = \sum_{i\in N} \mathbb{E}(I_i)x_i = \sum_{i\in N} p_i x_i$. Therefore, $E(nscp) = ns - E(\Gamma)$, where ns is the number of supported projects. Then, we calculated how much in percentage each performance measure is improved.

TABLE 3. % Improvements Achieved by PPS over Baseline Scenario

		E(score)			E($nscp$)			Budget Utilization		
Factor	Levels	Min	Avg	Max	Min	Avg	Max	Min	Avg	Max
bf	0.1	6.5	9.3	12.5	8.8	10.5	14.3	8.0	11.1	14.9
	0.2	6.7	9.2	11.9	6.5	10.2	13.1	8.7	11.8	15.2
p_i	U(0.01-0.1)	6.5	7.6	8.7	6.5	8.7	12.1	8.0	9.6	11.0
	U(0.01-0.2)	9.4	10.9	12.5	10.8	12.1	14.3	11.1	13.3	15.2
q	0.4	6.5	8.6	11.1	6.5	9.8	12.8	8.0	10.6	13.5
	0.5	7.8	9.9	12.5	8.2	10.9	14.3	9.5	12.2	15.2
θ	0.9	7.0	9.4	12.5	6.5	10.6	14.3	8.5	11.7	15.2
	0.95	6.5	9.0	11.8	6.5	10.1	14.0	8.0	11.2	14.7

We present the results in Table 3. Modeling uncertainty in budget spending and solving PPS has improved $E(score)$ by 9.25% which means, under the same call

budget, a research council can obtain higher scientific output. Similarly, $E(nscp)$ increases by 10.35%, which is because PPS selects more projects by decreasing budgetary slack, and selects the projects with higher success probability. Finally, budget utilization increases by 11.45% as PPS considers possible budget underspending of each project to create money for more projects.

When we have a tighter budget (small bf), PPS achieves higher improvement in $E(score)$ and $E(nscp)$, i.e. modeling uncertainty is more critical if we have less budget in a call. If cancellation probabilities (p_i) are higher or budget underspending probability for successful projects (q) is higher, then budgetary slack is more likely to occur and we see that PPS achieves significantly more improvement in expected total score, expected number of successful projects and budget utilization measures. Finally, we observe that as the decision maker takes a higher risk of overrun, i.e. smaller θ, then PPS achieves higher improvements in expected total score, expected number of successful projects and budget utilization. In the next section, we explore the computational performance of MISOCP model.

3.2. **Computational performance of MISOCP for practical size instances.** In Table 4, we give the results for medium ($|N|$ =250 or 500) and large size ($|N|$ =1000 or 2000) problem instances. We report 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 (%)). We calculate EmpCL and EmpBUR values in simulation runs. Each row presents averages of 2^4x5 = 80 instances. Note that we measure the mean CPU time for the instances solved to optimum in the given time limit.

TABLE 4. Computational Results

Size	Factor	Levels	opt (%)	gap (%)	CPU (s) mean	EmpCL	BEB	EmpBUR(%)
Medium ($\|N\|$ = 250 or 500)	$\|N\|$	250	100	-	419.9	0.9759	0.2970	93.9
		500	74	0.25	322.2	0.9774	0.2010	95.5
	bf	0.1	86	0.33	541.5	0.9763	0.2915	94.0
		0.2	88	0.17	217.6	0.9770	0.2065	95.4
	p_i	U(0.01-0.1)	75	0.26	443.4	0.9791	0.2987	95.3
		U(0.01-0.2)	99	0.06	329.0	0.9743	0.1993	94.1
	q	0.4	84	0.30	404.7	0.9770	0.2489	94.7
		0.5	90	0.19	353.9	0.9764	0.2491	94.7
	θ	0.90	91	0.19	362.1	0.9638	0.2471	95.2
		0.95	83	0.29	396.4	0.9895	0.2509	94.2
Large ($\|N\|$ = 1000 or 2000)	$\|N\|$	1000	63	0.11	909.8	0.9825	0.1400	96.5
		2000	60	0.04	684.9	0.9892	0.0991	97.3
	bf	0.1	60	0.10	953.0	0.9811	0.2100	95.4
		0.2	64	0.04	653.0	0.9839	0.1474	96.4
	p_i	U(0.01-0.1)	64	0.09	779.3	0.9841	0.2147	96.4
		U(0.01-0.2)	60	0.05	818.8	0.9809	0.1427	95.5
	q	0.4	63	0.07	993.8	0.9828	0.1785	95.9
		0.5	61	0.08	599.1	0.9822	0.1788	95.9
	θ	0.90	64	0.07	639.3	0.9730	0.1775	96.3
		0.95	60	0.07	967.6	0.9920	0.1799	95.5

As $|N|$ increases, opt(%) and CPU time performance of the model degrades, but gap(%) values get better. It becomes harder to solve instances to optimum (see Figure 3), but the optimality gap for feasible solutions gets smaller. We achieved the highest average gap, 0.25% when $|N| = 500$. The non-optimal solutions achieved by the model are all near optimal which is remarkable and practical for a large-scale nonlinear model. As $|N|$ gets larger, the budget utilization rate (EmpBUR) approaches to 97% (also shown in Figure 3). This is due to the aggregate random expenditure effect of increased number of supported projects. This observation also shows the

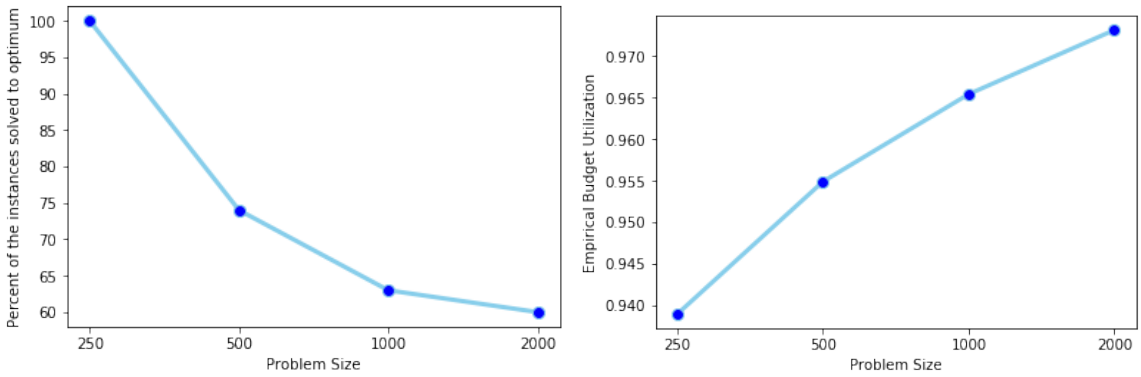


FIGURE 3. Hardness of the instances and budget utilization rate

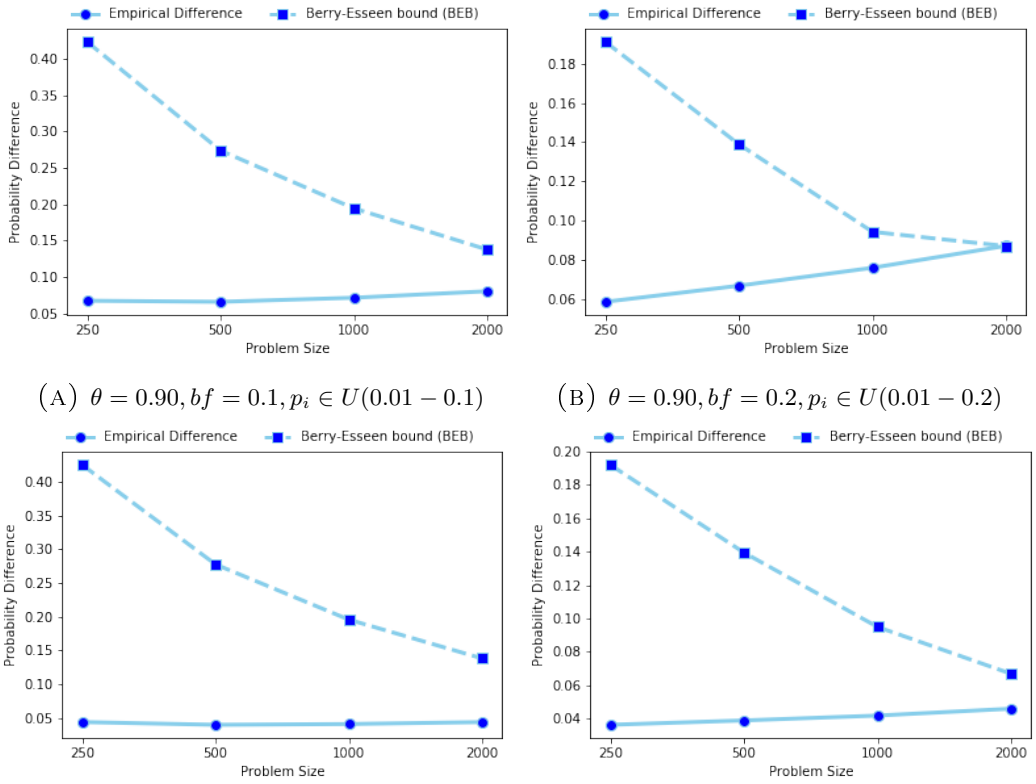
importance of the proposed model for large-scale project portfolio selection problems under expenditure uncertainty. In other words, if the expenditure uncertainty is modeled by a PPS along with an appropriate distribution, the budget utilization clearly improves when the number of supported projects in the portfolio gets larger. 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 Property 1).

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 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 (p_i) has a significant effect on BEB. The reason for this behavior is that as p_i values increase, the variance of the true unknown distribution increases and BEB decreases due to equation (13). 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 (EmpCL) as expected. When we solve instances with $\theta = 0.9$, the average empirical probability level is close to 0.97. When we solve instances with $\theta = 0.95$, the average empirical probability level is close to 0.99. High empirical probability levels are promising from a decision maker‘s perspective.

To get more insights about the approximation error in terms of probability levels, 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 p_i 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 the Normal distribution gives a good approximation. Note that empirical probability differences are less than or equal to 0.08 for $\theta = 0.9$ and they are less than or equal to 0.04 for $\theta = 0.95$. 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 p_i 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



(A) $\theta = 0.90, bf = 0.1, p_i \in U(0.01 - 0.1)$

(B) $\theta = 0.90, bf = 0.2, p_i \in U(0.01 - 0.2)$

(C) $\theta = 0.95, bf = 0.1, p_i \in U(0.01 - 0.1)$

(D) $\theta = 0.95, bf = 0.2, p_i \in U(0.01 - 0.2)$

FIGURE 4. Berry-Esseen bound and Empirical Difference

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, managerial insights for government research organizations are discussed.

3.3. **Managerial Insights for Governmental Research Organizations.** The proposed approach to model expenditure uncertainty in R&D projects can help research councils or governmental research organizations (in a broader context) to reduce unspent budgets at the aggregate level. The simulation study suggests that 100% budget utilization rate is hard to achieve, but 94-97% utilization rates are within reach if we rigorously model the budget uncertainty. This finding is an expected one since research activities are inherently uncertain in terms of outcomes and expenses. It is critical to understand that there is no exact true distribution model for any uncertain (stochastic) input and all the mathematical models are just approximation of the reality. On the other hand, if the budget uncertainty is not taken into account in the project selection model, the utilization rate could even drop to 81.6% in the worst case, which is relatively very low. In other words, if the expected utilization rate is 94% for a portfolio when we apply the proposed approach, then it could decrease by 15.2% (at the maximum) (refer to Table 3), and become 81.6% when we do not employ the proposed uncertainty modeling. Therefore, modeling the budget uncertainty rigorously with an analytical model pays off.

The best resource for the uncertainty modeling is the data of completed projects in various grant programs. Governmental research organizations should analyze those large data sets for the input modeling of expenditure uncertainties. Availability of commercial input modeling packages such as EasyFit help practitioners identify possible probability distributions. Expert opinions could also shed light upon identifying distributional characteristics. It is authors' view that most of the time expenditure behaviors could be categorized into different spending patterns, which entail the utilization of mixture distributions. For middle and large-size project portfolios in terms of possible number supported projects, Normal approximation could be employed as

a tractable model to solve the proposed chance-constrained optimization model along with the theoretical bound and simulation studies.

The decision makers of the governmental organizations are sensitive about spending grant programs' budgets. The being underbudget or overbudget implies non-positive messages for the upcoming budget requirements. Therefore, the public decision makers want to minimize the difference between the expenditure and the appropriation. In this regard, the approximation quality of the Normal distribution is important. Simulation studies along with the theoretical bound for the worst case behavior shed light upon practical probability of being within the budget. In practice, the decision makers love to work with high probabilities such as 90% and 95%. Our simulation studies empirically demonstrate that the optimal solutions of our proposed model eminently satisfy the desired probability level (i.e. 90% or 95%) of the chance constraint. Besides, the approximation errors are less than 4% for the 95% level and less than 8% for the 90% level. Although those approximation errors could change according to the various mixture distributions and different problem data, we believe that this observation also provides a practically highly relevant insight for the public decision makers.

4. Conclusions

In this study, for research councils we identify the unspent budget at the aggregate level as a practically important problem since the money that is not used could have been utilized to grant other promising projects. Therefore, we propose a new method to help research councils to enhance budget utilization. First, we model the expenditure uncertainty with a mixture distribution.

Second, we have shown that Normal distribution can be used to approximate the proposed model. This helps to reformulate PPS using conic constraints (MISOCP).

Our computational tests showed that even for 2000 projects MISOCP model can be solved to optimum (or near optimum) by commercial solvers such as IBM CPLEX in a reasonable amount of time.

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.

The main take away for the decision makers in a research council is that when we solve our chance constrained model (PPS) with the Normal approximation, the optimal solutions of our model eminently satisfy the desired probability level of the chance constraint (albeit with mild conservatism).

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 or other governmental research organizations at large to improve the budget utilization to support more projects. In 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

[1] A. D. Henriksen and A. J. Traynor. A practical R&D project-selection scoring tool. *IEEE Transactions on Engineering Management*, 46(2):158–170, 1999.

[2] N. G. Hall, D. Z. Long, J. Qi, and M. Sim. Managing underperformance risk in project portfolio selection. *Operations Research*, 63(3):660–675, 2015.

[3] T. C. L. Albano, E. C. Baptista, F. Armellini, D. Jugend, and E. M. Soler. Proposal and solution of a mixed-integer nonlinear optimization model that incorporates future preparedness for project portfolio selection. *IEEE Transactions on Engineering Management*, pages 1–13, 2019.

[4] R. Antle and G. D. Eppen. Capital rationing and organizational slack in capital budgeting. *Management Science*, 31(2):163–174, 1985.

[5] A. Dunk and H. Nouri. Antecedents of budgetary slack: A literature review and synthesis. *Journal of Accounting Literature*, 17:72–96, 1998.

[6] Q. Hu and J. Szmerekovsky. Project portfolio selection: A newsvendor approach. *Decision Sciences*, 48(1):176–199, 2017.

[7] TUBITAK. Activity reports. 2018. URL `http://www.tubitak.gov.tr/tr/icerik-faaliyet-raporlari`.

[8] NSF. National Science Foundation, Grant Policy Manual. 2005.

[9] NIH. National Institutes of Health, Grant Policy Statements. 2013. URL `grants.nih.gov/grants/policy/policy.htm`.

[10] G. H. A. Pereira, Denise A. B., and Mônica C. S. The truncated inflated beta distribution. *Communications in Statistics-Theory and Methods*, 41(5):907–919, 2012.

[11] L. A. Meade and A. Presley. R&D project selection using the analytic network process. *IEEE Transactions on Engineering Management*, 49(1):59–66, 2002.

[12] C. F. Chien. A portfolio-evaluation framework for selecting r&d projects. *R&D Management*, 32(4):359–368, 2002.

[13] G. A. B. Marcondes, R. C. Leme, and M. M. Carvalho. Framework for integrated project portfolio selection and adjustment. *IEEE Transactions on Engineering Management*, 66(4):677–688, 2019.

[14] H. Kellerer, U. Pferschy, and D. Pisinger. Knapsack problems. *Springer*, 2004.

[15] M. W. Dickinson, A. C. Thornton, and S. Graves. Technology portfolio management: optimizing interdependent projects over multiple time periods. *IEEE Transactions on Engineering Management*, 48(4):518–527, 2001.

[16] A. L. Medaglia, S. B. Graves, and J. L. Ringuest. A multiobjective evolutionary approach for linearly constrained project selection under uncertainty. *European Journal of Operational Research*, 179(3):869–894, 2007.

[17] A. Koç, D. P. Morton, E. Popova, S. M. Hess, E. Kee, and D. Richards. Prioritizing project selection. *Engineering Economist*, 54(4):267–297, 2009.

[18] S. Solak, J-P. B. Clarke, E. L. Johnson, and E. R. Barnes. Optimization of R&D project portfolios under endogenous uncertainty. *European Journal of Operational Research*, 207(1):420–433, 2010.

[19] M. Çağlar and S. Gürel. Public R&D project portfolio selection problem with cancellations. *OR Spectrum*, 39(3):659–687, 2017. URL `http://dx.doi.org/10.1007/s00291-016-0468-5`.

[20] S. Kavadias and R. O. Chao. *Resource allocation and new product development portfolio management*. Elsevier, Oxford, 2007.

[21] N. M. Arratia, F. I Lopez, S.E. Schaeffer, and L. Cruz-Reyes. Static R&D project portfolio selection in public organizations. *Decision Support Systems*, 84:53–63, 2016.

[22] M. Çağlar and S. Gürel. Impact assessment based sectoral balancing in public R&D project portfolio selection. *Socio-Economic Planning Sciences*, 66:68–81, 2019. URL `https://doi.org/10.1016/j.seps.2018.07.001`.

[23] S.A. Gabriel, S. Kumar, J. Ordonez, and A. Nasserian. A multiobjective optimization model for project selection with probabilistic considerations. *Socio-Economic Planning Sciences*, 40(4):297–313, 2006.

[24] N. G. Hall. Research and teaching opportunities in project management. *INFORMS TutORials in Operations Research*, pages 329–388, 2016.

[25] W. Feller. *An introduction to probability theory and its applications*, volume 2. Wiley, New York, 2 edition, 1971.

[26] M. Aljuaid and H. Yanikomeroglu. Investigating the Gaussian convergence of the distribution of the aggregate interference power in large wireless networks. *IEEE Transactions on Vehicular Technology*, 59(9):4418–4424, 2010.

[27] C. G. Esseen. A moment inequality with an application to the central limit theorem. *Skand. Aktuarietidskr.*, 39:160–170, 1956.

[28] I. G. Shevtsova. An improvement of convergence rate estimates in the lyapunov theorem. *Doklady Mathematics*, 82:862–864, 2010.

[29] Y. Hong. On computing the distribution function for the Poisson binomial distribution. *Computational Statistics and Data Analysis*, 59:41–51, 2013.

[30] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. New York: Wiley, 2nd edition, 1995.

Appendix A. Supplementary Material

A.1. **Proof of Property 2.**

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

$$\mathbb{E}(Y_i) = 0 \tag{14}$$

$$\begin{aligned}\mathbb{E}(Y_i^2) &= \mathbb{E}\left([\hat{b}_i - \mathbb{E}(\hat{b}_i)]^2\right) = \mathbb{E}\left[\hat{b}_i^2 - 2\hat{b}_i\mathbb{E}(\hat{b}_i) + (\mathbb{E}(\hat{b}_i))^2\right] = \mathbb{E}(\hat{b}_i^2) - [\mathbb{E}(\hat{b}_i)]^2 \\ &= \mathrm{Var}(\hat{b}_i) = b_i^2\mathrm{Var}(R_i)\end{aligned} \tag{15}$$

$$\begin{aligned}\mathbb{E}(|Y_i^3|) &= \mathbb{E}\left(\left|[\hat{b}_i - \mathbb{E}(\hat{b}_i)]^3\right|\right) = \mathbb{E}\left[\left|\hat{b}_i^3 - 3\hat{b}_i^2\mathbb{E}(\hat{b}_i) + 3\hat{b}_i[\mathbb{E}(\hat{b}_i)]^2 - [\mathbb{E}(\hat{b}_i)]^3\right|\right] \\ &= \left|\mathbb{E}(\hat{b}_i^3) - 3\mathbb{E}(\hat{b}_i^2)\mathbb{E}(\hat{b}_i) + 2[\mathbb{E}(\hat{b}_i)]^3\right| = \left|b_i^3\mathbb{E}(R_i^3) - 3b_i^3\mathbb{E}(R_i^2)\mathbb{E}(R_i) + 2b_i^3[\mathbb{E}(R_i)]^3\right|\end{aligned} \tag{16}$$

By using equation (24) and the property of gamma function such that $\Gamma(t+1) = t\Gamma(t)$, we derive:

$$\begin{aligned}\mathbb{E}(R_i^3) &= p_i\left(\tau_1^3\frac{\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_1+3)}{\Gamma(\alpha_1)\Gamma(\alpha_1+\beta_1+3)}\right) + (1-p_i-q) \\ &\quad + q\left(\sum_{k=0}^{k=3}\frac{3!}{k!(3-k)!}\tau_2^k(1-\tau_2)^{3-k}\frac{\Gamma(\alpha_2+\beta_2)\Gamma(\alpha_2+3-k)}{\Gamma(\alpha_2)\Gamma(\alpha_2+\beta_2+3-k)}\right) \qquad (17) \\ &= p_i\left(\tau_1^3\frac{(\alpha_1+2)(\alpha_1+1)\alpha_1}{(\alpha_1+\beta_1+2)(\alpha_1+\beta_1+1)(\alpha_1+\beta_1)}\right) + (1-p_i-q) \\ &\quad + q(1-\tau_2)^3\frac{(\alpha_2+2)(\alpha_2+1)\alpha_2}{(\alpha_2+\beta_2+2)(\alpha_2+\beta_2+1)(\alpha_2+\beta_2)} \\ &\quad + 3q\tau_2(1-\tau_2)^2\frac{(\alpha_2+1)\alpha_2}{(\alpha_2+\beta_2+1)(\alpha_2+\beta_2)} \\ &\quad + 3q\tau_2^2(1-\tau_2)\frac{\alpha_2}{(\alpha_2+\beta_2)} + q\tau_2^3 \qquad (18)\end{aligned}$$

By using equations (26), (27) and Remark 1, we can obtain inequality (13). □

A.2. **Property 3 and its proof.**

Property 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:

$$\mathbb{E}(W) = \frac{\alpha b + \beta a}{\alpha + \beta} \tag{19}$$

$$Var(W) = \frac{(b-a)^2 \alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{20}$$

$$\mathbb{E}(W^n) = \begin{cases} \sum_{k=0}^{k=n} \frac{n!}{k!(n-k)!} a^k (b-a)^{n-k} \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+n-k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n-k)} & \text{if } \ 0 < a < b \\ b^n \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+n)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n)} & \text{if } \ a = 0 \text{ and } b > 0 \end{cases} \tag{21}$$

Proof. 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. [30], 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:

$$\begin{aligned} \mathbb{E}(W) &= \mathbb{E}(a + (b-a)T) = a + (b-a)\mathbb{E}(T) = a + \frac{(b-a)\alpha}{\alpha+\beta} = \frac{a\alpha + a\beta + b\alpha - a\alpha}{\alpha+\beta} \\ &= \frac{\alpha b + \beta a}{\alpha+\beta} \end{aligned}$$

$$\begin{aligned} Var(W) &= Var(a + (b-a)T) = Var((b-a)T) = (b-a)^2 Var(T) \\ &= \frac{(b-a)^2\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

for $\ 0 < a < b$

$$\mathbb{E}(W^n) = \mathbb{E}((a+(b-a)T)^n) = \mathbb{E}\left(\sum_{k=0}^{k=n}\binom{n}{k}a^k(b-a)^{n-k}T^{n-k}\right)$$
$$= \sum_{k=0}^{k=n}\frac{n!}{k!(n-k)!}a^k(b-a)^{n-k}\mathbb{E}(T^{n-k})$$
$$= \sum_{k=0}^{k=n}\frac{n!}{k!(n-k)!}a^k(b-a)^{n-k}\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+n-k)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n-k)}$$

for $a = 0$ and $b > 0$

$$\mathbb{E}(W^n) = \mathbb{E}((a+(b-a)T)^n) = \mathbb{E}((bT)^n) = b^n\mathbb{E}(T^n) = b^n\frac{\Gamma(\alpha+\beta)\Gamma(\alpha+n)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n)}$$

□

Note that the n^{th} moment is derived to apply the Berry-Esseen theorem.

A.3. **Derivation of statistical measures for R_i.**

Property 4. The mean, variance and n^{th} moment of the random variable R_i are given as follows:

$$\mathbb{E}(R_i) = p_i\frac{\alpha_1\tau_1}{\alpha_1+\beta_1} + q\frac{\alpha_2+\beta_2\tau_2}{\alpha_2+\beta_2} + (1-p_i-q) \tag{22}$$

$$\begin{aligned} Var(R_i) = \; & p_i\left[\left(\frac{\alpha_1\tau_1}{\alpha_1+\beta_1}\right)^2 + \frac{\tau_1^2\alpha_1\beta_1}{(\alpha_1+\beta_1)^2(\alpha_1+\beta_1+1)}\right] \\ & + q\left[\left(\frac{\alpha_2+\beta_2\tau_2}{\alpha_2+\beta_2}\right)^2 + \frac{(1-\tau_2)^2\alpha_2\beta_2}{(\alpha_2+\beta_2)^2(\alpha_2+\beta_2+1)}\right] \\ & + (1-p_i-q) \\ & - \left[p_i\frac{\alpha_1\tau_1}{\alpha_1+\beta_1} + q\frac{\alpha_2+\beta_2\tau_2}{\alpha_2+\beta_2} + (1-p_i-q)\right]^2 \end{aligned} \tag{23}$$

$$\mathbb{E}(R_i^n) = p_i \left(\tau_1^n \frac{\Gamma(\alpha_1+\beta_1)\Gamma(\alpha_1+n)}{\Gamma(\alpha_1)\Gamma(\alpha_1+\beta_1+n)} \right) + (1-p_i-q) + q \left(\sum_{k=0}^{k=n} \frac{n!}{k!(n-k)!} \tau_2^k (1-\tau_2)^{n-k} \cdot \frac{\Gamma(\alpha_2+\beta_2)\Gamma(\alpha_2+n-k)}{\Gamma(\alpha_2)\Gamma(\alpha_2+\beta_2+n-k)} \right) \quad (24)$$

Proof. $\mathbb{E}(R_i) = \sum_{j=1}^{j=3} w_j m_j^1$ where m_1^1 is the mean (first moment) of the truncated distribution in open interval $(0, \tau_1)$, m_2^1 is the mean of the truncated distribution in open interval $(\tau_2, 1)$, m_3^1 is the mean of the degenerate distribution at point 1, by using equation in (19), we can obtain first moments (means) so that $m_1^1 = \frac{\alpha_1 \tau_1}{\alpha_1+\beta_1}$ and $m_2^1 = \frac{\alpha_2+\beta_2\tau_2}{\alpha_2+\beta_2}$, since mean of the constant value is itself, then $m_3^1 = 1$. Apparently, $w_1 = p_i$, $w_2 = q$, and $w_3 = (1-p_i-q)$. Therefore, we obtain: $\mathbb{E}(R_i) = p_i \frac{\alpha_1\tau_1}{\alpha_1+\beta_1} + q\frac{\alpha_2+\beta_2\tau_2}{\alpha_2+\beta_2} + (1-p_i-q)$. We know $Var(R_i) = \mathbb{E}(R_i^2) - [\mathbb{E}(R_i)]^2$.

We can write $\mathbb{E}(R_i^2) = \sum_{j=1}^{j=3} w_j m_j^2$ where m_1^2 is the second moment of the truncated distribution in open interval $(0, \tau_1)$, m_2^2 is the second moment of the truncated distribution in open interval $(\tau_2, 1)$, m_3^2 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:

$$\begin{aligned} Var(R_i) &= \mathbb{E}(R_i^2) - [\mathbb{E}(R_i)]^2 = p_i\mathbb{E}(W_1^2) + q\mathbb{E}(W_2^2) + (1-p_i-q) - [\mathbb{E}(R_i)]^2 \\ &= p_i \left([\mathbb{E}(W_1)]^2 + Var(W_1) \right) + q \left([\mathbb{E}(W_2)]^2 + Var(W_2) \right) + (1-p_i-q) - [\mathbb{E}(R_i)]^2 \end{aligned} \quad (25)$$

Hence, by using equations in (19), (20) and (22), we derive:

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

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

$$\begin{aligned} \mathbb{E}(R_i^n) = p_i & \left(\tau_1^n \frac{\Gamma(\alpha_1 + \beta_1)\Gamma(\alpha_1 + n)}{\Gamma(\alpha_1)\Gamma(\alpha_1 + \beta_1 + n)} \right) + (1 - p_i - q) \\ & + q \left(\sum_{k=0}^{k=n} \frac{n!}{k!(n-k)!} \tau_2^k (1 - \tau_2)^{n-k} \frac{\Gamma(\alpha_2 + \beta_2)\Gamma(\alpha_2 + n - k)}{\Gamma(\alpha_2)\Gamma(\alpha_2 + \beta_2 + n - k)} \right) \end{aligned}$$

□

A.4. **Theorem for the theoretical bound.**

Theorem 1. Berry–Esseen theorem for the maximum error of normal approximation: Let $Y_1, Y_2, ...Y_n$ be i.ni.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 $Q_n = \frac{\sum_{i=1}^{i=n} Y_i}{\sqrt{\sum_{i=1}^{i=n} \mathbb{E}(Y_i^2)}}$, G_n is the c.d.f of Q_n, Φ is the c.d.f of the standard normal distribution. Then, for the Kolmogorov distance is defined by

$$D_{\text{Kol}} = \sup_{z \in \mathbb{R}} |G_n(z) - \Phi(z)| . \tag{26}$$

there exists a constant C, such that $D_{\text{Kol}} \leq C\psi$ where

$$\psi = \left(\sum_{i=1}^{i=n} \mathbb{E}(|Y_i^3|) \right) \left(\sum_{i=1}^{i=n} \mathbb{E}(Y_i^2) \right)^{-3/2} . \tag{27}$$

Remark 1. Esseen [27] 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 [28] 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.