Multi-factor Dependence Modelling with Specified Marginals and Structured Association in Large-scale Project Risk Assessment

Byung-Cheol Kim[1]

[1] *Assistant Professor. Project and Supply Chain Management, Black School of Business, Penn State Erie, the Behrend College, 5101 Jordan Road, Erie, PA. 16563-1400.* buk70@psu.edu

Abstract

This paper examines the high-dimensional dependence modelling problem in the context of project risk assessment. As the dimension of uncertain performance units (i.e., itemized costs and activity times) in a project increases, specifying a feasible correlation matrix and eliciting relevant pair-wise information, either from historical data or with expert judgement, becomes practically unattainable or simply not economical. This paper presents a factor-driven dependence elicitation and modelling framework with scalability to large-scale project risks. The multi-factor association model (MFAM) accounts for hierarchical relationships of multiple association factors and provides a closed-form solution to a complete and mathematically consistent correlation matrix. Augmented with the structured association (SA) technique for systematic identification of hierarchical association factors, the MFAM offers additional flexibility of utilizing the minimum information available in standardized, ubiquitous project plans (e.g., work breakdown structure, resource allocation, or risk register), while preserving the computational efficiency and the scalability to high dimensional project risks. Numerical applications and simulation experiments show that the MFAM, further combined with extended analytics (i.e., parameter calibration and optimization), provides credible risk assessments (with accuracy comparable to full-scale simulation) and further enhances the realism of dealing with high-dimensional project risks utilizing all relevant information.

KEYWORDS: Project management, dependence modelling, large-scale risks, simulation.

1 Introduction

This paper tackles the problem of dependence modelling for large-scale project risk assessment. Dependence modelling constitutes an essential element of risk-adjusted project planning and predictive control, in particular for probabilistic cost estimates (GAO 2020; Garvey et al. 2016), stochastic network schedules (GAO 2015; Trietsch et al. 2012; van Dorp 2005), project-end outcome updates (Cho 2009; Kim 2015), and predictive performance tracking (Kim and Kwak 2018). Inter-dependence between project tasks is also one of the driving factors of project complexity along with the project size and the variety of tasks (Baccarini 1996; Tatikonda and Rosenthal 2000). Consequently, accounting for the nature and the degree of dependence is a demanding challenge for proper management of modern projects with increasing complexity and structural uncertainty (Mo et al. 2008; Williams 1999).

The need for quantitative risk assessment as a decision support tool has been well recognized since several seminal papers in capital investments (Hertz 1964) and operations research (Malcolm et al. 1959; Van Slyke 1963). In practice, however, projects often behave in a way that clashes with what is expected from the best practices and standards for successful completion in time (Love et al. 2013; Schonberger 1981) and within the budget (Flyvbjerg 2006; Love et al. 2013). Elementary statistics shows that project cost and time, as a risk-adjusted sum of random variables, tend to exceed the sum of their isolated marginal estimates when there exist positive inter-variable associations. Empirical data also suggest that (i) inter-variable correlations are commonly observed and (ii) ignoring correlation leads to systematic underestimation of the real risks (Chau 1995; Newton 1992; Skitmore and Ng 2002). Moreover, as the uncertainty dimension increases the percent underestimation of the total cost (or time) drastically increases (Garvey et al. 2016, p.322). Subsequently, a proper consideration of inter-variable dependence is widely emphasized as a crucial element of contingency settings and project risk assessment in general (GAO 2015, p.115; GAO 2020, p.155; NASA 2013, pp.33-37).

In theory, dependence modelling can be straightforward. In a narrow sense, a vector of dependent random variables can be specified as a mixture of univariate marginals ($\mathbf{X}$) and the corresponding correlation matrix ($\mathbf{\Sigma_X}$).

$$\mathbf{X}_\Sigma \sim \langle \mathbf{X}; \mathbf{\Sigma}_\mathbf{X} \rangle \tag{1}$$

The correlation-driven dependent vector ($\mathbf{X}_\Sigma$) in Eq. (1) provides a mathematically rigorous representation. However, specifying a feasible correlation matrix is a data-intensive process. In practice, the burden of data collection for correlation specification can be unattainably challenging (Lurie and Goldberg 1998). In particular, high-dimensional dependence modelling can be overly restrictive, mostly due to three well-known challenges, which can be collectively referred to as the curse of dimensionality. First of all, the number of pairwise correlations required to fully specify a correlation matrix increases quadratically. Although the general perception of large-scale projects changes over time, projects with thousands or more activities are becoming increasingly common in practice (GAO 2015, pp.102-104; Safran 2020). For example, a risk model with 1,000 variables requires assessments of ${}_{1000}C_2$ = 499,500 correlation coefficients. The burden of data collection in this scale, either from historical data or with expert judgment, would be practically unattainable, or simply not economical. Even more challenging, there are also situations where pair-wise correlations are restricted by the selection of marginal distributions (Demirtas and Hedeker 2011; Lurie and Goldberg 1998). A more detailed discussion on the curse of dimensionality will be presented in Section 2.

A sensible way of dealing with the curse of dimensionality is to reconstruct the problem in a way that reduces the data collection and elicitation burden (Morgan et al. 1992). A decision maker may conveniently avert the dimensionality issues by adopting drastic simplification assumptions, while sacrificing the flexibility of representing various dependence combinations (Goh and Sim 2011; Trietsch et al. 2012). In the project control literature, Bayesian networks were examined as an analytic framework for factor modelling and adaptive project time updating (Cho 2009; van Dorp 2020). Cho (2009) presented a single-factor Bayesian model in which all activities in a project are influenced by a single resource factor. van Dorp (2020) also proposed a single-factor dependence model that employs a new family of power distributions, the two-sided power distributions, to represent the mode of a PERT (program evaluation and review technique) distribution. As a robust solution for large-scale risk analysis, however, single-factor approaches can be overly restrictive in the way that all pairwise correlations in the analysis are calibrated by a single factor. In these regards, a dependence model can be considered more realistic when it offers the flexibility of accounting for multiple risk factors commonly observed in real project settings.

Methodologically, however, increasing the number of risk factors for dependence specification is also prohibited by the quantity and quality of the data available for corresponding parameter estimation. Whenever available, empirical data from past projects or expert assessments should be used. When a project is more predictable with a plethora of similar projects in the past, empirical data or subjective assessments by experts can be used. At the same time, there are more challenging projects with unique scope, innovative methodologies, and increased complexities in terms of component interfaces and project scales. These projects are as a rule less predictable and can be hardly characterized with quantitative data collected from past projects. Consequently, the nature and degree of risks inevitable in such one-of-a-kind projects cannot be fully quantified using empirical data alone. Here we observe a dilemma, somewhat inevitable in project risk assessment: the less there exist relevant empirical data from similar projects, the more the need for a sensible risk assessment increases. As a viable alternative, subjective assessments of the pair-wise correlations can be employed. Yet, the efficacy of subjective correlation assessments rapidly diminishes as the number of random variables increases mostly due to the mathematical consistency required for a feasible correlation matrix (See **Section 2.1** for more details of this issue).

These observations indicate that the robustness of a solution to the dependence modelling problem for large-scale project risk analysis can be enhanced with three analytical features: multi-factor capability, applicability under limited empirical data, and dimensional scalability. Accordingly, the objective of this article is set to present a multi-factor dependence modelling framework that provides the flexibility of addressing the limited data availability, while preserving the scalability to high-dimensional project risks. To achieve this goal, we investigate a dependent vector that can be fully specified with three input elements:

$$\mathbf{X}_{\mathbf{r}} \sim \langle \mathbf{b}; \mathbf{r}, \boldsymbol{\Psi} \rangle \tag{2}$$

where $\mathbf{b} = (b_1,\ldots,b_d)^T$ is a vector of observable random variables of which marginals, $f_{\mathbf{b}} = \{f(b_i)\}$, (i=1,…,d) are specified independently prior to accounting for possible dependence; $\mathbf{r} = (r_1,\ldots,r_K)^T$, is a vector of association factor (AF) variables that are elicited and specified as the proxy of the pairwise dependence between the base variables (**b**); and $\boldsymbol{\Psi}= [\psi_{ik}]$ (i=1,…,d; k=1,…,K) is a d×K allocation matrix (AM) of binary elements, which defines the relationships between **b** and **r**.

Specifically, we present an analytic framework with two stages: the structured association (SA) and the multi-factor association model (MFAM). First, the SA establishes a hierarchical structure of all relevant AFs identified in a project, providing a qualitative solution to the multi-factor capability and the applicability to limited data for a robust dependence model. Then, the MFAM transforms the qualitative SA information into a quantitative, mathematically consistent correlation matrix ($\boldsymbol{\Sigma}_{\mathbf{r}}$) of a vector of specified marginals. Adopting analytic (non-simulation) approaches (i.e., the second-moment approach), the MFAM offers a computationally efficient algorithm, readily scalable to high dimensional risk modelling and analysis.

Note that inter-variable association may arise due to causal relationship between variables or a common factor that may affects two or more variables concurrently (Bolstad 2007, p.3). The key premise underlying our approach is that by selecting the AFs wisely a decision maker is able to establish a balance between the data availability and the modelling flexibility, while effectively mitigating the curse of dimensionality. A selection of the AFs can be considered wise if the marginal distributions of the factors and the corresponding allocation matrix can be elicited and specified using all relevant information readily available in standard project settings. In this paper, we focus on establishing stochastic association between project performance units (i.e., itemized costs and activity times, hereinafter PUs) using all relevant information accessible in standard project environments, for example, the work breakdown structure (WBS), resource plans, and a risk register (PMI 2013, p.163). It should be properly emphasized that project risk information from various sources is often available in unstructured forms (e.g., drawings, organizational plans, and resource plans) (Xing et al. 2019). In particular, a risk register is used to identify and track all relevant risks in a project and their attributes relevant to project outcomes (PMI 2013). The information embedded in such project plans is conspicuously observable and thus objective. Yet project plans exist, as a rule, in qualitative formats. In this study, we adopt an ontological approach to transform any qualitative association information reflected in project plans into quantitative dependence information expressed as a correlation matrix. Ontology, as a branch of philosophy, offers a flexible perspective on the evolving nature of projects

(Morris 2013, p.236). In analytic settings, ontology allows a framework that represents the knowledge in a domain as a set of concepts and their relationships (Rodger 2013) and has attracted growing attention in risk studies, for instance, in safety (Xing et al. 2019), supply chain (Palmer et al. 2018), and environment (Scheuer et al. 2013).

The main contributions of this article can be highlighted in three aspects.

- The SA-MFAM approach enhances the realism of dependence modelling by offering the flexibility of accounting for multiple risk factors observed in individual projects based on all relevant information readily available in individual projects.
- The MFAM yields an analytic, closed-form solution to the correlation matrix, which can be further parameterized and calibrated meeting the limited data availability in individual projects.
- Adopting a factor-driven approach, the MFAM always generates a mathematically consistent matrix, preserving the scalability to high-dimensional risk modelling and analysis.

The rest of this article is organized as follows. The following section outlines the challenges in large-scale dependence modelling and presents the SA technique as a viable solution. Section 3 formulates the MFAM. In Section 4, we carry out a set of credibility tests and evaluate the performance of MFAM against Monte Carlo simulation. Section 5 demonstrates the implications of dependence (or ignoring dependence) in project decision making using three SA-MFAM applications. Conclusions and future research issues are summarized in Section 6.

2 Structured Association

2.1 Large-scale correlation assessment

Uncertainty models, in a generic setting, have a form of a joint distribution with three components: (i) a function of random variables in d-dimension, $G(\mathbf{X})$, $\mathbf{X} = (X_1,\ldots,X_d)^T$, (ii) a set of univariate marginal distributions, $f_{\mathbf{X}} = \{f(X_i)\}, i = 1,\ldots,d$ and (iii) a set of dependence parameters, mostly in terms of a correlation matrix (i.e., a symmetric, positive semi-definite matrix with unit diagonal elements), $\mathbf{\Sigma_X} = ([\rho_{ij}]: -1 \le \rho_{ij} \le 1; 1 \le i, j \le d)$.

In large-scale project risk analyses, the efforts to develop a fully-specified feasible correlation matrix are easily hindered by three factors: (i) high dimensionality, (ii) sample size required for adequate correlation estimates, and (iii) mathematical consistency required for a feasible correlation matrix (Bedford and Cooke 2002; Kim 2021; Kurowicka and Cooke 2006, p.84). As mentioned earlier, the high dimensionality alone causes a dire challenge in terms of the number of pairwise correlations. The schedule risk analysis for a medical center project discussed in GAO (2015, p.104), for example, involves 6,098 activities, in which the full-scale correlation matrix requires assessments of over 18.5 million correlation coefficients. There are more practical challenges involved in the correlation-driven dependence approach. Inter-variable dependence is a *delicate* statistical property and data collection for quantitative dependence estimation can be unattainably burdensome. For instance, the sample size required to detect a non-zero product-moment correlation between two random variables at a mild significance of $\alpha = 0.1$ and the power of $1 - \beta = 0.8$ is 617 for small effect size (0.10), 68 for medium effect size (0.30), and 22 for large effect size (0.50) (Cohen 1988, p.101). In large-scale projects with hundreds or thousands of PUs, collecting statistically meaningful data to elicit a full set of correlation coefficients can be unattainably challenging. Moreover, the temporary and unique nature of projects implies that objective data for dependence elicitation become less available for unique projects with rare constraints in which risk accounts are most needed. When objective data is not available, expert judgement can be a sensible alternative (Werner et al. 2017). However, matching statistical dependence measures with subjective judgements is not straightforward, even for trained statisticians (Clemen et al. 2000).

There is also a theoretical challenge. When a correlation matrix is constructed from independently conducted pair-wise assessments, the resulting correlation matrix is not guaranteed to be feasible. Consider three cost items in a project, say X_1, X_2, and X_3, of which correlation matrix is fully specified with three correlation coefficients: ρ_{12}, ρ_{13}, and ρ_{23}. When two of the three correlation coefficients are known, for instance, $\rho_{12} = 0.6$ and $\rho_{13} = 0.3$, the feasible range of ρ_{23} is restricted to [-0.58, 0.94], a sub-set of the range of partial correlation coefficient, [-1, 1]. The loss of a feasible correlation space in high-dimensional dependence models becomes substantial as the dimension increases (Ghosh and Henderson 2003), making correlation-driven approaches even less attractive.

2.2 Structured association

The itemized costs and activity times in a project are hardly isolated variables, statistically independent from each other. More likely, a certain degree of association tends to build up between closely linked PUs within a project, due to various common factors such as work types, combined contracts, renewable resources shared by multiple tasks, concurrent exposure to other internal (e.g., culture, organizational structure, project maturity) and external (e.g., weather, material suppliers, and labor markets) factors (GAO 2020, Chapter 12; Hulett 2011, Chapter 5; van Dorp 2005). It is sensible then to properly account for the inherent association binding multiple PUs as a coherent group, reflecting the distinct procedures of being planned and executed within a project.

Definition 1: Structured Association

Dependence in a broad sense is a measure of statistical association between random variables. The structured association (SA) is defined as the operational association between two or more performance units (PUs) in a project due to common factors emerging from, for example, (i) the inherent work attributes shared by the PUs and (ii) the way that PUs are planned and executed following the structured procedures and practices in project management.

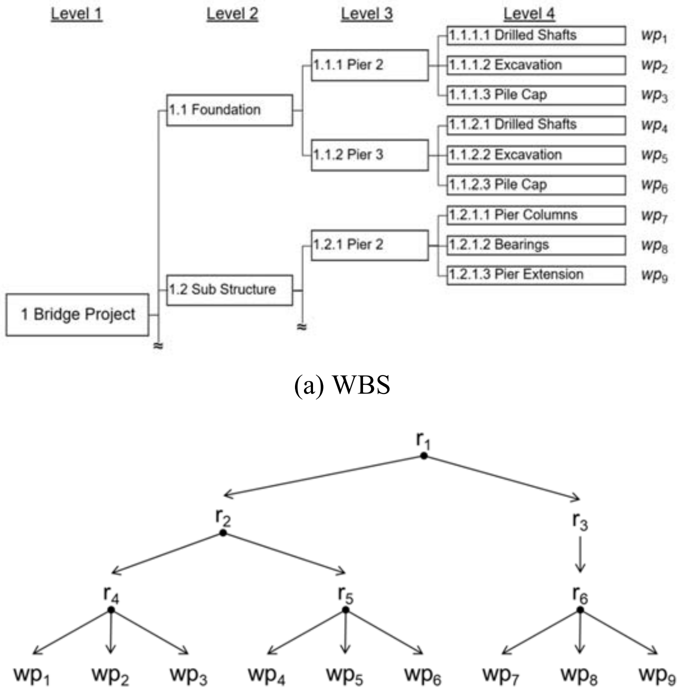
In this study, the SA is constructed as a set of hierarchical factors representing relative associations between project PUs. This section presents two distinct SA elicitation procedures: the top-down SA and the bottom-up SA. The former extracts relative association information from the WBS of individual projects, whereas the latter from the common resources shared by multiple PUs.

2.3 Top-down SA

The WBS of a project shows a hierarchical decomposition of the total scope of a project (PMI 2013). A WBS represents hierarchical association between work items in terms of levels and work packages. Starting with the project itself (Level 1), work components are divided into smaller, more controllable items until the work item can be properly planned, executed, and tracked by a single entity (i.e., a person or an organization). When a work item on the upper level ("parent") is decomposed into two or more, smaller items on a lower level ("children"), the 100 percent rule applies: the sum of the children's scope is equal to the scope of the parent (PMI 2013). Work packages (WPs) are the work items at the lowest level(s) of a WBS. Thus, the total scope of a project is determined as the sum of all WPs in the WBS. In a similar sense, the total (direct) project cost is estimated from the sum of all WP costs.

The WBS level method:

From the nature of the hierarchical decomposition, it can be inferred that a WBS is a qualitative representation of the relative association between work items (Reinschmidt 2009). That is, the actual performance of the *child* items from a same *parent* tends to share a certain degree of operational association. Consider the WBS of a bridge project in Fig. 1 (a). The WBS was reconstructed from an actual schedule of a $230 million bridge project. The WBS consists of four levels and more than 20 work packages (only the first nine are shown). Each item in the WBS is uniquely numbered with a hierarchical code system. For example, Item 1.1.2.1 is a WP at Level 4 with Item 1.1.2 as its *parent*.



(a) WBS

(b) SA tree of AFs

Fig. 1 Top-down SA tree for a bridge project.

Reinschmidt (2009) introduced a simplified way of creating a correlation matrix by grouping the WPs under the same parent in terms of a *level* parameter. That is, the degree of correlation between two WPs is specified by the number of parent tasks shared by them. Then, the degree of association between the WPs in Fig. 1 (a) can be expressed with three parameters: ρ_1, ρ_2, and ρ_3 for one, two, and three levels, respectively. Specifically, a degree of association function, $a(wp_i, wp_j)$, between two WPs, wp_i and wp_j, can be defined as the number of parent levels shared by the two WPs: $a(wp_i, wp_j) = \#[\{l_k\}, l_k \in \{l_k\}_{wp_i} \cap \{l_k\}_{wp_j}]$, where l_k denotes Level-k, $\{l_k\}_{wp_i}$ ($\{l_k\}_{wp_j}$) denotes the set of levels associated with wp_i (wp_j), and #[•] is the cardinality function that counts the number of elements in a set. For example, $a(wp_1, wp_2) = \#[\{l_1, l_2, l_3\}] = 3$, $a(wp_1, wp_4) = \#[\{l_1, l_2\}] = 2$, and $a(wp_1, wp_7) = \#[\{l_1\}] = 1$. Then the correlation coefficient between two WPs is parameterized accordingly, as $\rho(wp_1, wp_2 \,|\, a(wp_1, wp_2)) = \rho(wp_1, wp_2 \,|\, 3) = \rho_3$, $\rho(wp_1, wp_4 \,|\, a(wp_1, wp_4)) = \rho(wp_1, wp_4 \,|\, 2) = \rho_2$, and $\rho(wp_1, wp_7 \,|\, a(wp_1, wp_7)) = \rho(wp_1, wp_7 \,|\, 1) = \rho_1$.

The WBS code method:

In this paper, we generalize the WBS level method to a hierarchical multi-factor code method. Specifically, all parent items in a WBS are converted into independent AFs, $\mathbf{r} = (r_1, \ldots, r_K)^T$. The primary merits from the extension are twofold. First, we can account for uneven associations between the work items within the same level. More importantly, we can account for the mixed effects of multiple AFs according to the level of details of individual WBS. For instance, the level 3 under 'Sub Structure' in Fig. 1 (a) may induce different degrees of associations from the level 3 under 'Foundation', due to the distinct properties of the two work types. Specifically, the association between wp_1 (1.1.1.1) and wp_4 (1.1.2.1) is influenced by the marginal association induced by two *parent* items, 1.0 (Level 1) and 1.1 (Level 2), whereas the association between wp_4 (1.1.2.1) and wp_7 (1.2.1.1) is affected by the association embedded in the global parent item, *Bridge Project*. Then, the hierarchy embedded in the WBS can be systematically transformed into a hierarchical factor tree as in Fig. 1 (b). Subsequently, we generalize the association operator as a union of two sets of AFs: $a(wp_i, wp_j) = \{r_k\}_{k \in wp_i} \cap \{r_k\}_{k \in wp_j}$. For example, $a(wp_1, wp_4) = \{r_1, r_2, r_4\} \cap \{r_1, r_2, r_5\} = \{r_1, r_2\}$ and $a(wp_1, wp_7) = \{r_1, r_2, r_4\} \cap \{r_1, r_3, r_6\} = \{r_1\}$.

Note that the SA factors from the WBS code method can be grouped into two classes: the global factor and the local factors. The global factor works at the project level and influences all WPs in the project, whereas the local factors are applied to subgroups of the WBS. The global factor may represent soft performance factors at the project level (Williams 2003), for instance, the "project degree of dependence" in van Dorp (2020). From the combination of these factors, the top-down SA provides additional flexibility of specifying relative association imposed on project PUs in accordance with a pre-established WBS in a project.

2.4 Bottom-up SA

Once a WBS is developed for a project, detailed planning for individual WPs follows. Resource planning and leveling plays a central role at this stage of project planning. Resource planning is a process of identifying all required resources and allocating them to individual tasks in an optimal way that satisfies the project objectives and relevant constraints. Naturally, resource planning involves strategic trade-offs between alternative courses of actions and optimization of the expected outcomes. As a result, resource constraints across multiple activities build up in a project, creating additional dependence between activities, which may trigger additional sensitivity (Song et al. 2021) and increased risks in activity times and costs (Asadabadi and Zwikael 2021).

The bottom-up SA elicits inter-variable association information from such deliberate planning efforts of resource utilization and schedule optimization, which is broadly referred to as resource-leveling. It is worth noting that resource allocation plans in a project may change according to the priority of the project. As a result, the bottom-up SA approach may provide additional insights into the nature and properties of project-specific inter-variable association, which cannot be obtained from empirical data collected from past projects. Consider two alternative plans for a sub-project in Fig. 2. The sub-project has six WPs of which costs and times are estimated as, for illustration purposes, $1,000 and one week, uniformly across all WPs. The first plan (Alternative 1) involves one crew (Crew A) and one crane. That is, all the six WPs are carried out in a sequence by a single crew. In addition, one crane is used for wp_2 and wp_6. As a result, the sub-project duration is six weeks. Suppose that the sub-project forms the critical path of the whole project and the project considers accelerating the schedule as in Alternative 2. The project adopts the fast-tracking technique and divides the sub-project into two parallel groups: $\{wp_1, wp_2, wp_3\}$ and $\{wp_4, wp_5, wp_6\}$. Moreover, each group is carried out by independent crews, Crew A and Crew B, respectively. As a result, the sub-project may finish three weeks earlier.

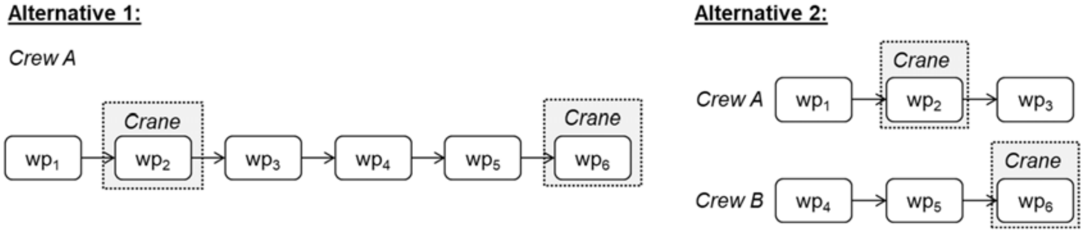


Fig. 2 Resource-driven bottom-up SA.

When two or more tasks share a resource in a project, a certain degree of association tends to build up in their actual performance. Such an association may not be purely statistical, but somewhat operational in that the degree of association is induced by the way that the tasks are planned and carried out at the operational level. Specifically, the WP times and costs are often intertwined via the production rate of the allocated resources. Consider the generic formulas for time and cost estimates of typical construction items:

$$time_i = \frac{q_i}{p_i};\; cost_i = (m_i + l_i + e_i)(1 + O\,\&\,P) = (m_i + u_i \times time_i)(1 + O\,\&\,P) \tag{3}$$

where q_i is the work quantity; p_i is the production rate of allocated resources (i.e., labor and equipment); m_i is the material cost; l_i is the labor cost; e_i is the equipment cost; u_i is the renewable resource (labor plus equipment) rate per time unit; and *O&P* is the overhead and profit rate. The *O&P* is typically determined at the project level, while all the other quantities are estimated for individual tasks.

Itemized estimates as in Eq. (3) indicate that task times and costs are influenced by the detailed resource allocation and the performance of the resources. Specifically, the production rate (p_i) is a crucial attribute of renewable resources. Skilled labors and crane operators would have better and more predictable production rates than those with limited experience or training. Likewise, a larger crane shared by multiple tasks would concurrently influence the production rates of the tasks. As a result, the production rate of a shared resource would affect both the times and costs of the associated activities. From these observations, the bottom-up SA establishes a hierarchical tree structure from detailed resource plans at the lower levels of an individual project. For example, schematic dependence structures underlying the two alternatives in Fig. 2 can be established in terms of three AFs: Crew A, Crew B, and Crane. The resulting factor loading diagrams in Fig. 3 indicate that the actual performance of the WPs may have different association structures depending on the specifics of the allocated resources. The cost and time performance of Alternative 1 are influenced by two AFs: r_1 (Crew A) and r_2 (Crane), whereas the performance of Alternative 2 is influenced by three AFs: r_1, r_2, and r_3 (Crew B). Then, an association function between two work packages can be defined accordingly. For example, $a(wp_2, wp_6; Alt.1) = \{r_1, r_2\} \cap \{r_1, r_2\} = \{r_1, r_2\}$, whereas $a(wp_2, wp_6; Alt.2) = \{r_1, r_2\} \cap \{r_2, r_3\} = \{r_2\}$.

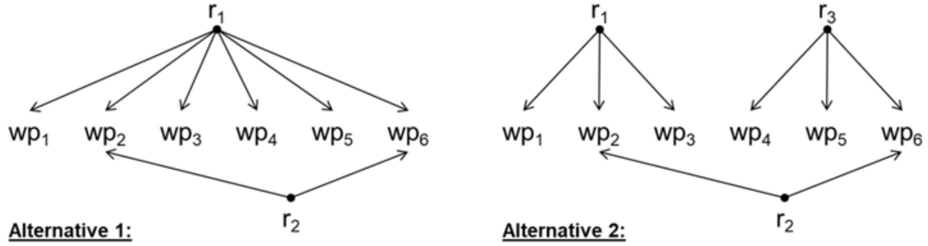


Fig. 3 Resource-driven factor loading. Note that r$_1$, r$_2$, and r$_3$ represent the SA induced by Crew A, Crane, and Crew B, respectively.

2.5 SA elicitation characteristics

The SA approaches provide a systematic way of parameterizing the relative degree of association between relevant PUs in a project. Note that a set of hierarchical factors can be constructed in a more trivial way. For instance, risk events identified in a risk register can be structured according to the risk breakdown structure. In this regard, the AFs and their relationships with the individual PUs can be constructed utilizing all relevant information available in individual projects. Accordingly, the SA transforms as-is project plans to a hierarchical dependence structure of a finite set of AFs. Specifically, the SA elicitation procedures can be characterized by three features.

- ***Project-specific association***. Each project is unique in one way or another, and the most reliable information on the internal dynamics between the PUs would be available inside. The SA utilizes qualitative information generated in individual projects. An ideal WBS for the same project may change across project organizations, reflecting the relative strengths and strategies of individual organization. Project resource plans are optimized in a way that maximizes the efficiency of available resources while minimizing detrimental effects on project outcomes. The SA systematically reflects such unique features of individual projects in terms of the structured AFs.
- ***As-is information***. The primary sources of the SA elicitation are standard documents available in most projects. The WBS-driven SA represents a top-down perspective on the inter-variable association structure, whereas the resource-driven SA accounts for a bottom-up perspective. The two approaches are not necessarily mutually exclusive but complementary by nature. Not discussed in detail here, but risks documented in a risk register would also serve as a reliable source of AFs.
- ***Dimensional scalability***. Utilizing the hierarchical nature of project planning procedures, SA approaches provide intuitive structures of which parameters can be systematically elicited by expert judgement or other relevant information. As a result, the SA approaches provide a robust solution scalable to large-scale projects with a marginal burden of factor elicitation and data collection.

The AFs in the SA approaches represent plausible association between the PUs in a project. An AF is not necessarily a cause that directly affects the uncertainty of individual PUs. In this regard, the AFs in this study need to be distinguished from the "systemic bias across a project" in the linear association (Trietsch 2005; Trietsch et al. 2012) or the risk drivers (Hulett 2011) as a cause of dependence. More specifically, these models are designed to induce dependence by influencing the marginal distributions (i.e., their expected values and/or variabilities). Conceptually, the AF in this paper is a broader concept, encompassing the causal relationships. As a result, the SA provides additional flexibility in the way that the dependence specification can be decoupled from the specification of the marginals. These issues will be further discussed in Section 3.4 and Section 5.

To be effective, the SA approaches need to overcome two methodological restrictions. The AFs elicited from project plans are qualitative by nature. The relative significance of individual AFs and their confounding effects on project performance need to be quantitatively assessed and transformed into coherent correlation coefficients. For instance, a potential difference between the dependence induced by $\{r_1\} \cap \{r_1, r_2\}$ for (wp_1, wp_2) and $\{r_1, r_2\} \cap \{r_2, r_3\}$ for (wp_2, wp_6) in the Alternative 2 of Fig. 3 needs to be explicitly accounted for. Also, the SA requires an analytic mechanism that numerically computes the combined effect of multiple AFs. To address these issues, a computationally efficient analytic (non-simulation) model is formulated and presented in the following section.

3 Multi-factor Association Model

The SA-induced dependence between PUs involves, as a rule, multiple factors. This section presents a multi-factor association (MFA) model that transforms the relative association information embedded in a hierarchical set of multiple AFs into a mathematically consistent correlation matrix. Characteristic properties and practical implementation options of the MFA model are also highlighted.

3.1 Multi-factor association

The key notations used in the derivation are summarized below.

- (•): a row vector of variables.
- {•}: a set of elements.
- $\mathbf{X} = (X_1,\ldots, X_d)^T$: a d×1 vector of factored variables.
- $\mathbf{b} = (b_1,\ldots, b_d)^T$: a d×1 vector of base variables.
- $\mathbf{r} = (r_1,\ldots, r_K)^T$: a K×1 vector of AFs.
- θ_i: a set of AFs associated with b_i.
- $R_i = \prod_{k\in\theta_i} r_{ik}$: an aggregate factor as a product of the AFs associated with b_i.
- $E[\bullet]$: the expectation operator that returns the mean value of a variable.
- $V[\bullet]$: an operator that returns the variance of a variable.
- $\sigma[\bullet]$: an operator that returns the standard deviation of a variable.
- $\rho[\bullet,\bullet]$: the correlation coefficient between two random variables.

Consider a vector of d random variables $\mathbf{X} = (X_1,\ldots, X_d)^T$. The marginal distributions and the dependence between the marginals are modeled following two steps. First, the *a priori* estimates of the marginals of **X** are represented in terms of d independent variables, $\mathbf{b} = (b_1,\ldots, b_d)^T$. Second, the possible association between the marginals is specified using a set of AFs, $\mathbf{r} = (r_1,\ldots, r_K)^T$. Then we model a dependent vector **X** in terms of the product of all AFs linked to individual base variables.

$$\{X_i\} = \{R_i b_i\},\ R_i = \prod_{k\in\theta_i} r_{ik} \tag{4}$$

where i =1,…,d; R_i is an aggregate factor defined as a product of the AFs (r_{ik}) linked to a variable b_i, and; θ_i is a set of AFs associated with b_i. Since the marginals of **X** are constructed from **b** multiplied by the relevant AFs, the former is referred to as 'factored' variables, whereas the latter is referred to as 'base' variables hereinafter. For ease of reference, a d-variate vector with K AFs is referred to as the K-th order d-variate MFA model and denoted as MFA[d, K]. Note that when two aggregate factors, R_i and R_j, share a set of common AFs the corresponding factored variables, X_i and X_j, become statistically dependent: $\theta_i \cap \theta_j \neq \varnothing \to \rho[X_i, X_j] > 0$, where $\rho[X_i, X_j]$ is the correlation coefficient of X_i and X_j. For example, a MFA[4,3] model is shown in Fig.4. The arrows in Fig. 4 represent the mapping relationships between the base variables and the AFs. From the relationships, four aggregate factors can be construction accordingly: $R_1 = r_1 \times r_2$, $R_2 = r_1 \times r_3$, $R_3 = r_1 \times r_2 \times r_3$, and $R_4 = r_1$.

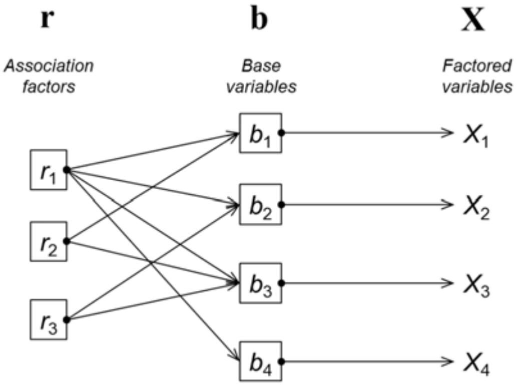


Fig. 4 MFA[4,3] model.

3.1.1 Normalizing AFs

The factored variable in Eq. (4) indicates that the effects of individual AFs on the base variables can be uneven with distinct combinations of relevant AFs ($\theta_i \neq \theta_j$). When multiple AFs are in play, the simplicity of dependence specification and the computational efficiency can be significantly improved by normalizing the R_i's with respect to the entire AF vector, **r**. To this end, we introduce the allocation matrix ($\boldsymbol{\Psi}$) as a systematic way of loading a set of K AFs onto a d-variate random vector.

$$\boldsymbol{\Psi} = \underset{d\times K}{\left[\psi_{ik}\right]}, \quad \psi_{ik} = \begin{cases} 1 & r_k \in \theta_i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $i = 1,\ldots,d$ and $k = 1,\ldots, K$. $\boldsymbol{\Psi}$ is a $d\times K$ matrix with binary indicators. For example, the MFA[4,3] model in Fig. 4 includes one global AF (r_1) that influences all the random variables and two local AFs (r_2 and r_3) that have limited impacts only on the relevant variables. Then the corresponding allocation matrix is constructed as:

$$\boldsymbol{\Psi} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

Given a set of K AFs and the corresponding allocation matrix, a factored variable (X_i) can be expressed without loss of generality as a product of the entire AFs, $\forall r_k \in \mathbf{r}$:

$$X_i = R_i b_i = \prod\nolimits_{k=1}^{K} r_k^{\psi_{ik}} b_i \tag{6}$$

Eq. (6) normalizes Eq. (4) in terms of the global K replacing the local association sets (θ_i's) for individual marginals.

3.1.2 MFA variable properties

An algebraic model for the mixed effects of multiple AFs on the pairwise correlation between random variables is formulated using the first two moments of the base variables and the AFs. The second-moment approach, also referred to as the method of moment (MOM), has been effectively used in uncertainty modelling (Morgan et al. 1992) and project risk studies (Kim 2019). Accordingly, the base variables and the AFs are specified in terms of their means and variances. That is, $b_i \sim \langle E[b_i], V[b_i] \rangle$ and $r_k \sim \langle E[r_k], V[r_k] \rangle$. Then the mean and variance of a factored variable are expressed as a product of independent variables.

Theorem 1. Product of Random Variables (Garvey et al. 2016)(p.204)

Consider the product of K random variables,

$$R = r_1 r_2 \cdots r_K \tag{7}$$

If $\{r_k\}$ (k=1,…,K) are independent and their marginal means and variances are known, the mean and variance of the product of the variables are computed as:

$$E[R] = \prod_{k=1}^{K} E[r_k] \tag{8}$$

$$V[R] = \prod_{k=1}^{K} E[r_k^2] - \prod_{k=1}^{K} E[r_k]^2 = \prod_{k=1}^{K} (V[r_k] + E[r_k]^2) - \prod_{k=1}^{K} E[r_k]^2 \tag{9}$$

Note that the mean of the product of independent variables is the product of the means of the variables, whereas the variance of the product of independent variables is the product of the means of the squared variables minus the product of the squared means of the variables.

Proof. See **Appendix A.** ■

The mean and variance in **Theorem 1** are exact no matter what the distributions over the component variables. If the r_k's are lognormal, the product variable will be exactly lognormal. Even if the r_k's are not lognormal, the central limit theorem indicates that the distribution of R will approximately approach to the lognormal as K increases, provided the r_k's are independent.

From **Theorem 1**, the first two moments of factored variables in Eq. (6) are expressed as:

$$E[X_i] = E[R_i b_i] = \prod_{k=1}^{K} E[r_k^{\psi_{ik}}] E[b_i] \tag{10.a}$$

$$V[X_i] = V[R_i b_i] = \prod_{k=1}^{K} E[(r_k^{\psi_{ik}})^2] E[b_i^2] - \prod_{k=1}^{K} E[r_k^{\psi_{ik}}]^2 E[b_i]^2 \tag{10.b}$$

Eqs. (10) are collectively referred to as the MFA formulas.

3.2 MFA correlation

The MFA formulas provide a *closed-form* solution to the construction of a complete correlation matrix for the factored variables. Given a set of AFs (**r**) and the corresponding allocation matrix (**Ψ**), the pairwise covariance between two factored variables, X_i and X_j, is $Cov[\prod_{k=1}^{K} r_k^{\psi_{ik}} b_i, \prod_{k=1}^{K} r_k^{\psi_{jk}} b_j]$. However, an analytic solution to the covariance is not trivial when the mapping of AFs to base variables is non-uniform. To solve this problem, we apply a subset approach. The goal is to formulate a computationally efficient algorithm that converts the factor-driven specification $\langle \mathbf{b}; \mathbf{r}, \mathbf{\Psi} \rangle$ to a complete correlation coefficient matrix, $\mathbf{\Sigma_r}$.

Let θ_i and θ_j denote two sets of AFs associated with b_i and b_j, , respectively. For example, the MFA[4,3] in Fig. 4 indicates that $\theta_1 = \{r_1, r_2\}, \theta_2 = \{r_1, r_3\}, \theta_3 = \{r_1, r_2, r_2\}$, and $\theta_4 = \{r_1\}$. Without loss of generality, θ_i and θ_j can be expressed as a union of two sub-sets:

$$\theta_i = \theta_{i\cap j} \cup \theta_{i-j} \quad \text{and} \quad \theta_j = \theta_{i\cap j} \cup \theta_{j-i} \tag{11}$$

where $\theta_{i\cap j}$ is the intersection of θ_i and θ_j, whereas θ_{i-j} (or θ_{j-i}) is the set difference of θ_i and θ_j (or θ_j and θ_i). That is, $\theta_{i-j} = \theta_i \setminus \theta_j = \{r_k \in \theta_i \mid r_k \notin \theta_j\}$ (the relative complement of θ_j in θ_i). For example, if $i = j$, $\theta_{i\cap j} = \theta_i = \theta_j$ and $\theta_{i-j} = \theta_{j-i} = \varnothing$ hold. Subsequently, for $i \neq j$, the covariance between two factored variables linked by arbitrary sets of AFs can be represented with three aggregate factors:

$$R_{i\cap j} = \prod\nolimits_{k\in\theta_{i\cap j}} r_k,\ R_{i-j} = \prod\nolimits_{k\in\theta_{i-j}} r_k, and\ R_{j-i} = \prod\nolimits_{k\in\theta_{j-i}} r_k \tag{12}$$

Then, the covariance of two factored variables is generalized into:

$$Cov[X_i, X_j] = Cov\left[\prod\nolimits_{k=1}^{K} r_k^{\psi_{ik}} b_i, \prod\nolimits_{k=1}^{K} r_k^{\psi_{jk}} b_j\right] = Cov[R_{i\cap j} R_{i-j} b_i, R_{i\cap j} R_{j-i} b_j] \tag{13}$$

where $R_{i\cap j}, R_{i-j}, R_{j-i}$,b_i, and b_j are independent, by definition. Introducing $g_i = R_{i-j} b_i$ and $g_j = R_{j-i} b_j$, the covariance of X_i and X_j is expressed as

$$Cov[X_i, X_j] = Cov[R_{i\cap j} g_i, R_{i\cap j} g_j] = E[(R_{i\cap j} g_i - E[R_{i\cap j} g_i])(R_{i\cap j} g_j - E[R_{i\cap j} g_j])] \tag{14}$$

which can be further elaborated as:

$$\begin{aligned} Cov[R_{i\cap j} g_i, R_{i\cap j} g_j] &= E[R_{i\cap j}{}^2 g_i g_j] - E[R_{i\cap j}]^2 E[g_i] E[g_j] \\ &= E[R_{i\cap j}{}^2] E[g_i] E[g_j] - E[R_{i\cap j}]^2 E[g_i] E[g_j] = V[R_{i\cap j}] E[g_i] E[g_j] \end{aligned} \tag{15}$$

Then, from $E[g_i] = E[R_{i-j} b_i]$ and $E[g_j] = E[R_{j-i} b_j]$,

$$Cov[X_i, X_j] = V[R_{i\cap j}] E[R_{i-j}] E[b_i] E[R_{j-i}] E[b_j] \tag{16}$$

where $E[R_{i-j} b_i] = E[R_{i-j}] E[b_i]$ (and $E[R_{j-i} b_j] = E[R_{j-i}] E[b_j]$) holds because R_{i-j} (R_{j-i}) and b_i (b_j) are independent.

Eq. (16) indicates that a non-zero covariance is induced by the common aggregate factor, $R_{i\cap j}$. This is intuitive because, if $\theta_i \cap \theta_j \equiv \varnothing$, the covariance should be zero because $V[R_{i\cap j}] = 0$.

From Eqs. (10) and (16), a correlation coefficient between factored variables is computed by

$$\rho[i,j]_{i\neq j} = Correl[X_i, X_j]_{i\neq j} = \frac{V[R_{i\cap j}] E[R_{i-j}] E[b_i] E[R_{j-i}] E[b_j]}{\sqrt{V[R_i b_i]}\sqrt{V[R_j b_j]}} \tag{17}$$

Eq. (17) indicates that the correlation between two factored variables can be systematically computed using only the first two moments of the AFs and the corresponding base estimates, b_i and b_j. Specifically,

$$V[R_{i\cap j}]_{i\neq j} = \prod\nolimits_{k\in\theta_{i\cap j}} E[r_k^2] - \prod\nolimits_{k\in\theta_{i\cap j}} E[r_k]^2 \tag{18.a}$$

$$E[R_{i-j}]_{i\neq j} = \prod\nolimits_{k\in\theta_{i-j}} E[r_k] \quad \text{and} \quad E[R_{j-i}]_{i\neq j} = \prod\nolimits_{k\in\theta_{j-i}} E[r_k] \tag{18.b}$$

3.3 MFA programming

The computation of the aggregate factor moments in Eq. (18) can be cumbersome, involving products of uneven sub-sets of AFs for *i* and *j*. The additional burden can be immense in high dimensional risk modelling. Yet, the computational efficiency and the scalability to large-scale cases can be considerably improved by adopting the Hadamard formulation (Fallat and Johnson 2007; Styan 1973).

3.3.1 Hadamard formulation

Given $\mathbf{\Psi}$ and two factored variables X_i, and X_j, we introduce a union identity vector, $\mathbf{D}_{ij} = (d_{1;ij}, ..., d_{K;ij})$, which is defined as a vector of binary indicators of the elements of $\theta_{i\cap j}$ with respect to the entire AFs.

$$\{d_{k;ij}\}_{k=1,...,K} = \begin{cases} 1 & if\, \psi_{ik} = \psi_{jk} = 1 \\ 0 & otherwise \end{cases} \tag{19}$$

Note that $d_{k;ij} = \psi_{ik} \times \psi_{jk}$ holds and $\mathbf{D}_{ij}$ can be algebraically generated as the Hadamard product of two row vectors of $\mathbf{\Psi}$, $\psi_{i\bullet}$ and $\psi_{j\bullet}$, where $\psi_{i\bullet}$ ($\psi_{j\bullet}$) is a vector of the *i*-th (*j*-th) row of $\mathbf{\Psi}$. Then, given $\mathbf{\Psi}$, a union identity vector of an arbitrary pair of factored variables can be systematically constructed. Specifically, the aggregate factors in Eq. (12) can be consistently represented with respect to the whole set of AFs.

$$R_{i\cap j} = \prod\nolimits_{k=1}^{K} r_k^{d_{k;ij}} \; ; \; R_{i-j} = \prod\nolimits_{k=1}^{K} r_k^{\psi_{ik} - d_{k;ij}} \; ; \text{and } \; R_{j-i} = \prod\nolimits_{k=1}^{K} r_k^{\psi_{jk} - d_{k;ij}} \tag{20}$$

Moreover, the correlation coefficient in Eq. (17) can be rewritten as:

$$\rho[i,j]_{i\neq j} = \frac{\left(\prod E[r_k^2]^{d_{k;ij}} - \prod E[r_k]^{2d_{k;ij}}\right) E[\prod r_k^{\psi_{ik} - d_{k;ij}}] E[\prod r_k^{\psi_{jk} - d_{k;ij}}] E[b_i] E[b_j]}{\left(\prod E[r_k^2]^{\psi_{ik}} E[b_i^2] - \prod E[r_k]^{2\psi_{ik}} E[b_i]^2\right)^{0.5} \times \left(\prod E[r_k^2]^{\psi_{jk}} E[b_j^2] - \prod E[r_k]^{2\psi_{jk}} E[b_j]^2\right)^{0.5}} \tag{21}$$

where all the products hold for *k* = 1,…, *K*.

Formulating correlation coefficients in terms of unified indices as in Eq. (21) provides a significant computational advantage. Note that the product function is available in most programming languages (e.g., **prod**() function in Matlab and R). In particular, the **PRODUCT**() in Microsoft Excel is a viable option. To this end, the correlation coefficient in Eq. (21) can be rewritten in a vector format as:

$$\rho[i,j]_{i\neq j} = \frac{\left(\prod \mathbf{E}_{rs}^{\circ(\psi_{i\bullet}\circ\psi_{j\bullet})} - \prod \mathbf{E}_r^{\circ 2(\psi_{i\bullet}\circ\psi_{j\bullet})}\right) \prod \mathbf{E}_r^{\circ(\psi_{i\bullet} - \psi_{i\bullet}\circ\psi_{j\bullet})} \prod \mathbf{E}_r^{\circ(\psi_{j\bullet} - \psi_{i\bullet}\circ\psi_{j\bullet})} (\mathbf{E}_b)_i (\mathbf{E}_b)_j}{\left[(\prod \mathbf{E}_{rs}^{\circ\psi_{i\bullet}})(\mathbf{E}_{bs})_i - (\prod \mathbf{E}_r^{\circ 2\psi_{i\bullet}})(\mathbf{E}_b)_i^2\right]^{0.5} \times \left[(\prod \mathbf{E}_{rs}^{\circ\psi_{j\bullet}})(\mathbf{E}_{bs})_j - (\prod \mathbf{E}_r^{\circ 2\psi_{j\bullet}})(\mathbf{E}_b)_j^2\right]^{0.5}} \tag{22}$$

where $\mathbf{E}_{rs}$ and $\mathbf{E}_r$ are vectors of $\{E[r_k^2]\}$ and $\{E[r_k]\}$ (*k* = 1,…,*K*), respectively, and ∘ is the Hadamard product operator for entry-wise computation (Fallat and Johnson 2007). For instance, $\psi_{i\bullet}\circ\psi_{j\bullet}$ in Eq. (22) returns a vector of ($\psi_{i1}\times\psi_{j1},\ldots,\psi_{iK}\times\psi_{jK}$), which is identical to the union identity vector ($\mathbf{D}_{ij}$) in Eq. (19). Also, the Hadamard power of a vector is defined as an entry-wise power function (Fallat and Johnson 2007). For instance, $\mathbf{E}_r^{\circ 2(\psi_{i\bullet}\circ\psi_{j\bullet})}$ in Eq. (22) returns $(E[r_1]^{2d_{1;ij}}, ..., E[r_K]^{2d_{K;ij}})$, which is identical to the index formulation in Eq. (21).

3.3.2 Vector formulation of correlation matrix

The Hadamard formulation of the AF-induced correlation can be readily programmed into a closed-form vector formulation of a correlation matrix. For instance, in Microsoft Excel, the product term of $\prod \mathbf{E}_r^{\circ 2(\psi_{i\bullet} \circ \psi_{j\bullet})}$ is computed by "{=PRODUCT(**E**$_r$^(2*INDEX(**Ψ**,*i*,0)*INDEX(**Ψ**,*j*,0)))}", where $\mathbf{E}_r$ is a (K×1) cell range and $\mathbf{\Psi}$ is a ($d \times K$) cell range of the allocation matrix. Note that the "{" and "}" in the formula are automatically inserted when the formula is entered by pressing Ctrl+Shift+Enter in Excel, forcing the arrays in the formula be computed entry-wise.

A generic algorithm for the MFA correlation computation is presented below.

```
MFA Algorithm (MFAA):
S.1 Read input parameters $\langle \mathbf{b}; \mathbf{r}, \mathbf{\Psi} \rangle$.
    §1.1 Read the base variables, $\mathbf{b} = (b_1,\ldots, b_d)^T$: the means ($\mathbf{E}_b$) and variances ($\mathbf{V}_b$).
    §1.2 Read the association factors, $\mathbf{r} = (r_1,\ldots, r_K)^T$: the means ($\mathbf{E}_r$) and variances ($\mathbf{V}_r$).
    §1.3 Read the allocation matrix, $\mathbf{\Psi}$ ($d \times K$).
S.2 Compute auxiliary parameter vectors $\mathbf{E}_{rs}$ and $\mathbf{E}_{bs}$.
    §2.1 $\mathbf{E}_{bs} = \{E[b_i^2]\} = \mathbf{V}_b + \mathbf{E}_b^{\circ 2}$
    §2.2 $\mathbf{E}_{rs} = \{E[r_k^2]\} = \mathbf{V}_r + \mathbf{E}_r^{\circ 2}$
S.3 Compute the means and variances of the factored variables, $\mathbf{X} = (X_1,\ldots, X_d)^T$.
    §3.1 $\mathbf{E}_X = \left\{ (\prod \mathbf{E}_r^{\circ \psi_{i\bullet}})(\mathbf{E}_b)_i \right\}, \quad i = 1,\ldots,d$
    §3.2 $\mathbf{V}_X = \left\{ (\prod \mathbf{E}_{rs}^{\circ \psi_{i\bullet}})(\mathbf{E}_{bs})_i - (\prod \mathbf{E}_r^{\circ 2\psi_{i\bullet}})(\mathbf{E}_b)_i^2 \right\}, \quad i = 1,\ldots,d$
S.4 Compute the aggregate factor moments.
    §4.1 $V[R_{i \cap j}] = \prod \mathbf{E}_{rs}^{\circ(\psi_{i\bullet} \circ \psi_{j\bullet})} - \prod \mathbf{E}_r^{\circ 2(\psi_{i\bullet} \circ \psi_{j\bullet})}$
    §4.2 $E[R_{i-j}] = \prod \mathbf{E}_r^{\circ(\psi_{i\bullet} - \psi_{i\bullet} \circ \psi_{j\bullet})}$
    §4.3 $E[R_{j-i}] = \prod \mathbf{E}_r^{\circ(\psi_{j\bullet} - \psi_{i\bullet} \circ \psi_{j\bullet})}$
S.5 Generate the correlation coefficient matrix ($i$ =1,…,$d$; $j$ =1,…,$d$)
    §5.1 $\mathbf{\Sigma}_\mathbf{r}(i, j \mid \mathbf{b}, \mathbf{r}, \mathbf{\Psi}) = \begin{cases} \dfrac{V[R_{i \cap j}] E[R_{i-j}] E[R_{j-i}] (\mathbf{E}_b)_i (\mathbf{E}_b)_j}{\left[ (\mathbf{V}_X)_i \times (\mathbf{V}_X)_j \right]^{0.5}} & if \;\; i \neq j \\ 1 & if \;\; i = j \end{cases} \qquad (23)$
```

3.4 Factor-driven project cost and time risk modelling

3.4.1 Project cost risk

A salient advantage of the closed-form correlation solution in Eq. (23) is that the effects of AFs on the marginals can be separated from the effects of the AFs on inter-variable dependence. As a result, the factor-induced correlation matrix ($\mathbf{\Sigma}_\mathbf{r}$) can be used either with the base random vector (**b**) or the factored random vector (**X**) in order to generate a vector of dependent variables. In project risk studies, a dependent vector of specified marginals (**b** or **X**) and a feasible correlation matrix ($\mathbf{\Sigma}_\mathbf{r}$) provides additional flexibility for modelling large-scale project cost risks. Specifically, when project cost is estimated as the sum of the itemized estimates of the project's work packages (GAO 2020; Garvey et al. 2016), the MFA offers two options for robust project cost risk assessment.

- Base project cost ($C_{\mathbf{b}}$) with the factor-induced correlation ($\mathbf{\Sigma_r}$):

$$E[C_{\mathbf{b}}] = \sum_{i=1}^{d} E[b_i] \text{ and } V[C_{\mathbf{b}}] = \boldsymbol{\sigma}_{\mathbf{b}}{}^{T} \mathbf{\Sigma}_{\mathbf{r}} \boldsymbol{\sigma}_{\mathbf{b}} \tag{24}$$

- Factored project cost ($C_{\mathbf{X}}$) with the factor-induced correlation ($\mathbf{\Sigma_r}$):

$$E[C_{\mathbf{X}}] = \sum_{i=1}^{d} E[X_i] \text{ and } V[C_{\mathbf{X}}] = \boldsymbol{\sigma}_{\mathbf{X}}{}^{T} \mathbf{\Sigma}_{\mathbf{r}} \boldsymbol{\sigma}_{\mathbf{X}} \tag{25}$$

where $\boldsymbol{\sigma}_{\mathbf{b}} = (\sqrt{V[b_1]},\ldots,\sqrt{V[b_d]})^T$ and $\boldsymbol{\sigma}_{\mathbf{X}} = (\sqrt{V[X_1]},\ldots,\sqrt{V[X_d]})^T$.

3.4.2 Project time risk

Project schedule risk is different. Project duration is determined by the longest path (the critical path) in a network schedule. With uncertainty into account, the project schedule risk is also influenced by the network complexity (i.e., multiple paths) along with the variabilities and possible correlation of activity times (Schonberger 1981). A robust way to capture the impacts of network complexity is Monte Carlo simulation, which inevitably involves random sampling of dependent vectors (Van Slyke 1963).

The MFA model provides *factor-driven* sampling procedures, which are distinguished from the conventional correlation-driven sampling in two notable ways. First, the MFA approach provides a computationally efficient procedure for high-dimensional random sampling. Factor-driven sampling generates dependent random vectors arithmetically from simple multiplication of independent distributions, without any constraints on the resulting correlation matrix or other sophisticated computations, for example, the Cholesky decomposition for normal joint distributions or the "NORmal To Anything" (NORTA) method (Cario and Nelson 1997). Second, more importantly, the MFA approach generates pair-wise correlations from dependent random samples (See Eq. (23)), not the other way around as in the correlation-driven sampling. As a result, the resulting correlation coefficient matrix is always mathematically consistent, satisfying the positive semi-definite property.

Proposition 1.

A factor-driven correlation matrix, $\mathbf{\Sigma}_{\mathbf{r}}(i, j \mid \mathbf{b}, \mathbf{r}, \mathbf{\Psi})$ in Eq. (23), is always mathematically consistent (i.e., symmetric, positive semi-definite, and with unit diagonal elements) in that there exist ainfinite number of dependent random samples that asymptotically match the correlation matrix.

Proof: Given a vector of independent marginals $\mathbf{b} = (b_1,\ldots, b_d)^T$ and a vector of AFs $\mathbf{r} = (r_1,\ldots, r_K)^T$, the MFA can draw an infinite number of dependent random vectors by multiplying the variables (Eq. (4)) according to the allocation matrix $\mathbf{\Psi}=[\psi_{ik}]$ ($i = 1,\ldots,d$; $k = 1,\ldots,K$). Therefore, the existence of a non-zero dependent random vector is guaranteed by the arithmetic nature of the sampling procedure. Then, the corresponding correlation matrix is always positive semi-definite in the sense that there always exist an infinite number of corresponding dependent random vectors.

3.4.3 Decoupling dependence from marginals

Proposition 1 suggests that the MFA-induced correlation matrix can be seamlessly implemented with any dedicated simulation software without any restrictions of mathematical consistency check or correlation adjustments. In practice, the MFA sampling can be implemented in two modes:

- **Full-MFA sampling**: Dependent random vectors are generated for the factored variables $\{X_i\}$, $i = 1,\ldots,d$, while satisfying the factor-induced correlation $\mathbf{\Sigma_r}$.
- **Base-MFA sampling**: Dependent random vectors are generated for the base variables $\{b_i\}$, $i = 1,\ldots,d$, while satisfying the factor-induced correlation $\mathbf{\Sigma_r}$. In this case, the marginals are decoupled from the correlation matrix.

The full-MFA sampling accounts for the effects of AFs on the variance of the marginals. That is, the AFs affect both the marginals and the inter-variable correlations. Specifically, Eq. (10.b) indicates that the variance of a factored variable is always greater than or equal to the variance of the corresponding base variable, $V[X_i] \geq V[b_i]$, where the equality holds when $\theta_i = \varnothing$. In practice, there are situations where an increase of marginal variances due to AFs is supported by empirical evidence. For instance, the AFs can be calibrated using historical cost overruns (or schedule delays). A possible discrepancy between a base estimate and the factored variable can be attributed to the biases in individual marginals as well as the factor-induced dependence.

The base-MFA sampling offers an alternative to the full-MFA sampling, numerically decoupling marginals from dependence specification. In practice, there are also situations in which a variance increase in individual marginals can be considered *duplicative* in that possible effects of the AFs on the aggregate uncertainty are accounted for primarily in terms of the factor-induced dependence. The base-MFA sampling eliminates the risk of unwarranted variance increase. The base-MFA sampling proceeds following three steps. First, generate a large set of dependent random vectors using the full-MFA sampling. Then, scale rates (τ_i) are numerically computed for individual marginals from the simulated samples of the factored variables. Lastly, the simulated random samples of the factored variables are numerically adjusted using the scale rates.

Base-MFA Sampling Algorithm:

S.1 Generate dependent random vectors using the full-MFA sampling.

1.a Draw random samples from the base estimates $\tilde{\mathbf{b}}_i = \{\tilde{b}_{i,s}\}$, ($s = 1,\ldots,S$), where S is the sample size in a simulation.

1.b Draw random samples from the AFs $\tilde{\mathbf{r}}_k = \{\tilde{r}_{k,s}\}$, ($s = 1,\ldots,S$).

1.c Generate dependent random samples: $\tilde{\mathbf{X}}_i = \{\tilde{X}_{i,s}\} = \{\tilde{R}_{i,s}\tilde{b}_{i,s}\}$, where $\tilde{R}_{i,s} = \prod_{k=1}^{K} \tilde{r}_{k,s}^{\psi_{ik}}$.

S.2 Compute the scale rates for individual marginal variables:

$$\tau_i = (V[\tilde{\mathbf{X}}_i] / V[b_i])^{0.5}$$

where $V[\tilde{\mathbf{X}}_i]$ is the variance of the factored samples and $V[b_i]$ is the variance of the base marginals.

S.3 Adjust the variance of simulated random numbers to match the base variable variances:

$$\tilde{y}_{i,s} = E[b_i] + (\tilde{X}_{i,s} - E[\tilde{\mathbf{X}}_i])/\tau_i \tag{26}$$

where $E[\tilde{\mathbf{X}}_i]$ is the mean of the factored samples and $\tilde{y}_{i,s}$ is a random number that matches the variance equality constraint: $V[\tilde{\mathbf{y}}_i] = V[b_i]$, where $\tilde{\mathbf{y}}_i = \{\tilde{y}_{i,s}\}$, ($s$ =1,…,S).

4 Credibility Test

A test project is analyzed to demonstrate the performance (i.e., accuracy, robustness, and computational efficiency) of the MFA analysis as compared against Monte Carlo simulation. The project parameters are designed in a way that challenges the primary premise underlying the MFAM (i.e., the MOM approximation). Specifically, the test settings account for three control factors: (i) asymmetricity of variables, (ii) effects of base distribution types (PERT-beta vs. triangular distribution), and (iii) effects of AF distribution types: (normal vs. lognormal).

4.1 Project settings

Consider a project with four WPs: wp_1, wp_2, wp_3, and wp_4. The base costs of the WPs (b_1, b_2, b_3, and b_4) are estimated using the three-point technique: a lower bound (o), a most likely estimate (m), and an upper bound (p). The project has identified three AFs: r_1, r_2, and r_3. The base WP costs are set uniformly as $o_i = 3$, $m_i = 4$, and $p_i = 9$ (i =1,…,4). Note that the base estimates are strongly skewed to the pessimistic side with the one-to-five ratio to the lower and upper bounds. The mean and standard deviation of the AFs are set by expert judgment as: E[r_1] = 1.1 and σ[r_1] = 0.1; E[r_2] = 1.1 and σ[r_2] = 0.2; E[r_3] = 1.1 and σ[r_3] = 0.3, which account for a uniform, on average, 10% cost growth in all WPs as well as uneven standard deviations between 0.1 and 0.3. Lastly, we adopt the allocation matrix shown in Fig. 4. As a result, possibly uneven degrees of dependence are induced across the six correlations (ρ_{12}, ρ_{13}, ρ_{14}, ρ_{23}, ρ_{24}, and ρ_{34}) between the WP costs. Consequently, four test cases are set up, as summarized in Table 1.

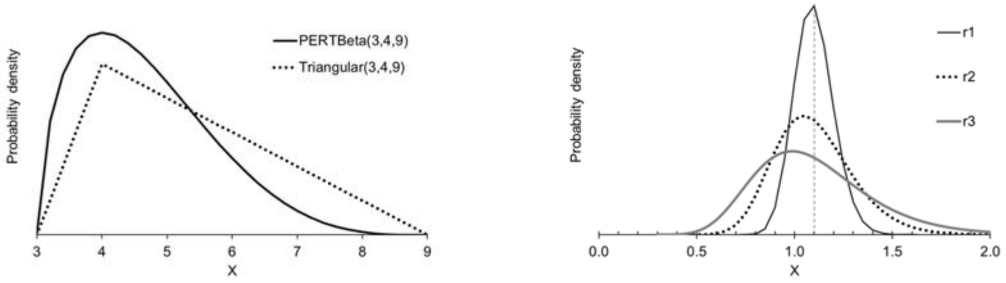
Table 1. Credibility test settings.

Cases	Base WP costs, (i =1,…,4)		Association factors, (k =1,2,3)		Allocation matrix
	Distribution types	(E[b_i], σ[b_i])	Distribution types	(E[r_k], σ[r_k])	
Case 1	b_i ~ PERTBeta	(4.667, 1.016)	r_k ~ Normal	r_1 ~ (1.1, 0.1) r_2 ~ (1.1, 0.2) r_3 ~ (1.1, 0.3)	$\mathbf{\Psi} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix}$
Case 2	b_i ~ PERTBeta	(4.667, 1.016)	r_k ~ LogNormal		
Case 3	b_i ~ Triangular	(5.333, 1.312)	r_k ~ Normal		
Case 4	b_i ~ Triangular	(5.333, 1.312)	r_k ~ LogNormal		

The two base distributions and the lognormal AF distributions used in the analysis are presented in Fig. 5. Note that the MFAM yields a Pearson's product moment correlation matrix using only the first two moments of the base and AF distributions. As a result, the mean and standard deviation of the base WP costs change according to the distribution type. When the PERTBeta distribution is used to match the three-point estimates (3,4,9) as in Case 1 and 2, E[b_i] = 4.667 and σ[b_i] = 1.016. These values need to be adjusted to E[b_i] = 5.333 and σ[b_i] = 1.312 for the Triangular WP costs in Case 3 and 4. More specifically, the beta shape parameters (α, β) matching a PERT three-point estimate (o, m, p) are solved by imposing the mean and mode equality conditions on the beta(α,β,o,p) and the PERT(o,m,p): $E[x|\text{Beta}(\alpha,\beta,o,p)] = E[x|\text{PERT}(o,m,p)]$ and $Mode[x|\text{Beta}(\alpha,\beta,o,p)] = Mode[x|\text{PERT}(o,m,p)]$.

$$o_i + (p_i - o_i)\frac{\alpha_i}{\alpha_i + \beta_i} = \frac{o_i + 4\times m_i + p_i}{6} \quad \text{and} \quad \frac{o_i(\beta_i - 1) + p_i(\alpha_i - 1)}{\alpha_i + \beta_i - 2} = m_i \tag{27}$$

For example, given (o_i,m_i,p_i)=(3,4,9), $\alpha_i = 1 + 4(m_i - o_i)/(p_i - o_i)$ =1.667 and $\beta_i = 1 +$ $4(p_i - m_i)/(p_i - o_i)$ = 4.333. Then the corresponding mean and variance of the PERTBeta are solved as 4.667 ($(o_i + 4\times m_i + p_i)/6$) and 1.016^2 ($(p_i - o_i)^2 \alpha_i \beta_i / [(\alpha_i + \beta_i + 1)(\alpha_i + \beta_i)^2]$), respectively.



(a) Base distributions

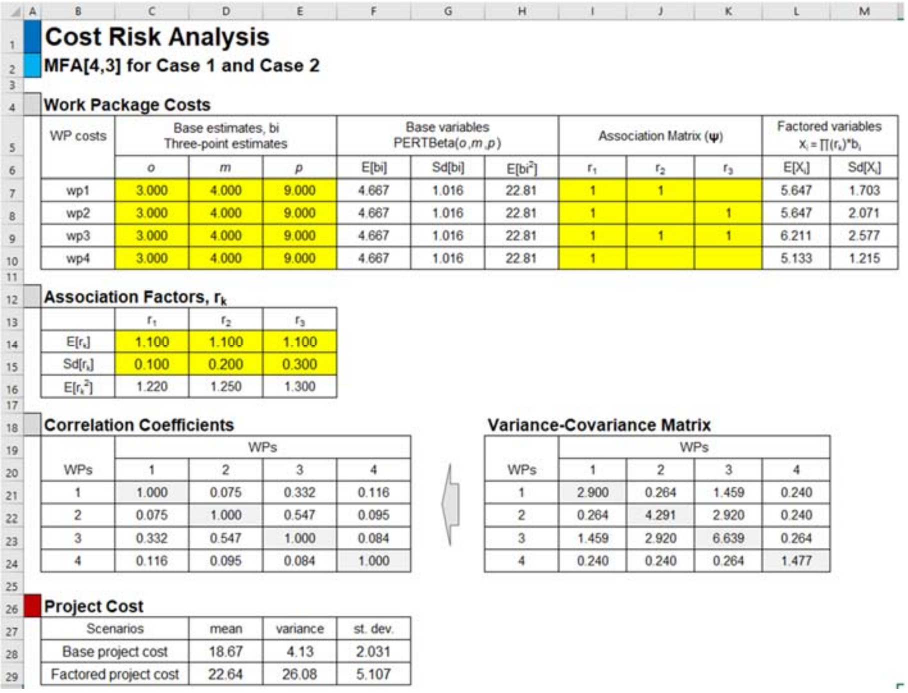
(b) AF distributions (lognormal)

Fig. 5. The base estimates and the AF distributions.

4.2 MFA analysis

Two MFA[4,3] models (for Case 1 and 2, and Case 3 and 4, respectively) are programmed in Excel spreadsheet. The PERTBeta MFA analysis for Case 1 and 2 is shown in Fig. 6. It is noteworthy that the MFA[4,3] model in Fig. 6 shows the MFA analysis in its entirety, and the MFA algorithm was fully programmed using only the basic functions in Microsoft Excel. Details of the spreadsheet formulas are presented in Appendix B. The spreadsheet MFA model illustrates the compactness and computational efficiency of the MFA approach. From a practical perspective, the most noteworthy point would be that the MFAM provides a closed-form solution to the factor-induced correlation matrix with only a marginal burden of data collection and modelling.

The MFA[4,3] analysis in **Fig. 6** also shows the summary statistics of the project cost (the sum of all the WP costs) under two scenarios: (i) the base project cost from the independent base WP costs and (ii) the factored project cost from the factored WP costs. Given a complete correlation matrix, the mean and variance of the base project cost are computed as: $\sum_{i=1}^{4} E[b_i]$ = 18.68 (the sum of Cells "F7:F10") and $\sum_{i=1}^{4} \sigma[b_i]^2$ = 4.13 (the sum of the squares of Cells "G7:G10"). In contrast, the mean and variance of the factored project cost are computed by Eq. (25) $\sum_{i=1}^{4} E[X_i]$ = 22.64 (the sum of Cells "L7:L10") and $\boldsymbol{\sigma}_{\mathbf{X}}^{T}\boldsymbol{\Sigma}_{\mathbf{r}}\boldsymbol{\sigma}_{\mathbf{X}}$ = 26.08, respectively, where $\boldsymbol{\sigma}_{\mathbf{X}}$ is a vector of the standard deviations of the factored WP costs in Cells "M7:M10". The difference between the two project costs shows the effects of the AFs on the whole project cost risk. Specifically, the results indicate that the MFA[4,3] model effectively captures uneven impacts of the individual AFs on the expected project cost and the associated uncertainty. That is, the expected project cost and the corresponding standard deviation increase by 21% (from 18.67 to 22.64) and 151% (from 2.031 to 5.107), respectively.



Cost Risk Analysis

MFA[4,3] for Case 1 and Case 2

Work Package Costs

WP costs	Base estimates, bi Three-point estimates			Base variables PERTBeta(*o*,*m*,*p*)			Association Matrix (ψ)			Factored variables $X_i = \prod(r_k)^*b_i$	
	o	*m*	*p*	E[bi]	Sd[bi]	E[bi²]	r_1	r_2	r_3	E[X_i]	Sd[X_i]
wp1	3.000	4.000	9.000	4.667	1.016	22.81	1	1		5.647	1.703
wp2	3.000	4.000	9.000	4.667	1.016	22.81	1		1	5.647	2.071
wp3	3.000	4.000	9.000	4.667	1.016	22.81	1	1	1	6.211	2.577
wp4	3.000	4.000	9.000	4.667	1.016	22.81	1			5.133	1.215

Association Factors, r_k

	r_1	r_2	r_3
E[r_k]	1.100	1.100	1.100
Sd[r_k]	0.100	0.200	0.300
E[r_k^2]	1.220	1.250	1.300

Correlation Coefficients

	WPs			
WPs	1	2	3	4
1	1.000	0.075	0.332	0.116
2	0.075	1.000	0.547	0.095
3	0.332	0.547	1.000	0.084
4	0.116	0.095	0.084	1.000

Variance-Covariance Matrix

	WPs			
WPs	1	2	3	4
1	2.900	0.264	1.459	0.240
2	0.264	4.291	2.920	0.240
3	1.459	2.920	6.639	0.264
4	0.240	0.240	0.264	1.477

Project Cost

Scenarios	mean	variance	st. dev.
Base project cost	18.67	4.13	2.031
Factored project cost	22.64	26.08	5.107

Fig. 6 MFA[4,3] model for Case 1 and Case 2.

4.3 Simulation validation

Monte Carlo simulation provides a robust solution while fully replicating the asymmetric properties imposed on the base estimates and the lognormal AFs. The MFA[4,3] analysis results were compared against simulation results by a dedicated simulation suite: @Risk from Palisade (www.palisade.com). First released in 1984, @Risk has been widely used in various business sectors and provides robust options and a generic user-interface for general-purpose simulation modelling.

4.4 Performance evaluation

The results of the four cases from the MFA and the simulation are summarized in Table 2. Note that all simulation results in the table are based on 100,000 iterations. The comparison is made in three categories: factored WP costs (X_1 through X_4), correlation coefficients ($\rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34}$), and the factored project cost. Moreover, the deviations are measured by the absolute percentage deviation of the simulation from the MFA. That is,

$$|\%\Delta| = \left|100(Sim - MFA)/MFA\right| \tag{28}$$

The results in Table 2 strongly indicate that the MFA risk assessments are substantially consistent with those from the simulation. The maximum |%Δ| of the expected value and the standard deviation of the project cost from the four cases are practically zero with 0.044% and 0.568% (Case 2), respectively. The high precision of the MFA estimates is attributed to the precision observed in the factored WP costs and the associated correlations. The maximum |%Δ| of the expected value and the standard deviation of the factored WP costs across the four cases are 0.000% (all cases) and 0.531% (Case 2 X_2), respectively. This result indicates that the MFA[4,3] model presented credible solutions to the factor-induced dependence with asymmetric, non-Gaussian variables. At the same time, the correlation coefficients show a higher order of deviations with the maximum %Δ of 7.143% (Case 1 ρ_{34}) among the 24 correlation coefficients from the four cases. Note that the corresponding absolute deviation is yet merely 0.0060 (7.143% of ρ_{34} = 0.084), which shall be considered immaterial. Nevertheless, greater than 1% deviations are observed in all correlation coefficients except ρ_{23}. In order to further examine the accuracy of the MFA correlation, we set up a supplementary experiment on the correlation coefficient distributions from the simulation and the effects of sample size on the simulation correlation assessment, of which details are presented below.

Table 2. Project cost risks. (|%Δ| in bold font is the maximum value of the four cases**.**)

Cases	Methods	Factored WP costs $E[X_i]/\sigma[X_i]$				Correlation coefficients						Factored Project Cost
		X_1	X_2	X_3	X_4	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}	E[C]/σ[C]
Case 1	MFA[4,3]	5.647 1.703	5.647 2.072	6.211 2.577	5.133 1.215	0.075	0.332	0.116	0.547	0.095	0.084	22.64 5.107
	@Risk	5.647 1.704	5.647 2.072	6.211 2.570	5.133 1.214	0.074	0.330	0.116	0.546	0.091	0.078	22.64 5.092
	|%Δ|	**0.000** 0.059	**0.000** 0.000	**0.000** 0.272	**0.000** 0.082	1.333	0.602	0.000	0.183	**4.211**	**7.143**	0.000 0.294
Case 2	MFA[4,3]	*Identical to Case 1*										
	@Risk	5.647 1.706	5.647 2.083	6.211 2.588	5.133 1.217	0.078	0.337	0.123	0.547	0.094	0.085	22.65 5.136
	|%Δ|	0.000 0.176	0.000 **0.531**	0.000 **0.427**	0.000 **0.165**	4.000	**1.506**	**6.034**	0.000	1.053	1.190	**0.044** **0.568**
Case 3	MFA[4,3]	6.453 2.088	6.453 2.490	7.099 3.068	5.867 1.545	0.066	0.298	0.097	0.499	0.081	0.073	25.87 6.035
	@Risk	6.453 2.092	6.453 2.481	7.099 3.066	5.867 1.543	0.063	0.297	0.093	0.499	0.080	0.072	25.87 6.021
	|%Δ|	0.000 **0.192**	0.000 0.361	0.000 0.065	0.000 0.129	**4.545**	0.336	4.124	0.000	1.235	1.370	0.000 0.232
Case 4	MFA[4,3]	*Identical to Case 3*										
	@Risk	6.453 2.087	6.453 2.490	7.099 3.070	5.867 1.544	0.065	0.299	0.097	0.502	0.078	0.069	25.87 6.034
	|%Δ|	0.000 0.048	0.000 0.000	0.000 0.065	0.000 0.065	1.515	0.336	0.000	**0.601**	3.704	5.479	0.000 0.017

4.5 Robustness assessment

Simulation results fluctuate over random sampling and the degree of fluctuation depends, mostly, on the size of the random samples. An extended simulation experiment was conducted to further evaluate the accuracy of MFA correlation assessment against Monte Carlo analysis. Specifically, we investigate two robustness issues: (i) How accurate is the MFA[4,3] correlation assessment as compared against simulation solutions?, and (ii) How is the MFA[4,3] correlation estimate accuracy influenced by the number of iterations in a simulation?

The factor-driven project cost simulation is programmed in the Visual Basic for Application (VBA) in Excel, which replaces @Risk. Using the VBA simulation, we carry out n_χ = 1,000 simulations of the

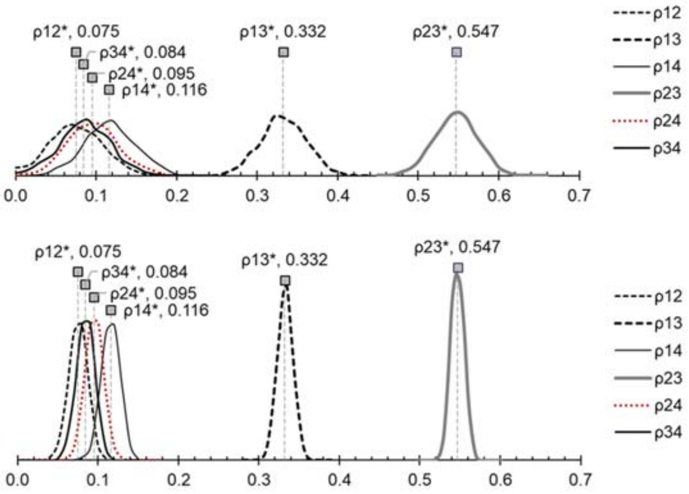
Case 2 experiment using two sets of iterations: χ = 1,000 and χ = 10,000, where χ is the number of iterations in a simulation. We introduce the outperformance indicator (ω) in order to quantify the credibility of the MFA correlation estimates as an unbiased estimate of the true means. Let $\omega(i,j|n_\chi,\chi)$ denote the probability that a simulation with χ iterations generates a correlation coefficient estimate with an error (from the aggregate mean of n_χ simulations) greater than the corresponding MFA correlation estimate error.

$$\omega(i,j\,|\,n_\chi,\chi)=\Pr\left[\left|\Delta_{n_\chi}\right|\geq\left|\Delta_{MFA}\right|\right]=\frac{1}{n_\chi}\sum_{s=1}^{n_\chi}\left[\mathbf{I}\left(\left|\rho_s-\overline{\rho}_{n_\chi}\right|\geq\left|\rho_{MFA}-\overline{\rho}_{n_\chi}\right|\right)\right] \qquad (29)$$

where ρ_s is the correlation coefficient from the s-th simulation of a total n_χ simulations; $\overline{\rho}_{n_\chi}$ is the mean of correlation coefficients from n_χ simulations; ρ_{MFA} is the correlation coefficient from the MFA model; Δ_{n_χ} is the deviation of the s-th simulated correlation coefficient from $\overline{\rho}_{n_\chi}$; Δ_{MFA} is the deviation of the MFA correlation coefficient from $\overline{\rho}_{n_\chi}$; and **I** is the indicator function that returns 1 if its argument is true, and 0 otherwise. Note that for clarity, correlation indices (i, j) are dropped in Eq. (29). Simply, $\omega(i,j|n_\chi,\chi)$ measures the probability that the MFA provides a correlation coefficient estimate with a less deviation from the mean estimate of n_χ simulations with a chosen number of χ.

From the extended simulations, distributions of the six correlation coefficients were constructed as presented in Fig. 7. The results reveal two advantages that the MFA model may offer in practice. First, the precision of the factor-driven correlations obtained using simulation depends heavily on the sample size in each simulation. The correlation coefficient distributions in Fig. 7 show that the range of possible correlation coefficients can be significantly larger with χ = 1,000 compared against with χ = 10,000. For instance, the 90% confidence interval of ρ_{23} is reduced by nearly 66% to [0.53, 0.56] with χ = 10,000 from [0.50, 0.59] with χ = 1,000. Note that even with the increasing computational power available in practice, the sample size still can be a decisive factor for efficient random sampling as the number of random variables increases. Second, the MFA correlation estimates provide practically unbiased (with respect to repeated simulations) estimates of the aggregate means of the correlation coefficients.

Table 3 outlines the summary statistics of three performance measures: the maximum simulation deviation ($\text{Max}\,|\,\Delta_{n_\chi}\,|$), the MFA deviation ($|\Delta_{MFA}|$), and the outperformance indicator ($\omega(i,j|n_\chi,\chi)$). The results on the six correlation coefficients consistently suggest that the MFA correlations are practically unbiased estimates of the aggregate means from the 1,000 simulations regardless of the sample sizes. With χ = 10,000, the maximum value of $|\Delta_{MFA}|$ is 0.0009 in ρ_{13}. In contrast, the $\text{Max}\,|\,\Delta_{n_\chi}\,|$ values indicate that the simulation fluctuation heavily depends on the sample size. The $\omega(i,j|n_\chi,\chi)$ results show that the MFA correlation estimates outperform, with the minimum chance of 92.6% (ρ_{34}), individual simulation estimates regardless of the sample sizes.



(a) χ = 1,000

(b) χ = 10,000

Fig. 7 Effects of the sample size (χ) on simulated correlation.
(n_χ = 1,000 ; ρ* values are from the MFA[4,3].)

Table 3. Credibility of MFA correlation assessment with respect to 1,000 simulations.

Sample size (χ)	Measures	ρ_{12}	ρ_{13}	ρ_{14}	ρ_{23}	ρ_{24}	ρ_{34}
χ = 1,000	$\bar{\rho}_{n_\chi}$	0.0745	0.3324	0.1170	0.5467	0.0960	0.0864
	Max $\vert\Delta_{n_\chi}\vert$	0.1020	0.1139	0.1135	0.0938	0.1205	**0.1240**
	$\vert\Delta_{MFA}\vert$	0.0005	0.0004	0.0010	0.0003	0.0010	**0.0024**
	$\omega(i,j\vert n_\chi,\chi)$	0.9880	0.9920	0.9670	0.9890	0.9760	**0.9260**
χ = 10,000	$\bar{\rho}_{n_\chi}$	0.0750	0.3329	0.1155	0.5467	0.0953	0.0841
	Max $\vert\Delta_{n_\chi}\vert$	0.0348	0.0298	**0.0379**	0.0260	0.0325	0.0302
	$\vert\Delta_{MFA}\vert$	0.0000	**0.0009**	0.0005	0.0003	0.0003	0.0001
	$\omega(i,j\vert n_\chi,\chi)$	0.9990	0.9290	0.9550	0.9790	0.9870	0.9820

5 Applications

Empirical data from previous projects or general experiences provide valuable, although often incomplete, dependence information relevant to a new project (Ranasinghe 2000; Touran and Wiser 1992; Wang and Huang 2000). It would be sensible then to make full use of all relevant information whenever possible. The MFA model's closed-form correlation matrix provides a robust solution to the problem of utilizing incomplete dependence information for coherent risk analysis. This section presents three illustrative examples of minimum information dependence modelling in typical project settings. First, a WBS-driven dependence model for project cost risk is presented for more general project settings. Then, effects of large-scale dependence on the amount of contingency funds are examined. Lastly, the closed-form MFA correlation is parameterized and calibrated using an optimization technique for a partially specified correlation matrix.

5.1 WBS-SA project cost analysis

Consider a WBS of a generic building project, as in Fig. 8. The project consists of twelve work packages (*wp*$_1$ through *wp*$_{12}$) on four levels, starting with the project (*r*$_1$) on Level-1. Level-2 includes three work items: *Site Works & Foundation* (r_2), *Structures* (r_3), and *Interior & Finishing* (r_4). The WP costs are estimated in terms of the three-point estimate, (o_i,m_i,p_i) = (3, 4.5, 9), which are considered 'base estimates', (b_1,… b_{12}). The triangular distribution is used for the WP costs and their means and variances are computed as $E[b_i]=(o_i+m_i+p_i)/3$ = 5.5 and $V[b_i]=(o_i^2+m_i^2+p_i^2-o_im_i-m_ip_i-p_io_i)/18$ =1.625.

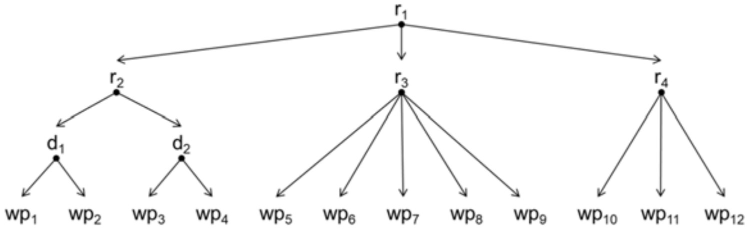


Fig. 8 A WBS of a building project with hierarchical AFs. d_1 = Site-works; and d_2 = Foundation. Note that d_1 and d_2 are dummy factors that exist in the WBS but are not used as an AF.

Dependence elicitation:

For illustrative purposes, we consider a situation where the relative degree of association between the WPs are represented with four AFs: One global association (GA) factor at the project level (GA1 for r_1) and three local association (LA) factors at the intermediate levels (LA1, LA2, and LA3 for r_2, r_3, and r_4, respectively). The local factors represent potentially different degrees of association due to distinct properties of intermediate work items. For instance, WPs under *Site Works & Foundation* in a construction project are commonly influenced by weather and site conditions, whereas actual costs of the *Structures* WPs tend to be influenced by material types (e.g., concrete or steel) and their market rates. In addition, the GA1 at the project level accounts for the effect of soft factors of the overall project performance (van Dorp 2020; Williams 2003). Then the allocation matrix is constructed as:

$$\mathbf{\Psi}^T=\left[\psi_{ik}\right]^T_{d\times K}=\begin{bmatrix}1&1&1&1&1&1&1&1&1&1&1&1\\1&1&1&1&0&0&0&0&0&0&0&0\\0&0&0&0&1&1&1&1&1&0&0&0\\0&0&0&0&0&0&0&0&0&1&1&1\end{bmatrix} \tag{30}$$

AF parameter estimates:

Given the AFs and the corresponding allocation matrix, the dependence modelling problem becomes a task of estimating the AF parameters. Let $\{r_k\}$ (k =1,…,4) denote the four AFs. First, likely impacts of the AFs on WP costs are assessed. A company with accumulated historical data in house may use such data to extract the means and variances of individual AFs (Trietsch et al. 2012). In more general cases, the AF can be estimated by expert judgment. In this example, we consider a combination of high, medium, and low impacts in terms of the relative coefficients of variation (COVs) of the AFs compared to the COV of the base estimates, 0.23.

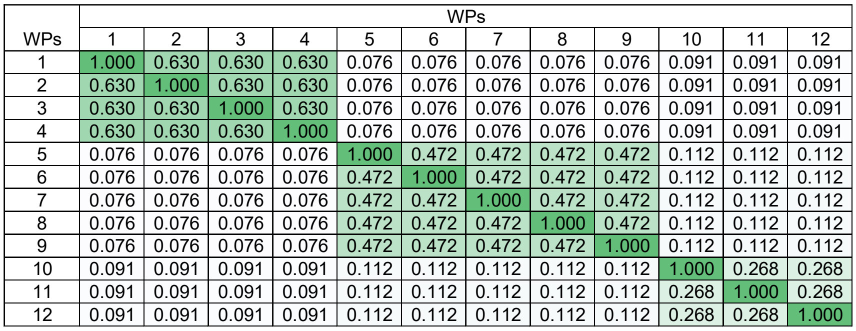
- $r_1 \sim E[r_1]$= 1.0 and $\sigma[r_1]$ = 0.1 for GA1 (low association)
- $r_2 \sim E[r_2]$= 1.0 and $\sigma[r_2]$ = 0.3 for LA1 (high association)

- $r_3 \sim E[r_3]$= 1.0 and $\sigma[r_3] = 0.2$ for LA2 (medium association)
- $r_4 \sim E[r_4]$= 1.0 and $\sigma[r_4] = 0.1$ for LA3 (low association)

Here $E[r_k]$= 1.0 ($k = 1,\ldots,4$) represents a situation where there is no systematic cost growth.

Correlation matrix:

Given the three input elements $\langle \mathbf{b}; \mathbf{r}, \mathbf{\Psi} \rangle$ of the MFA[12,4], a complete correlation coefficient matrix between the WP costs is computed using the MFA Algorithm (Section 3.3). The resulting correlation matrix is presented in Fig. 9.



	WPs											
WPs	1	2	3	4	5	6	7	8	9	10	11	12
1	1.000	0.630	0.630	0.630	0.076	0.076	0.076	0.076	0.076	0.091	0.091	0.091
2	0.630	1.000	0.630	0.630	0.076	0.076	0.076	0.076	0.076	0.091	0.091	0.091
3	0.630	0.630	1.000	0.630	0.076	0.076	0.076	0.076	0.076	0.091	0.091	0.091
4	0.630	0.630	0.630	1.000	0.076	0.076	0.076	0.076	0.076	0.091	0.091	0.091
5	0.076	0.076	0.076	0.076	1.000	0.472	0.472	0.472	0.472	0.112	0.112	0.112
6	0.076	0.076	0.076	0.076	0.472	1.000	0.472	0.472	0.472	0.112	0.112	0.112
7	0.076	0.076	0.076	0.076	0.472	0.472	1.000	0.472	0.472	0.112	0.112	0.112
8	0.076	0.076	0.076	0.076	0.472	0.472	0.472	1.000	0.472	0.112	0.112	0.112
9	0.076	0.076	0.076	0.076	0.472	0.472	0.472	0.472	1.000	0.112	0.112	0.112
10	0.091	0.091	0.091	0.091	0.112	0.112	0.112	0.112	0.112	1.000	0.268	0.268
11	0.091	0.091	0.091	0.091	0.112	0.112	0.112	0.112	0.112	0.268	1.000	0.268
12	0.091	0.091	0.091	0.091	0.112	0.112	0.112	0.112	0.112	0.268	0.268	1.000

Fig. 9 MFA-induced correlation matrix.

Results:

Now the independence assumption can be dropped and a more realistic cost risk assessment becomes attainable. Specifically, we compare three scenarios of project cost risks.

- Independence (IND) Scenario. The project cost (C_{Ind}) of independent WP costs is generated using the base costs (**b**). That is, $E[C_{Ind}] = \sum_{i=1}^{12} E[b_i] = 66.0$ and $V[C_{Ind}] = \sum_{i=1}^{12} V[b_i] = 19.5$.
- Base-MFA (B-MFA). The base project cost ($C_{\mathbf{b}}$) is computed using the base costs (**b**), and the MFA correlation ($\mathbf{\Sigma_r}$) in Fig. 9. From Eq. (24), $E[C_{\mathbf{b}}] = \sum_{i=1}^{12} E[b_i] = 66.0$ and $V[C_{\mathbf{b}}] = \boldsymbol{\sigma}_{\mathbf{b}}^T \mathbf{\Sigma}_{\mathbf{r}} \boldsymbol{\sigma}_{\mathbf{b}} = 63.7$.
- Full-MFA (F-MFA). The factored project cost (C_X) is computed using the factored costs (**X**) and the $\mathbf{\Sigma_r}$. From Eq. (25), $E[C_{\mathbf{X}}] = \sum_{i=1}^{12} E[X_i] = 66.0$ and $V[C_{\mathbf{X}}] = \boldsymbol{\sigma}_{\mathbf{X}}^T \mathbf{\Sigma}_{\mathbf{r}} \boldsymbol{\sigma}_{\mathbf{X}} = 141.5$.

The resulting project cost risk profiles are shown in Fig. 10, which illustrate the effects of the MFA-induced dependence on the project cost risk. Note that the three analytic risk profiles (IND, Base-MFA, and Full-MFA) are fitted using the lognormal distribution. Fig. 10 also includes the cost risk profile obtained using a @Risk simulation model (with 50,000 iterations) for the Full-MFA. The graphs indicate overall a good match between the MFA and the simulation. The results demonstrate that the MFA approach provides flexibility in dealing with the potential effects of different dependence scenarios in project risk assessment. Specifically, the standard deviation of the project cost increases by 80% in Base-MFA ($\sigma[C_{\mathbf{b}}] = 7.98$) from IND ($\sigma[C_{Ind}] = 4.42$), and by 169% in Full-MFA ($\sigma[C_{\mathbf{X}}] = 11.90$).

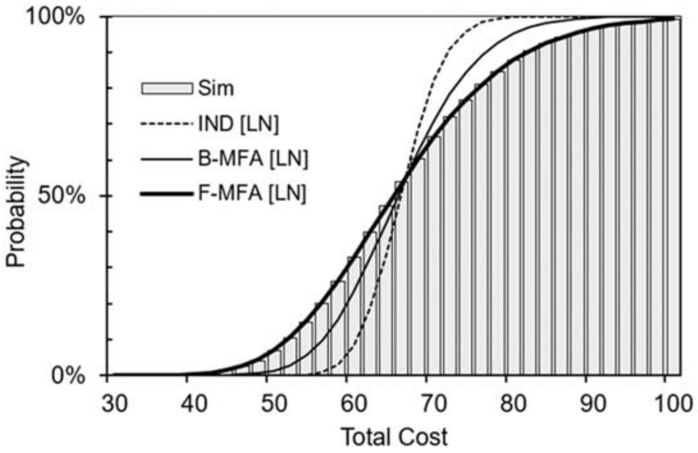


Fig. 10 Project cost risk profiles.

5.2 Impact of dependence and project size

This section investigates the robustness of MFA in large-scale project settings. We examine the effects of dependence on the amount of contingency reserve, which is added to the expected project cost in order to establish a risk-adjusted budget at a chosen confidence level. We also evaluate possible effects of project size on the contingency underestimation due to the independence assumption using four benchmark projects with varying scales. The four projects are generated by incrementally duplicating the MFA[12,4] model in Fig.8 to larger projects with 24-WPs, 48-WPs, 120-WPs, and lastly 1,200-WPs. Note that the four risk factors in the original MFA[12,4] are preserved in the scaled-up projects so that pairwise correlations between WPs remain the same. In each project, we estimate the contingency amounts under two conditions: the independence (IND) contingency and the full-MFA contingency. Note that adding contingency reserve to expected project cost or time is a widely endorsed practice for risk-driven project planning and control (GAO 2015, p.113; GAO 2020, p.158; PMI 2013).

The results are summarized in Table 4. Given the mean and standard deviation of a project cost (either IND or F-MFA), the risk-adjusted budget at the 90 percent confidence level (P-90 budget) is estimated from a lognormal distribution. Then, the contingency reserve (CR) is determined as the gap between the P-90 value and the expected project cost ($E[C]$). Note also that we directly measure the effects of project size on the factor-induced CR in terms of the percent underestimation (%U) of the P-90 contingency under the independent assumption (CR-90_{IND}) with respect to the corresponding P-90 contingency with risk factors (CR-$90_{\text{F-MFA}}$). That is, %U = (CR-$90_{\text{F-MFA}}$ – CR-90_{IND})/CR-$90_{\text{F-MFA}}$. Moreover, the accuracy of the MFAM for the 1,200-WPs project is compared against the simulation results by @Risk. From the results, two important observations can be made regarding the practical implications of the MFA approaches in large-scale contingency assessment.

Table 4. Effects of project size on the risk-adjusted contingency reserve.

Cases		Project Cost, C			Contingency at P-90 (CR-90)	
		$E[C]$	$\sqrt{V[C]}$	P-90	CR-90	%U
12 WPs	IND	66.0	4.42	71.7	5.74	63.4%
	F-MFA[12,4]	66.0	11.90	81.7	15.68	-
24 WPs	IND	132.0	6.24	140.1	8.09	73.2%
	F-MFA[24,4]	132.0	22.9	162.2	30.18	-
48 WPs	IND	264.0	8.83	275.4	11.41	80.7%
	F-MFA[48,4]	264.0	44.9	323.2	59.15	-
120 WPs	IND	660.0	13.96	678.0	17.99	87.7%
	F-MFA[120,4]	660.0	110.9	806.0	146.02	-
1,200 WPs	IND	6600.0	44.16	6656.7	56.69	96.1%
	F-MFA[1200,4]	6600.0	1100.3	8049.0	1449.01	-
	Sim	6599.5	1094.7	8045.8	1446.33	-

Note: CR is the contingency reserve added to the expected project cost; %U represent the percent underestimation of the IND CR from the F-MFA CR. The simulation results are from 50,000 iterations.

First, the contingency underestimation error due to the independence assumption increases as the project size increases. The %U values indicate that given the risk factors remain unchanged, the CR is underestimated and the degree of underestimation increases as the project size increases, up to the excessive level of the 96.1% in the 1,200-WPs project. In other words, the cost risk of the 1,200-WPs is dominated by the factor-induced dependence rather than the marginal variations. Second, the comparison of the MFA results and the simulation (the 1,200-WPs project) indicates that the accuracy of the MFAM is reasonably preserved regardless of the project sizes. It should also be noted that due to the analytical (i.e., non-simulation and algebraic) nature of the MFA analysis, additional computational or economical burdens for large-scale applications are practically negligible as compared against the full-scale simulation.

5.3 Partially specified correlation calibration

The burden of dependence elicitation and data collection increases as the number of uncertain variables increases. (For example, the 1,200-WPs project requires 719,400 pairwise correlation assessments.) A sensible approach to the large-scale dependence specification problem is to use all relevant information in various formats, rather than conveniently adopting the independence assumption or unrealistically restrictive assumptions (e.g., single factor approaches). In construction, for instance, some activities (e.g., earthworks and foundations) are more common in general projects and it would be easier to collect standardized performance data for statistical correlation assessment. The closed-form MFA correlation allows the decision maker to infer the entire correlation coefficients in accordance with a limited number of pairwise correlation coefficients known in a project. In this context, the partially specified correlation problem can be formulated as a two-step MFA calibration problem: (i) identify driving parameters in the MFA model, which can be used to fully calibrate the MFA inputs and (ii) determine the optimal values of the calibration parameters in a way that satisfies all relevant constraints and observed data.

Step 1. Identify calibration parameters

Consider a situation where only four of the ${}_{12}C_2$ correlation coefficients of the 12-WP project cost problem, for instance, $\rho_{i,j}^{\#}$, $(i,j) \in \{(1,2),(2,4),(3,5),(4,7)\}$, are elicited from relevant sources. Then a set of four calibration parameters can be identified as follows: $s_1 = E[r_k]/\alpha_k$ (k=1,2), $s_2 = E[r_k]/\alpha_k$ (k=3,4), $s_3 = \sigma[r_k]/\beta_k$ (k=1,2), and $s_4 = \sigma[r_k]/\beta_k$ (k=3,4), where $\alpha_k \in \{1,1,1,1\}$ and $\beta_k \in \{0.1,0.3,0.2,0.1\}$. The s_1 (s_2) represents a constant scale for the means of r_1 and r_2 (r_3 and r_4), whereas s_3 (s_4) represents a constant scale for the standard deviations of r_1 and r_2 (r_3 and r_4).

Step 2. Determine the optimal values of the calibration parameters

Then the optimization problem is set up as:

- Minimize: $$\phi\left[\mathbf{\Sigma}^{\#}(\rho_{i,j}^{\#}) - \mathbf{\Sigma}_{MFA}^{\#}(\mathbf{b};\mathbf{r}(s_1,s_2,s_3,s_4),\mathbf{\Psi})\right] \tag{31}$$
- Subject to: $s_1 > 0;\ s_2 > 0;\ s_3 > 0;\ s_4 > 0$

where the objective function $\phi[\bullet]$ measures the distance between the partially specified correlation matrix, $\mathbf{\Sigma}^{\#}$, and the corresponding elements of the parameterized MFA[12,4] correlation matrix, $\mathbf{\Sigma}_{MFA}^{\#}$. We adopt for the distance measure the sum of the squared distances of the specified correlation coefficients between two correlation matrices. Then, the objective function is established as

$$\phi\left[s_1,s_2,s_3,s_4\right] = \sum\nolimits_{\{(i,j)\}} \left\{\rho_{i,j}^{\#} - \tilde{\rho}_{i,j}(\mathbf{b};\mathbf{r}(s_1,s_2,s_3,s_4),\mathbf{\Psi})\right\}^2 \tag{32}$$

where $\tilde{\rho}_{i,j}$ is a correlation coefficient from the parameterized MFA[12,4] and $(i,j) \in \{(1,2),(2,4),(3,5),(4,7)\}$ as identified above.

The calibration problem was solved for four cases: Case I for overall low correlations, Case II for moderate correlations, and Case III and IV for mixed correlations. The solutions from using the Solver in Microsoft Excel are summarized in Table 5. The results show that optimal parameters with a perfect match can be found in three cases: Case I, II, and IV. Obviously, a perfect calibration is not guaranteed, as in Case III. At the same time, the results demonstrate the use of the MFA model as a means of effective analytic inference for missing dependence information.

Table 5. Dependence calibration for partially specified correlation matrix.

Cases	Known correlations $(\rho_{12},\rho_{24},\rho_{35},\rho_{47})$	Optimal parameters (s_1,s_2,s_3,s_4)	Calibrated correlations $(\tilde{\rho}_{12},\tilde{\rho}_{24},\tilde{\rho}_{35},\tilde{\rho}_{47})$	Evaluation $\phi\left[s_1,s_2,s_3,s_4\right]$
I	(0.3, 0.3, 0.3, 0.3)	(1.34, 1.28, 0.56, 0.80)	(0.3, 0.3, 0.3, 0.3)	0.0
II	(0.6, 0.6, 0.6, 0.6)	(1.10, 0.91, 0.88, 1.09)	(0.6, 0.6, 0.6, 0.6)	0.0
III	(0.6, 0.3, 0.6, 0.6)	(1.92, 1.24, 1.24, 1.51)	(0.5, 0.5, 0.6, 0.5)	600.0
IV	(0.6, 0.6, 0.3, 0.6)	(1.04, 1.53, 0.83, 0.80)	(0.6, 0.6, 0.3, 0.6)	0.0

6 Conclusions

Proper dependence consideration is crucial for realistic project risk assessment and making informed decisions under uncertainty. We present an analytic framework that combines a systematic way of accounting for multiple risk factors (the SA) and a quantitative dependence model based on the second moment approach (the MFAM). The SA-MFAM offers an analytic, closed-form, and computationally tractable alternative to the correlation-driven approaches for dependence modelling. Specifically, the SA-MFAM (i) adopts a multi-factor approach that enhances the realism of project risk modelling with the flexibility of accounting for various association factors observed in individual projects (Section 2 and Section 4); (ii) provides an analytic, closed-form solution to the correlation matrix generation problem, which can be further parameterized and calibrated to comply with project-specific data availability (Section 3 and Section 5); and (iii) provides a computationally efficient and tractable algorithm that guarantees a mathematically consistent correlation matrix, while preserving its scalability to high-dimensional risk problems.

Numerical applications demonstrate, in conjunction with more general analyses (e.g., optimization and calibration), the robustness of the SA-MFAM approaches in real project settings, emphasizing its scalability under limited data availability. Implementing the SA-MFAM in a project can be straightforward yet adaptive. The three input elements $\langle \mathbf{b}; \mathbf{r}, \mathbf{\Psi} \rangle$ can be readily obtained from standard practices and project documents (e.g., WBS, resource plans, or risk register). Once a complete and mathematically consistent correlation matrix is constructed, the MFAM offers extended flexibility of integrating the correlation with the base (factor-free) marginals or the factored marginals, which further enhances the realism of dealing with marginals decoupled from the risk factors.

The SA-MFAM approach would offer several advantages in the practice and research of project risk assessment. The computational experiments and illustrative applications show that the SA-MFAM is fairly robust, under mild assumptions. It is reasonably unbiased (with respect to repeated simulations), tractable, and readily scalable to high-dimensional settings while leveraging limited information and preserving the accuracy comparable to full-scale simulation. In practice, large-scale projects with thousands of activities are becoming increasingly common. It would be worth noting that risk-driven project planning is a repetitive and computationally intensive process of identifying various sources of uncertainties and prioritizing their impacts on project outcomes under diverse scenarios. With the analytical robustness and an affordable computational burden, the SA-MFAM would be able to enhance the realism for risk-driven project planning and control under uncertainty. The SA-MFAM can also be an efficient tool for project risk studies, providing deeper insights into the nature and implications of factor-induced dependence. For instance, the large-scale experiments in Section 5.2 indicate that the degree of contingency underestimation due to the independence assumption depends not only on the risk factors but also on the project size, to the extent that can be grossly misleading (e.g., the contingency underestimation of 96.1% in the 1,200 WP project).

Quantitative risk modelling has been attracting increasing attention in the project management literature. There are several research opportunities worthy of further investigation. First, our model assumes risk factors to be represented with their means and variances. This second-moment assumption can be extended to account for factors in more sophisticated formats, for example, risk events with a probability of occurring (Hulett 2011). Second, the nature and characteristics of the trade-offs between the parameter reduction and the resulting loss of accuracy need to be further investigated. Another, possibly more urgent, research agenda is to study the implications of factor-driven dependence under various decision-making settings in project control. Extensive simulation experiments with controlled parameters would be a viable approach to pursue this line of research.

Appendix A. The second moment of the product of random variables

Consider a product of two ($K = 2$) independent random variables, $R_2 = r_1r_2$. The mean and variance of the individual random variables are given as $\{E[r_k]\}$ and $\{V[r_k]\}$, $k = 1,2$, respectively. Then the mean and variance of R_2 are determined as

$$E[R_2] = \iint_{R_2} r_1 r_2 f_{r_1 r_2}(r_1, r_2)dr_1 dr_2 = \int_{r_1} r_1 f_{r_1}(r_1)dr_1 \int_{r_2} r_2 f_{r_2}(r_2)dr_2 = E[r_1]E[r_2] \tag{A.1}$$

$$V[R_2] = E[(r_1r_2)^2] - E[r_1r_2]^2 = E[r_1^{\ 2}]E[r_2^{\ 2}] - E[r_1]^2 E[r_2]^2 \tag{A.2}$$

where $f_{r_1r_2}(r_1, r_2) = f_{r_1}(r_1)f_{r_2}(r_2)$ holds because r_1 and r_2 are independent and $E[(r_1r_2)^2] = E[r_1^{\ 2}]E[r_1^{\ 2}]$ holds because r_1^2 and r_2^2 are also independent.

Consider a product of K (≥ 3) independent random variables, $R_K = r_1 \cdots r_{K-1} r_K$. From $R_K = R_{K-1} r_K$ and Eqs. (A.1) and (A.2), the mean and variance of R_K are

$$E[R_K] = E[R_{K-1}]E[r_K] \tag{A.4}$$

$$V[R_K] = E[R_{K-1}^{\ 2}]E[r_K^{\ 2}] - E[R_{K-1}]^2 E[r_K]^2 \tag{A.5}$$

where R_{K-1} and r_K are also independent. Since $E[R_{K-1}] = E[r_1 \cdots r_{K-1}] = E[r_1] \cdots E[r_{K-1}]$ and $E[R_{K-1}^{\ 2}] = E[r_1^{\ 2} \cdots r_{K-1}^{\ 2}] = E[r_1^{\ 2}] \cdots E[r_{K-1}^{\ 2}]$, the mean and variance of a project of K (≥ 3) independent random variables are

$$E[R_K] = \prod_{k=1}^{K} E[r_k] \tag{A.6}$$

$$V[R_K] = \prod_{k=1}^{K} E[r_k^{\ 2}] - \prod_{k=1}^{K} E[r_k]^2 = \prod_{k=1}^{K} \left(V[r_k] + E[r_k]^2\right) - \prod_{k=1}^{K} E[r_k]^2 \tag{A.7}$$

End of the proof of Eqs. (8) and (9). ■

Appendix B. Spreadsheet model of MFA[4,3]

Once an analyst enters the three input elements of MFA $\langle \mathbf{b}; \mathbf{r}, \mathbf{\Psi} \rangle$, $(\{b_i\} \sim PERT(o_i, m_i, p_i), i = 1,...,4$ in Cells “C7:E10”, $\{r_k\} \sim (E[r_k], \sigma[r_k])$, k=1, 2, 3 in Cells “C14:E15”, and $\mathbf{\Psi}$ in Cells “I7:K10”, the resulting correlation matrix ($\mathbf{\Sigma_r}$) in Cells “C21:F24” is computed by the MFA algorithm (Section 3.3). For instance, the correlation between wp_2 and wp_3 in Cell “E22” is:

$$\mathbf{\Sigma_r}(2,3) = \frac{V(R_{2\cap 3})E(R_{2-3})E(R_{3-2})(\mathbf{E}_b)_2(\mathbf{E}_b)_3}{[(\mathbf{V}_X)_2 \times (\mathbf{V}_X)_3]^{0.5}} = \frac{(0.1219)(1)(1.1)(4.667)(4.667)}{(2.072)(2.577)} = 0.547$$

Each element of the covariance term is computed using the Hadamard computation as follows.

- $V(R_{2\cap 3}): \{= \text{RODUCT}(\mathbf{E}_{rs}\text{\^{}}(\text{INDEX}(\boldsymbol{\Psi},2,0) * \text{INDEX}(\boldsymbol{\Psi},3,0))) -$

  $\text{PRODUCT}(\mathbf{E}_{r}\text{\^{}}(\text{INDEX}(\boldsymbol{\Psi},2,0) * \text{INDEX}(\boldsymbol{\Psi},3,0)))\text{\^{}}\,2\}$ (MFAA §4.1)

- $E(R_{2-3}): \{= \text{PRODUCT}(\mathbf{E}_{r}\text{\^{}}(\text{INDEX}(\boldsymbol{\Psi},2,0) - \text{INDEX}(\boldsymbol{\Psi},2,0) * \text{INDEX}(\boldsymbol{\Psi},3,0)))\}$ (MFAA §4.2)

- $E(R_{3-2}): \{= \text{PRODUCT}(\mathbf{E}_{r}\text{\^{}}(\text{INDEX}(\boldsymbol{\Psi},3,0) - \text{INDEX}(\boldsymbol{\Psi},2,0) * \text{INDEX}(\boldsymbol{\Psi},3,0)))\}$ (MFAA §4.3)

- $(\mathbf{E}_b)_2: \; = \text{INDEX}(\mathbf{E}_b,2,0)\,; \quad (\mathbf{E}_b)_3: \; = \text{INDEX}(\mathbf{E}_b,3,0)$ (MFAA §1.1)

- $(\mathbf{V}_X)_2: \{= (\mathbf{E}_{bs})_2 * \text{PRODUCT}(\mathbf{E}_{rs}\text{\^{}}\text{INDEX}(\boldsymbol{\Psi},2,0)) - (\mathbf{E}_b)_2^2 * \text{PRODUCT}(\mathbf{E}_{r}\text{\^{}}(2 * \text{INDEX}(\boldsymbol{\Psi},2,0)))\}$

- $(\mathbf{V}_X)_3: \{= (\mathbf{E}_{bs})_3 * \text{PRODUCT}(\mathbf{E}_{rs}\text{\^{}}\text{INDEX}(\boldsymbol{\Psi},3,0)) - (\mathbf{E}_b)_3^2 * \text{PRODUCT}(\mathbf{E}_{r}\text{\^{}}(2 * \text{INDEX}(\boldsymbol{\Psi},3,0)))\}$

  (MFAA §1.1)

where $\mathbf{E}_{rs}$ = Cells "C16:E16", $\mathbf{E}_{r}$ = Cells "C14:E14", $\mathbf{E}_{b}$ = Cells "F7:F10", $\mathbf{E}_{bs}$ = Cells "H7:H10". ■

References

Asadabadi, M. R., and Zwikael, O. (2021). "Integrating Risk into Estimations of Project Activities' Time and Cost: A Stratified Approach." *European Journal of Operational Research*. 291(2), 482-490

Baccarini, D. (1996). "The concept of project complexity—a review." *International journal of project management*, 14(4), 201-204.

Bedford, T., and Cooke, R. M. (2002). "Vines: A new graphical model for dependent random variables." *Annals of Statistics*, 30(4), 1031-1068.

Bolstad, W. M. (2007). *Introduction to Bayesian Statistics*, John Wiley & Sons, Inc., Hoboken, New Jersey.

Cario, M. C., and Nelson, B. L. (1997). "Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix." Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

Chau, K. W. (1995). "Monte Carlo simulation of construction costs using subjective data." *Construction Management & Economics*, 13(5), 369.

Cho, S. (2009). "A linear Bayesian stochastic approximation to update project duration estimates." *European Journal of Operational Research*, 196(2), 585-593.

Clemen, R. T., Fischer, G. W., and Winkler, R. L. (2000). "Assessing dependence: Some experimental results." *Management Science*, 46(8), 1100-1115.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, Routledge.

Demirtas, H., and Hedeker, D. (2011). "A practical way for computing approximate lower and upper correlation bounds." *The American Statistician*, 65(2), 104-109.

Fallat, S. M., and Johnson, C. R. (2007). "Hadamard powers and totally positive matrices." *Linear Algebra and its Applications*, 423(2-3), 420-427.

Flyvbjerg, B. (2006). "From Nobel prize to project management: Getting risks right." *Project Management Journal*, 37(3), 5-15.

GAO (2015). "GAO Schedule assessment guide: Best practices for project schedules." *Applied Research and Methods*, U.S. Government Accountability Office (GAO), Washington, DC, 240.

GAO (2020). "GAO Cost estimating and assessment guide: Best practices for developing and managing capital program costs." *Applied Research and Methods*, U.S. Government Accountability Office (GAO), Washington, DC.

Garvey, P. R., Book, S. A., and Covert, R. P. (2016). *Probability methods for cost uncertainty analysis: A systems engineering perspective*, CRC Press.

Ghosh, S., and Henderson, S. G. (2003). "Behavior of the NORTA method for correlated random vector generation as the dimension increases." *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(3), 276-294.

Goh, J., and Sim, M. (2011). "Robust optimization made easy with ROME." *Operations Research*, 59(4), 973-985.

Hertz, D. B. (1964). "Risk analysis in capital investment." *Harvard Business Review*, 42(1), 95-106.

Hulett, D. T. (2011). *Integrated cost-schedule risk analysis*, Gower Publishing, Ltd.

Kim, B.-C. (2015). "Integrating risk assessment and actual performance for probabilistic project cost forecasting: A second moment Bayesian model." *IEEE Transactions on Engineering Management*, 62(2), 158-170.

Kim, B.-C. (2021). "Dependence Modeling for Large-scale Project Cost and Time Risk Assessment: Additive Risk Factor Approaches." *IEEE Transactions on Engineering Management*, Early Access: https://ieeexplore.ieee.org/abstract/document/9332228.

Kim, B.-C., and Kwak, Y. H. (2018). "Improving the accuracy and operational predictability of project cost forecasts: an adaptive combination approach." *Production Planning & Control*, 29(9), 743-760.

Kim, B.-C. P., Jeffrey K. (2019). "What CPI = 0.85 really means: A probabilistic extension of the estimate at completion." *ASCE Journal of Management in Engineering*, 35(2), 1-18.

Kurowicka, D., and Cooke, R. M. (2006). *Uncertainty analysis with high dimensional dependence modelling*, John Wiley & Sons.

Love, P., Wang, X., Sing, C., and Tiong, R. (2013). "Determining the probability of project cost overruns." *Journal of Construction Engineering and Management*, 139(3), 321-330.

Love, P., Sing, C., Wang, X., Edwards, D. J., and Odeyinka, H. (2013). "Probability distribution fitting of schedule overruns in construction projects." *Journal of the Operational Research Society*, 64(8), 1231-1247.

Lurie, P. M., and Goldberg, M. S. (1998). "An approximate method for sampling correlated random variables from partially-specified distributions." *Management science*, 44(2), 203-218.

Malcolm, D. G., Roseboom, J. H., Clark, C. E., and Fazar, W. (1959). "Application of a technique for research and development program evaluation." *Operations Research*, 7(5), 646-669.

Mo, J., Yin, Y., and Gao, M. (2008). "State of the art of correlation-based models of project scheduling networks." *IEEE Transactions on Engineering Management*, 55(2), 349-358.

Morgan, M. G., Henrion, M., and Small, M. (1992). *Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis*, Cambridge university press.

Morris, P. W. (2013). *Reconstructing project management*, John Wiley & Sons.

NASA (2013). "Analytic method for probabilistic cost and schedule risk analysis." *National Aeronautics and Space Administration (NASA), Washington, DC.*

Newton, S. (1992). "Methods of analysing risk exposure in the cost estimates of high quality offices." *Construction Management & Economics*, 10(5), 431-449.

Palmer, C., Urwin, E. N., Niknejad, A., Petrovic, D., Popplewell, K., and Young, R. I. (2018). "An ontology supported risk assessment approach for the intelligent configuration of supply networks." *Journal of Intelligent Manufacturing*, 29(5), 1005-1030.

PMI (2013). *A guide to the project management body of knowledge*, Project Management Institute, Inc., Newtown Square, PA.

Ranasinghe, M. (2000). "Impact of correlation and induced correlation on the estimation of project cost of buildings." *Construction Management & Economics*, 18(4), 395-406.

Reinschmidt, K. F. (2009). "Project risk management: Class notes." Zachry Department of Civil Engineering, Texas A&M University, College Station, TX.

Rodger, J. A. (2013). "A fuzzy linguistic ontology payoff method for aerospace real options valuation." *Expert Systems with Applications*, 40(8), 2828-2840.

Safran (2020). "Pacific Northwest National Laboratory (PNNL)." <https://www.safran.com/case-studies/pacific-northwest-national-laboratory?hsCtaTracking=deabe776-2577-43d0-bb60-2096ab40a608%7C15f483a4-11a7-49fe-a243-a8e25fef0f86>. (October 17,2020).

Scheuer, S., Haase, D., and Meyer, V. (2013). "Towards a flood risk assessment ontology–Knowledge integration into a multi-criteria risk assessment approach." *Computers, Environment and Urban Systems*, 37, 82-94.

Schonberger, R. J. (1981). "Why projects are “always” late: a rationale based on manual simulation of a PERT/CPM network." *Interfaces*, 11(5), 66-70.

Skitmore, M., and Ng, S. (2002). "Analytical and approximate variance of total project cost." *Journal of Construction Engineering and Management*, 128(5), 456-460.

Song, J., Martens, A., and Vanhoucke, M. (2021). "Using schedule risk analysis with resource constraints for project control." *European Journal of Operational Research*, 288(3), 736-752.

Styan, G. P. (1973). "Hadamard products and multivariate statistical analysis." *Linear algebra and its applications*, 6, 217-240.

Tatikonda, M. V., and Rosenthal, S. R. (2000). "Technology novelty, project complexity, and product development project execution success: a deeper look at task uncertainty in product innovation." *IEEE Transactions on engineering management*, 47(1), 74-87.

Touran, A., and Wiser, E. P. (1992). "Monte Carlo technique with correlated random variables." *Journal of Construction Engineering and Management*, 118(2), 258-272.

Trietsch, D. (2005). "The effect of systemic errors on optimal project buffers." *International Journal of Project Management*, 23(4), 267-274.

Trietsch, D., Mazmanyan, L., Gevorgyan, L., and Baker, K. R. (2012). "Modeling activity times by the Parkinson distribution with a lognormal core: Theory and validation." *European Journal of Operational Research*, 216(2), 386-396.

van Dorp, J. R. (2005). "Statistical dependence through common risk factors: With applications in uncertainty analysis." *European Journal of Operational Research*, 161(1), 240-255.

van Dorp, J. R. (2020). "A dependent project evaluation and review technique: A Bayesian network approach." *European Journal of Operational Research*, 280(2), 689-706.

Van Slyke, R. M. (1963). "Monte Carlo methods and the PERT problem." *Operations Research*, 11(5), 839-860.

Wang, C.-H., and Huang, Y.-C. (2000). "A new approach to calculating project cost variance." *International Journal of Project Management*, 18(2), 131-138.

Werner, C., Bedford, T., Cooke, R. M., Hanea, A. M., and Morales-Nápoles, O. (2017). "Expert judgement for dependence in probabilistic modelling: a systematic literature review and future research directions." *European Journal of Operational Research*, 258(3), 801-819.

Williams, T. M. (1999). "The need for new paradigms for complex projects." *International journal of project management*, 17(5), 269-273.

Williams, T. M. (2003). "The contribution of mathematical modelling to the practice of project management." *IMA Journal of Management Mathematics*, 14(1), 3-30.

Xing, X., Zhong, B., Luo, H., Li, H., and Wu, H. (2019). "Ontology for safety risk identification in metro construction." *Computers in Industry*, 109, 14-30.