Decoupled Copula Models

Expert Belief Aggregation through Question-Invariant Spaces

Master Thesis - TU Delft

Hugo Save

Delft University of Technology

TU Delft

Decoupled Copula Models: Expert Belief Aggregation through Question-Invariant Spaces

by

Hugo Save

to obtain the degree of Master of Science in Applied Mathematics
at the Delft University of Technology,

to be defended publicly on July 8, 2025 at 15:00.

Project duration:	November 19, 2024 – July 8, 2025	
Thesis committee:	Dr. ir. G.F. Nane,	TU Delft, Supervisor
	Dr. A.F.F. Derumigny,	TU Delft

An electronic version of this thesis is available at
`http://repository.tudelft.nl/`.

TU Delft

"Science is the belief in the ignorance of experts." [1]

— Richard Feynman

"You must trust and believe in people or life becomes impossible." [2]

— Anton Chekhov

Acknowledgements

The completion of this thesis would not have been possible without the support of many individuals. First and foremost, I extend my sincere gratitude to my supervisor Dr. Tina Nane for introducing me to the fascinating field of expert judgment more than a year ago, supervising this thesis, and providing invaluable guidance and support throughout this research journey. It has been a privilege to work under her mentorship. Additionally, Dr. Alexis Derumigny deserves special thanks for being part of the thesis committee and for his time and expertise in evaluating this work.

Special recognition goes to my friend and colleague Johannes Taraz, who generously invested time in understanding my work and provided valuable feedback and fresh perspectives throughout these months. His insights have significantly enriched this research.

The exceptional teachers I have had at TU Delft deserve acknowledgment for preparing me for where I am today and contributing to making my academic experience at TU Delft truly valuable. Their dedication to education and research has shaped my academic journey.

Finally, this work would not have been possible without the unwavering support, patience, and encouragement of my family, partner, and friends during this intenstive yet rewarding period.

I acknowledge the use of the large language model Claude to assist with proofreading and language revision of this thesis. All ideas, analysis, and conclusions remain entirely my own work.

[1] "What is Science?" speech to the 15th annual meeting of the National Science Teachers Association, New York City (1966).

[2] Original Russian: "Надо всем верить, иначе жить нельзя." *Uncle Vanya* (1897), act 2.

Abstract

This thesis presents the decoupled copula model, a novel theoretical framework for aggregating expert judgments in structured expert judgment (SEJ) studies. The model's key innovation lies in transforming expert assessments into a "decoupled space" where systematic biases can be identified and corrected while capturing potential inter-expert dependencies. Unlike existing Bayesian SEJ methods, which are limited to linear error metrics, our framework accommodates flexible dependency measures with rigorous theoretical criteria and practical tests for their evaluation. While previous Bayesian approaches acknowledged the possibility of bias correction, they lacked practical procedures for implementing these corrections using historical test data. Our framework addresses these limitations by enabling flexible metric choices for measuring biases and dependencies. Captured biases include under- and overconfidence as well as consistent over- and underestimation. Additionally, we introduce novel calibration criteria that have been proven necessary for perfect aggregation methods, along with interpretable calibration metrics that measure discrepancies from these criteria. Empirical evaluation on 47 real-world SEJ studies demonstrates superior calibration properties while maintaining competitive predictive performance compared to established methods like the Classical Model. The empirical analysis reveals that inter-expert dependency modeling provides limited benefits, suggesting that systematic bias correction, rather than dependency modeling, drives improvements in aggregation performance in practical applications of our model.

List of Acronyms

SEJ Structured Expert Judgement

DM Decision Maker

PE Perfect Expert

IID Independent and Identically Distributed

CDF Cumulative Distribution Function

ECDF Empirical Cumulative Distribution Function

PDF Probability Density Function

KL Kullback-Leibler

MAP Maximum a Posteriori

MCMC Markov Chain Monte Carlo

MLE Maximum Likelihood Estimation

LOOCV Leave-One-Out Cross Validation

CM Classical Model

DP Density Product

EW Equal Weights

JC Jouini-Clemen Model

Contents

Chapter 1

Introduction

Expert judgment plays a critical role in decision-making under uncertainty, particularly in domains where empirical data is scarce, expensive, or unavailable [1]. In such situations, Structured Expert Judgement (SEJ) methods are employed to formally elicit and combine the beliefs of subject matter experts [1]–[5]. A key principle of SEJ is that both the elicitation and aggregation of expert opinions should follow a clearly defined and transparent process, designed to reduce cognitive biases and promote the clarity and reproducibility of the results [6]. Rather than relying on ad hoc or informal consultations, SEJ encourages a systematic approach that makes expert input suitable for scientific analysis. Building on this systematic foundation, this work presents a novel belief aggregation model that can (1) capture inter-expert dependencies, (2) correct for various experts' biases, (3) accommodate different types of quantities for measuring these biases and dependencies, and (4) provide both theoretical criteria and practical tests for evaluating these quantities. Here, 'bias' refers to the distributional difference between a question-invariant quantity derived from an expert's stated belief and the expected distribution of that same quantity if the expert's belief perfectly reflected the true underlying distribution.

This contribution sits within a diverse landscape of aggregation methodologies that can broadly be classified into behavioral and mathematical aggregation strategies [7], the latter being the focus of this work. Behavioral aggregation involves procedures aimed at reaching a consensus through direct or mediated group interaction [3], or facilitated discussions. While these methods can benefit from expert dialogue and the exchange of justifications, they are often susceptible to group dynamics, social pressures, and conformity effects, which may compromise the independence and diversity of opinions [8], [9]. In contrast, mathematical aggregation applies formal statistical or probabilistic rules to combine individual assessments without requiring interaction among experts. These methods, such as the Classical Model (CM) [1] or the Jouini-Clemen Model (JC) model [2], offer transparency, reproducibility, and resistance to social bias but come with their own assumptions about the nature of the expert distributions, dependencies, and weighting schemes. For a comprehensive overview of aggregation approaches, see, for instance, [10].

Among mathematical aggregation strategies, pooling algorithms are a widely adopted approach due to their computational simplicity, proven robustness, and interpretability [11]. Linear pooling combines expert assessments as weighted averages, formally expressed as $f^{\mathrm{DM}}(q) = \sum_{e=1}^{E} w_e f^e(q)$, where $f^e(q)$ represents expert e's subjective probability density for a random variable Q, and w_e are aggrega-

tion weights. Alternatively, logarithmic pooling uses multiplicative combinations: $f^{\mathrm{DM}}(q) = k \prod_{e=1}^{E} (f^e(q))^{\alpha_e}$, where k is a normalization constant and α_e are expert-specific exponents. While these methods offer transparency and computational efficiency, they typically assume independence between expert assessments, potentially overlooking correlations that may arise from shared training, common information sources, or similar analytical approaches [12]. Furthermore, by their mathematical construction, they cannot directly adjust for expert biases, such as consistent over- or underestimation of their believed median, on an individual level, but they can limit the influence of experts with undesirable biases. Two notable linear pooling methods are the widely used CM, where the weights are based on the experts' performance on a set of test questions, and the Equal Weights (EW) scheme where $w_e = 1/E$ for all experts [13].

In contrast to the postulated aggregation structure of linear or algorithmic pooling, Bayesian aggregation methods treat expert opinions as data to inform posterior beliefs about the quantity of interest [11]. The specific structure of a Bayesian DM varies depending on the specific model assumptions made to connect the belief distributions to the quantity of interest [2], [4]. Notable examples of Bayesian models include Winkler's model [4] and the copula-based JC model. Both of these models try to address the limitation of expert independence that is present in existing linear pooling models. Winkler's model focuses on the means of expert beliefs and models the linear errors between experts through a multivariate distribution. This has, among others, the consequence that all posterior densities will be unimodal. JC's approach uses, in contrast, the medians of expert assessments to model expert dependencies through the difference between the median and the realization. Instead of assuming a multivariate normal form, JC employs copulas to combine arbitrarily shaped beliefs together with inter-expert dependency modeling through the choice of copula.

For both of these models, however, several important limitations constrain their practical applicability. First, both Winkler's and JC's methods model expert dependencies exclusively through linear error structures, which may or may not be the most reliable way to capture expert dependencies. Second, while both approaches acknowledge that model parameters could be learned from historical performance data, they lack explicit numerical procedures for doing so. Third, neither model attempts to correct for potential systemic biases in experts, such as under- or overconfidence.

To address these limitations, this work introduces the *decoupled copula model* for expert belief aggregation that extends beyond the constraints of existing Bayesian approaches. The proposed method offers several key contributions: First, it provides flexibility in modeling expert dependencies through various statistical measures beyond linear error structures. Second, it presents explicit numerical procedures for learning model parameters from historical expert performance data. Third, the work includes both theoretical criteria and empirical tests for comparing different dependency measures, enabling systematic evaluation of aggregation performance.

A comparison of key characteristics across different expert judgment models is presented in Chapter 1, highlighting how the proposed decoupled approach addresses the limitations of existing methods.

The remainder of this thesis is structured as follows: Chapter 2 provides background on SEJ and establishes the mathematical foundation for copula-based aggregation. Chapter 3 presents formal DM definitions and a novel calibration metric to

Table 1.1: Comparison of expert judgment models: Winkler's model, the Jouini-Clemen model, the Classical model and, the decoupled copula model.

Model	**Type**	**Dependencies captured through**	**Mono or multi modal**	**Numerical procedures**
Winkler's	Bayesian	Linear error	Mono	N/A
JC's Copula	Bayesian	Linear error	Multi	N/A
Classical	Linear	N/A	Multi	Exists
Decoupled	Bayesian	Flexible	Multi	Exists

evaluate DMs. Chapter 4 presents the decoupled copula model with its theoretical derivation and numerical procedures for parameter estimation. Chapter 5 presents empirical evaluation and benchmarking of the decoupled copula model against established methods using data from real expert judgment studies. Chapter 6 discusses the implications of the findings, limitations and, future research directions.

The implementation of the methods and computational procedures described throughout this thesis are available in the accompanying code repository[1].

[1]Available at: `https://github.com/HugoSave/Decoupled`

Chapter 2

Background

This chapter establishes the theoretical foundation and practical context necessary for developing the decoupled copula model presented in Chapter 4. We begin by examining the practical considerations of expert elicitation and its inherent trade-offs, then review existing mathematical aggregation approaches, with particular attention to the CM and the JC model. The chapter concludes with essential mathematical background on copulas and dependency measures that will underpin our proposed framework.

2.1 Practical Considerations of SEJ Studies

In developing a new model, we want not only a mathematically correct framework but also a practically relevant one. For this, we need to consider the practical considerations of SEJ studies and their operational context.

When eliciting experts, there is often a trade-off between the costs of the elicitation and the information extracted from the experts. The trade-off comes both in the form of how many questions to ask, as well as in how much detail to elicit the experts' uncertainty about each question. The advantage of more questions comes from receiving more statistical data to make robust judgments of experts from, and the advantage of eliciting the uncertainty is that this tends to allow us to use more performant aggregation methods [14].

Focusing on a single quantity q, one way to model an expert's uncertainty about q, is to assume that the expert has a personal probability distribution for that quantity that represents the expert's belief about it. This modeling choice is the same as part of Savage's work on decision theory and personal probabilities [15]. In the scenario of q being continuous, we can denote expert e's probability distribution by the density function f^e and Cumulative Distribution Function (CDF) F^e. With this definition, we can more concretely characterize the tradeoff between costs and uncertainty elicitation, as the trade-off between time and exhaustion of the expert, against gaining a more accurate image of f^e.

One way to settle this trade-off, which is done in the classical model [1], is to elicit a limited amount of quantiles, $m_1, \dots, m_D$, of F corresponding to probabilities $p_1, \dots, p_D$. For example, one could choose $(p_1, p_2, p_3) = (5\%, 50\%, 95\%)$ and then elicit the respective quantiles m_d for $d = 1, \dots, 3$. m_d thus would have the property of $F(m_d) = p_d$. Finally, using these values, we construct a CDF function $\hat{F}(q)$ that has the property of $\hat{F}(m_d) = p_d$ for all d, that hopefully also approximates F

also for all other values that are not m_d. One way to do so is to choose the "least informative" distribution in the sense that it maximizes the Shannon entropy. For the case of finite support and quantiles constraints, this turns out to be a piecewise uniform distribution [16]. Another approach is to choose a reference distribution that is considered to be least informative and then choose $\hat{F}$ to minimize the Kullback-Leibler (KL) divergence between these while preserving the quantile and support constraints. These two approaches coincide when the reference distribution is the uniform distribution [16]. Another choice of reference distribution could be the log uniform distribution sometimes used in the CM which leads to a piecewise log-uniform distribution for $\hat{F}$ [1].

The practical constraints inherent in SEJ studies introduce several complexities that must be considered when developing aggregation methods. For SEJ models that require historical test questions with known answers, elicitation costs necessitate that the number of test questions is typically limited, often being fewer than 15 [14]. Additionally, it is not uncommon in applied SEJ studies for the ratio between the largest and smallest test question realizations to exceed 10^5 in magnitude [17]–[21], creating difficulties for statistical modeling across vastly different scales.

These practical considerations lead to a few fundamental challenges when designing models for SEJ studies:

- **Limited sample sizes**: The number of questions available for learning is constrained by practical and economic factors, requiring methods that can perform effectively with sparse data.
- **Scale heterogeneity**: Questions within a study may span multiple orders of magnitude, necessitating approaches that can handle diverse scales and range effectively.
- **Restricted belief information**: Expert uncertainty is typically captured through a limited number of quantiles, representing a balance between the richness of elicited information and practical feasibility constraints.

2.2 Belief Estimation from Quantile Data

As mentioned in the previous section, it can be practically needed to estimate expert beliefs from quantile data. Given a matrix of D quantiles from E experts, $\mathbf{m}_{Quantiles} \in \mathbb{R}^{E\times D}$, whose columns belong to strictly increasing accumulated probabilities $p_1, \dots, p_D$, we use the following procedure to generate E distributions, $\hat{F}^1, \dots, \hat{F}^E$. These distributions aim to estimate the true expert beliefs the $\mathbf{m}_{Quantiles}$ are elicited from. It is assumed that $p_1 \neq 0$, $p_D \neq 1$ and instead these cutoff points are estimated from $\mathbf{m}_{Quantiles}$.

1. Calculate L and U as the minimum and maximum of the quantiles in $\mathbf{m}_{Quantiles}$:

$$L = \min_{i,e} m_i^e, \quad U = \max_{i,e} m_i^e.$$

2. Define two non-negative overshoot parameters k_L, k_U and the extended support parameters to L^* and U^* as

$$L^* = L - k_L(U - L), \quad U^* = U + k_U(U - L).$$

If the support of the question of interest, Q, is known a priori, then k_L and k_U should be restricted to ensure $L^*, U^* \in \text{support}(Q)$. For example, if Q is a proportion, then L^*, U^* should be restricted to $[0, 1]$. In this work, however, we let both k_L and k_U be constants equal to 10% at all times.

3. For each expert e, estimate the distribution function $\hat{F}^e$ as the linear interpolation between domain values $(m_0^e, \dots, m_{D+1}^e)$ and codomain values $(p_0, \dots, p_{D+1})$, where $m_0^e = L^*, m_{D+1}^e = U^*$ and $p_0 = 0, p_{D+1} = 1$. Let $\hat{F}^e(q) = 0$ for $q \leq m_0^e$ and $\hat{F}^e(q) = 1$ for $q \geq m_{D+1}^e$. Explicitly this can also be written as

$$\hat{F}^e(q) = \begin{cases} 0 & \text{if } q \leq m_0^e \\ p_{d-1} + (p_d - p_{d-1})\frac{q - m_{d-1}^e}{m_d^e - m_{d-1}^e} & \text{if } q \in (m_{d-1}^e, m_d^e] \text{ for } d = 1, \dots, D+1 \\ 1 & \text{if } q > m_{D+1}^e \end{cases}$$

Which has density

$$\hat{f}^e(q) = \begin{cases} 0 & \text{if } q \leq m_0^e \\ \frac{p_d - p_{d-1}}{m_d^e - m_{d-1}^e} & \text{if } q \in (m_{d-1}^e, m_d^e] \text{ for } d = 1, \dots, D+1 \\ 0 & \text{if } q \geq m_{D+1}^e \end{cases}$$

Expert-specific support variation: An alternative approach calculates the support parameters individually for each expert. Instead of global L and U, we compute expert-specific bounds:

$$L^e = \min_d m_d^e, \quad U^e = \max_d m_d^e$$

and define $L^{e*} = L^e - k_L(U^e - L^e)$ and $U^{e*} = U^e + k_U(U^e - L^e)$ for each expert e. This generates expert-specific distributions $\hat{f}^e_{\text{E.Sup.}}$ that respect individual expert ranges rather than forcing a common global support across all experts. When clarification is needed, we denote the global support variant as $\hat{f}^e_{\text{G.Sup.}}$, but when no subscript is written, we refer to the global support version by default.

The choice between global and expert-specific support estimation methods can be motivated by both empirical and pragmatic considerations. Empirical motivations could come from better performance according to some chosen metric. Pragmatic motivations emerge from the requirements or limitations of the chosen SEJ method. For instance, when using logarithmic pooling then the support of the final DM distribution equals the intersection of individual expert belief supports. Thus, it might be desired to ensure that all experts share a common support to prevent the aggregated distribution from having empty support.

2.3 Existing Models

While a large range of existing models exist, we will here introduce a selection that our proposed model will later be benchmarked against in Chapter 5. In particular, we will benchmark against the commonly used CM, the JC model, and against a set of simple baseline models: EW, Density Product (DP), and uniform.

2.3.1 Classical Model

The CM, [1], is a SEJ model that combines expert assessments with a performance-based weighting scheme. The model is based on the principle that experts should be weighted according to their demonstrated ability to provide calibrated and informative probability assessments. This is achieved by weighing experts depending on their performance on a set of test questions for which the true answers are known to the analyst but not to the experts themselves. Experts provide their uncertainty assessments for these test questions in the form of probability distributions, typically by specifying percentiles (commonly the 5th, 50th, and 95th percentiles: $m_{5\%}$, $m_{50\%}$, and $m_{95\%}$).

The framework of the Classical Model centers on two key scoring metrics derived from the test question assessments: the calibration score and the information score. The calibration score measures statistical accuracy by first dividing the probability space into intervals based on the provided percentiles. For example, with three percentiles (5th, 50th, 95th), this creates four intervals: $(-\infty, m_{5\%}]$, $(m_{5\%}, m_{50\%}]$, $(m_{50\%}, m_{95\%}]$, and $(m_{95\%}, \infty)$. The empirical probability is then calculated as $p_i = \frac{s_i}{N}$, where s_i is the number of realizations falling in the interval i and N is the total number of test questions. The calibration score is calculated as:

$$\text{Cal}(e) = 1 - \chi^2_R[2NI(s,p)] \tag{2.1}$$

where R is the number of intervals, $s = (s_1, \dots, s_{R+1})$ is the sample distribution over the $R+1$ probability intervals, $p = (p_1, \dots, p_{R+1})$ are the theoretical probabilities for each interval, χ^2_R is the cumulative chi-square distribution with R degrees of freedom, and $I(s,p)$ is the relative information

$$I(s,p) = \sum_{i=1}^{N} s_i \log \frac{s_i}{p_i}. \tag{2.2}$$

In addition to the calibration score, the information score quantifies the concentration of an expert's distributions relative to a background measure. In case the background measure is the uniform measure then the information score is calculated as

$$I(e) = \frac{1}{N} \sum_{i=1}^{N} \left[\ln(m_{iR+1} - m_{i0}) + \sum_{r=1}^{R+1} p_r \ln \left(\frac{p_r}{m_{ir} - m_{ir-1}} \right) \right] \tag{2.3}$$

where m_{ir} represents the r-th percentile value for test question i, with m_{i0} and m_{iR+1} being the intrinsic range bounds, and p_r is the probability mass in interval r. These scores combine multiplicatively to create the combined score:

$$w'_e = \text{Cal}(e) \times I(e) \times \mathbf{1}_{\{\text{Cal}(e) > \alpha\}} \tag{2.4}$$

where α is the significance threshold (typically 0.05). The final normalized weights are:

$$w_e = \frac{w'_e}{\sum_{j=1}^{E} w'_j}. \tag{2.5}$$

The final aggregation step constructs the Decision Maker (DM) distribution through weighted linear pooling. For each test question, the aggregated distribution is:

$$F_{\text{DM}}(q) = \sum_{e=1}^{E} w_e F^e(q) \tag{2.6}$$

where F_e is the belief of expert e of the test question. The beliefs F^e are estimated according to Section 2.2 from belief quantiles of the test question.

Optimized Significance Threshold

The classical model offers flexibility in determining the significance threshold α. While α can be set as a predetermined value, the model also provides an optimization approach. In this alternative method, if a weight w_{DM} is assigned to the DM on the test questions using the same weighting scheme applied to experts, then α is chosen to maximize this weight value. This optimization variant, which we refer to as the optimized classical model, serves as one of the benchmark methods in our later comparative analysis.

2.3.2 The Jouini-Clemen Model

The JC model from 1996 [2] introduces a copula-based model for expert aggregation with inter-expert dependency modeling. For the JC model, we see the target question of interest as a random variable Q and we will also see the expert beliefs as random objects. This probabilistic framework enables the modeling of dependencies between expert beliefs through copula structures. Because our work with the decoupled copula model is inspired by the JC model, we will end this section by going through some of its limitations and potential improvements.

Regarding expert dependency modeling, it would arguably be preferable to model the dependencies between the belief functions F^e. The difficulty with this, however, is that these complete function dependencies are complex to model and capture. As a simplifying modeling assumption, Jouini and Clemen assume that relevant dependencies can be captured by examining only the medians m^{e*} of F^e, where the asterisk notation differentiates the observed median values from free argument variables m^e. To model inter-expert dependencies, they let M^e be the random variable that generated m^{e*}, which may be interpreted as experts having an inherently random element in their opinion formation. Using the notation $\mathbf{M} = (M^1, \dots, M^E)$ and $\mathbf{m}^* = (m^{1*}, \dots, m^{E*})$, the Bayesian perspective seeks $f_{Q|\mathbf{M}=\mathbf{m}^*}(q)$, which with a flat prior is proportional to $f_{\mathbf{M}|Q=q}(\mathbf{m}^*)$.

To facilitate this modeling, Jouini and Clemen define the linear error variable $\mathcal{E}^e = Q - M^e$ with realizations $\varepsilon^e = q - m^{e*}$, grouped as $\boldsymbol{\mathcal{E}} = (\mathcal{E}^1, \dots, \mathcal{E}^E)$ and $\boldsymbol{\varepsilon} = (\varepsilon^1, \dots, \varepsilon^E)$. They propose that an appropriate density for $\mathcal{E}^e$ is $f^e(\varepsilon + m^{e*})$, the elicited belief centered at zero, and implicitly assume that $\mathcal{E}^e$ and Q are independent. The corresponding CDF of $\mathcal{E}^e$ is then $F^e(\varepsilon + m^{e*})$.

The connection between conditional median densities and errors can be derived through:

$$\begin{aligned} F_{M^e|Q=q}(m) &= P(M^e \le m \mid Q = q) = P(q - \mathcal{E}^e \le m \mid Q = q) \\ &= 1 - P(\mathcal{E}^e < q - m \mid Q = q) = 1 - F_{\mathcal{E}^e|Q=q}(q - m) \end{aligned} \tag{2.7}$$

which together with the independence assumption leads to the conditional distribution

$$F_{M^e|Q=q}(m) = 1 - F^e(q - m + m^{e*}) \tag{2.8}$$

and corresponding density

$$f_{M^e|Q=q}(m) = f^e(q - m + m^{e*}). \tag{2.9}$$

Having derived the marginal densities of $M^e \mid Q = q$, we can construct the joint density $f_{\mathbf{M}|Q=q}$ through copula decomposition as in Corollary 1.1 to be:

$$f_{\mathbf{M}|Q=q}(m^1, \dots, m^E) = c_{\mathbf{M}|Q=q}\left(F_{M^1|Q=q}(m^1), \dots, F_{M^E|Q=q}(m^E)\right) \prod_{e=1}^{E} f_{M^e|Q=q}(m^e) \tag{2.10}$$

where $c_{\mathbf{M}|Q=q}$ is the conditional copula density of $\mathbf{M}$ conditional on $Q = q$.

Evaluating this expression at the observed medians, where $F_{M^e|Q=q}(m^{e*}) = 1 - F^e(q)$ and $f_{M^e|Q=q}(m^{e*}) = f^e(q)$, yields:

$$f_{Q|\mathbf{M}=\mathbf{m}^*}(q) \propto c_{\mathbf{M}|Q=q}\left(1 - F^1(q), \dots, 1 - F^E(q)\right) \prod_{e=1}^{E} f^e(q). \tag{2.11}$$

Finally, Jouini and Clemen make the explicit assumption that $c_{\mathbf{M}|Q=q}$ is independent of q and implicitly assume that the copula dependence of the medians is equivalent to that of the errors, leading to the final result:

$$f_{Q|\mathbf{M}=\mathbf{m}^*}(q) \propto c_{\boldsymbol{\mathcal{E}}}(1 - F^1(q), \dots, 1 - F^E(q)) \prod_{e=1}^{E} f^e(q). \tag{2.12}$$

For the error copula $c_{\boldsymbol{\mathcal{E}}}$, they use a Frank copula parameterization. To assess the single parameter, JC proposes having someone knowledgeable about the experts answer questions such as (paraphrased): "If experts 1 and 2 give their median estimates on two questions, what is the probability that the median errors of the second question will be both greater than or both less than their errors in the first question?". From such probability statements, they estimate pairwise Kendall tau correlations and determine the Frank copula parameter.

In this work, we are more interested in methods that do not require personal knowledge about each expert, instead drawing conclusions solely from assessed beliefs and test questions. Therefore, we will not cover this elicitation process in detail and will instead consider a variation where we fit the copula using Maximum Likelihood Estimation (MLE). Given N test questions, we denote the observed errors from question i as $\boldsymbol{\varepsilon}_i = (\varepsilon_i^1, \dots, \varepsilon_i^E)$ where ε_i^e is the observed error of expert e for question i. We denote the random variable of ε_i^e as $\mathcal{E}_i^e$. To fit the copula, we transform these errors to $[0, 1]$ bounded variables $\mathbf{u}_1, \dots \mathbf{u}_N$ where $\mathbf{u}_i = (u_i^1, \dots, u_i^E)$ through

$$u_i^e = F_{\mathcal{E}_i^e}(\varepsilon_i^e) = F_i^e(\varepsilon_i^e + m_i^{e*})$$

where we have denoted F_i^e as the observed belief of expert e on question i and m_i^{e*} as the median of this belief. These $\mathbf{u}_i$ samples are then used as in the Frank MLE estimation explained in Section 4.3.5.

Motivation for the Decoupled Copula Model

While the JC model introduced copula-based dependency modeling into the SEJ field, some limitations and potential extensions motivate the development of our proposed approach in Chapter 4:

- **Dependence on expert knowledge assessment:** The original JC framework requires study researchers to have someone who can reliably assess the

error dependencies between elicited experts through subjective probability statements. This places an additional burden on the research team and introduces potential subjectivity in dependency parameter estimation.

- **Theoretical assumptions:** The JC model makes several assumptions that lack explicit theoretical justification. First, it assumes that the copula dependence structure of expert medians is equivalent to that of the error variables. Second, the choice of $f^e(\varepsilon + m^{e*})$ as an appropriate density for the error variable $\mathcal{E}^e$ is not motivated theoretically.

- **Limited scope of dependency modeling:** The JC approach focuses exclusively on dependencies between expert medians, potentially missing other important aspects of expert belief relationships. Dependencies may exist at different quantile levels or involve other distributional characteristics beyond central tendencies.

- **Utilization of available data:** When historical test questions with known realizations are available, this information could potentially be leveraged to make better estimates of error distributions and dependencies than the approach proposed in the original model.

- **Restriction to linear error measures:** The focus on linear error measures of the form $Q - M^e$ may or may not be the most suitable to capture dependencies across different contexts and question types. A more flexible framework allowing for alternative error measures could prove more robust and generally applicable.

These limitations suggest the need for a more general framework that can capture expert dependencies through data-driven methods while allowing for greater flexibility in both the choice of dependency measures and the underlying mathematical structure.

2.3.3 Baseline Models

In addition to the classical model we also define the following baseline models.

Equal Weights

Let $w_e = 1/E$ where E is the number of experts and create the DM according to Eq. (2.6).

Density Product

Let f^e be beliefs estimated according to Section 2.2 and then define the DP DM density as being equal to

$$C \prod_{e=1}^{E} f^e(q)$$

where C is a normalization constant.

Uniform

For the uniform DM we calculate the support $[L^*, U^*]$ from the expert quantiles as in Section 2.2 and then define the uniform DM to be the continuous uniform distribution on $[L^*, U^*]$.

2.4 The Dataset

For empirical evaluations, we will use the expert judgment dataset introduced in [14] that contains 49 expert judgment studies conducted between 2006 and 2019. Among these 49 studies, there are two studies with 30 or more experts in it that we excluded for computational ease. We will refer to the collection of the remaining 47 studies as the "dataset" throughout this work.

In the (47 studies) dataset, there are a total of 548 test questions and 446 experts. For each expert and each question in every study, the dataset contains the expert's assessed 5%, 50%, and 95% percentiles, providing a quantile-based representation of each expert's belief distribution. The configurations of the number of experts and test questions per study can be seen in Fig. 2.1.

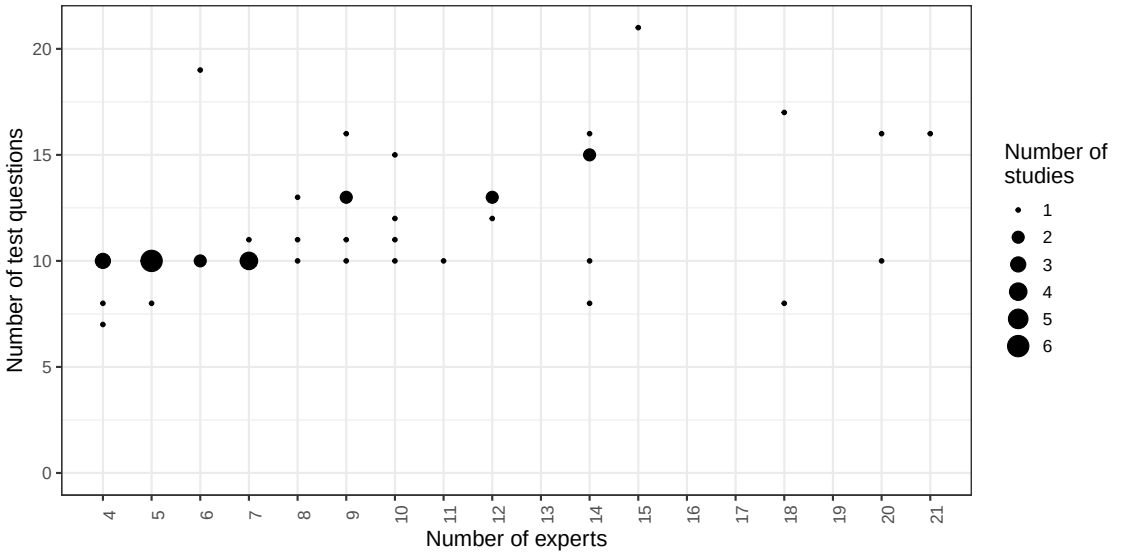


Figure 2.1: Plot showing the configurations of number of experts and test questions present in the 47 different studies of the dataset.

2.5 Leave-One-Out Cross-Validation

Leave-One-Out Cross Validation (LOOCV) is used throughout this work as the primary method validation framework. The specific implementation of LOOCV in this context operates at the study level rather than across the entire dataset. This comes from how the expert judgment methods in general are defined per study and require the same set of experts to answer all questions. Thus, for each study s in the dataset containing N_s questions and E_s experts, LOOCV is performed as follows:

1. For each question $t \in \{1, \ldots, N_s\}$ in study s:

(a) **Training set**: Use the remaining $N_s - 1$ questions with their known realizations and expert beliefs

(b) **Test question**: Hold out question t with its known realization q_t and expert beliefs $\mathbf{F}_t$

(c) **Model fitting**: Estimate model parameters using only the training questions

(d) **Prediction**: Apply the fitted model to the expert beliefs for question t to obtain a predictive distribution

(e) **Evaluation**: Compare the predictive distribution against the known realization q_t

2. Repeat for all studies in the dataset

This approach generates $\sum_s N_s$ total train/test evaluations across the dataset. For our dataset, this equals 548 questions over the 47 studies.

2.6 Mathematical Background

This section will work as a reference section for existing theorems that will be used later in the work.

2.6.1 Copulas

Copulas are a way to model multivariate distributions that decouple the dependency modeling from the marginal distributions. In this section, we will introduce the definition of the copula together with properties and common parameterizations that will be used throughout the paper. For more details and historical context, see, for example, [22].

Definition 1. A d-dimensional copula is a d-dimensional CDF, $C(u_1, \dots, u_d)$, with support on the d-dimensional hypercube, $[0,1]^d$, and whose marginals are uniformly distributed on $[0,1]$.

With $C(u_1, \dots, u_d)$ being a CDF we can also associate any random vectors with univariate components, $\bar{U} = [U_1, \dots, U_d]$ with $U_i \sim U(0,1)$ for $i = 1, \dots d$, to have a copula distribution, $\bar{U} \sim C$. The versatility of the copulas expands further than just univariate random vectors, however, as shown by Sklar's theorem from 1959, [23].

Theorem 1. Let F be a d-dimensional CDF with univariate marginals $F_1, \dots F_d$. Then there exists a unique d-dimensional copula, C, such that for all $(x_1, \dots, x_d) \in \mathbb{R}^d$,

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)). \tag{2.13}$$

Sklar's theorem shows that arbitrary CDFs with continuous marginals can be decomposed into their marginal distributions and a copula function, which captures the dependency structure between the variables. This allows for flexible modeling of dependencies independently of the choice of marginals. Taking the derivative of this expression gives us similar results for the density.

Corollary 1.1. Let F be a d-dimensional CDF with continuous marginals $F_1, ... F_d$, and let f be the joint Probability Density Function (PDF) of F with marginals $f_1, ..., f_d$. Then there exists a unique d-dimensional copula, C, such that for all $(x_1, ..., x_d) \in \mathbb{R}^d$,

$$f(x_1, ..., x_d) = c(F_1(x_1), ..., F_d(x_d))f_1(x_1) \cdots f_d(x_d) \tag{2.14}$$

where $c(u_1, ..., u_d) = \frac{\partial^d}{\partial u_1 ... \partial u_d} C(u_1, ..., u_d)$.

2.6.2 Copula Families

While the literature contains numerous copula families, we present here a selection of commonly used bivariate copulas that will be employed in our subsequent analysis. For a comprehensive treatment of copula families, we refer to Nelsen [24].

The copulas we consider belong to different mathematical classes. The Clayton, Gumbel, Frank, and Joe copulas are members of the Archimedean family, characterized by their construction through generator functions. In contrast, the Gaussian copula belongs to the elliptical family, derived from elliptical distributions. While we do not delve into the theoretical distinctions between these classes, this diverse selection ensures that the Dißmann algorithm described in Section 2.6.3 can identify appropriate copulas for various dependency patterns in the data.

Table 2.1 presents the copulas used in this work, along with their functional forms and parameter ranges.

Table 2.1: Copula families used in this study. The Gaussian and Frank copula are given in their multivariate form for dimensions $d \geq 2$.

Copula	$C(u, v)$ or $C(u_1, ..., u_d)$	**Parameter**
Independence	uv	-
Gaussian	$\Phi_R(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d))$	$R \in \mathbb{R}^{d \times d}$
Clayton	$(u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$	$\theta \in (0, \infty)$
Gumbel	$\exp\{-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}\}$	$\theta \in [1, \infty)$
Frank	$-\frac{1}{\theta} \ln\left(1 + \frac{\prod_{i=1}^{d}(e^{-\theta u_i} - 1)}{(e^{-\theta} - 1)^{d-1}}\right)$	$\theta \in \mathbb{R} \setminus \{0\}$
Joe	$1 - [(1-u)^\theta + (1-v)^\theta - (1-u)^\theta(1-v)^\theta]^{1/\theta}$	$\theta \in [1, \infty)$

The Gaussian copula is parameterized by a correlation matrix R, where Φ denotes the univariate standard normal CDF and Φ_R represents the multivariate normal CDF with correlation matrix R. The independence copula serves as a baseline, representing the absence of dependence between variables. For the bivariate case, all copulas except the independence copula are single-parameter families, with the Gaussian copula's correlation matrix R reducing to a single correlation coefficient.

2.6.3 Regular Vine Copulas

Regular vine copulas provide a flexible framework for modeling complex multivariate dependence structures by decomposing them into a cascade of bivariate copulas [25], making them particularly valuable for applications such as modeling financial returns where traditional approaches often fail to capture complex dependency patterns [26].

The key insight behind regular vines is that complex multivariate dependencies can be systematically broken down into a series of bivariate relationships. Rather than attempting to specify a single high-dimensional copula directly, regular vines build up the joint distribution by modeling how variables relate in pairs, and then how they relate conditionally given other variables, proceeding in a hierarchical manner through multiple tree levels.

A central part of using regular vines is determining the tree structure. Dißmann et al. [26] developed a sequential selection algorithm utilizing maximum spanning trees that prioritizes the strongest dependencies. In their approach, the first tree captures the strongest unconditional pairwise dependencies between variables, with subsequent trees building upon these relationships. Put loosely the procedure iteratively constructs each tree level through:

1. Computing empirical Kendall's tau coefficients for all valid variable pairs that are allowed to be part of the sub-tree.

2. Constructing a maximum spanning tree where edge weights are given by the absolute Kendall tau correlations

3. Selecting appropriate bivariate copula families for each edge and estimating their parameters

This process continues for each tree level until the vine structure is complete. While this sequential strategy does not guarantee globally optimal structures, it effectively prioritizes modeling the strongest dependencies in initial trees, aiming to capture the most significant dependency relations.

For implementation of this algorithm we used the rvinecopulalib package [27]. More detailed code signatures and fitting procedures are provided in Appendix A.5. For the detailed mathematical formulation of regular vine copulas and the Dißmann algorithm, see [26].

2.6.4 Inference Functions for Margins

When estimating parameters for copula models with continuous marginals, two main approaches are available: full MLE and multi-step procedures. In full MLE, all parameters (marginal and copula) are estimated simultaneously by maximizing the joint likelihood. An alternative approach involves sequential estimation stages.

Copula estimation inherently depends on the choice of marginal estimation method, as the marginal transformations directly affect the copula inputs. This two-step procedure of first estimating the marginal parameters individually and then the copula parameters is called the *method of inference functions for margins* (IFM) [28]. The IFM approach first estimates marginal distribution parameters then uses these estimates to transform observations to uniform margins via the probability integral transform, and finally estimates copula parameters from the transformed data.

Denoting θ_0 the set of parameters that generated the data, and $\hat{\theta}_{IFM}$ the parameters estimated from IFM, we have, under suitable regularity conditions, that

$$\sqrt{N}(\theta_0 - \hat{\theta}_{IFM}) \to \mathcal{N}(0, \nu(\theta_0)) \tag{2.15}$$

where N is the number of samples the marginals and copula are estimated from and $\nu(\theta_0)$ is a matrix defined in [28]. This result establishes the asymptotic normality of IFM estimators. Proof and details can be seen in [28]. This applies to both MLE and Bayesian estimation setups for the marginals and copula.

The IFM approach offers computational advantages for high-dimensional problems, as it avoids optimizing complex multivariate likelihood functions. However, it may be less statistically efficient than full MLE due to the sequential nature of parameter estimation.

2.6.5 Measure of Dependence - Distance Correlation

For the parameter selection of the upcoming decoupled copula model, we will need a measure of dependence between random variables of different dimensions. For this purpose we will use the modified distance correlation (dCor) introduced by Székely and Rizzo in 2013, [29], as a high-dimensional adjustment to the regular distance correlation metric introduced by Székely et al. in [30]. Put loosely, the dCor statistic measures a weighted L^2 distance between the joint characteristic function and the product of the marginal characteristic functions. The dCor statistic has the advantage of being applicable to random vectors of different dimensions, and it has an asymptotic Student t asymptotic distribution under the independence assumption. This consistent behavior for varying dimension sizes will turn out to be important for our scenario because we will want to compare dependencies between studies where dimensions will differ. Also, the asymptotic distribution will allow us to do formal hypothesis testing. What follows is a short summary of some of the definitions and results originally presented by Székely and Rizzo. These will later be used in Section 4.3.1.

Let $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$ be random vectors with dimension p and q. We are interested in whether $\mathbf{X}$ and $\mathbf{Y}$ are independent. Let $\mathbf{X_i}, \mathbf{Y}_i$ be n samples, $i = 1, \dots, n$, of the same distribution of $\mathbf{X}$ and $\mathbf{Y}$. For notational ease, let $|\cdot|$ be the Euclidean norm and define

$$a_{ij} = |\mathbf{X}_i - \mathbf{X}_j|, \quad i, j = 1, \dots, n,$$
$$a_{i.} = \sum_{k=1}^{n} a_{ik}, \quad a_{.j} = \sum_{k=1}^{n} a_{kj}, \quad \bar{a}_i = \bar{a}_{i.} = \frac{1}{n} a_{i.},$$
$$a_{..} = \sum_{i,j=1}^{n} a_{ij}, \quad \bar{a} = \frac{1}{n^2} \sum_{i,j=1}^{n} a_{ij} \quad \bar{a}_j = \bar{a}_{.j} = \frac{1}{n} a_{.j}.$$

and let $b_{i,j}$, $b_{i.}$, $b_{.j}$, $\bar{b}_i$, $b_{..}$, $\bar{b}$ and $\bar{b}_j$ be defined in a respective manner for $\mathbf{Y}_i$. Using these quantities we can define the non-modified sample distance covariance, $\mathcal{V}_n$, as

$$\mathcal{V}_n(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{i,j}^{n} A_{i,j} B_{i,j}$$

where

$$A_{i,j} = a_{ij} - \bar{a}_i - \bar{a}_j + \bar{a}, \quad B_{i,j} = b_{ij} - \bar{b}_i - \bar{b}_j + \bar{b}.$$

For the modified version, however, we modify A and B according to

$$A^*_{i,j} = \begin{cases} \frac{n}{n-1}\left(A_{i,j} - \frac{a_{ij}}{n}\right), & i \neq j; \\ \frac{n}{n-1}(\bar{a}_i - \bar{a}), & i = j, \end{cases} \qquad B^*_{i,j} = \begin{cases} \frac{n}{n-1}\left(B_{i,j} - \frac{b_{ij}}{n}\right), & i \neq j; \\ \frac{n}{n-1}(\bar{b}_i - \bar{b}), & i = j. \end{cases}$$

and define $\mathcal{V}_n^*$ as

$$\mathcal{V}_n^* = \frac{1}{n(n-3)} \left\{ \sum_{i,j=1}^{n} A_{i,j}^* B_{i,j}^* - \frac{n}{n-2} \sum_{i=1}^{n} A_{i,i}^* B_{i,i}^* \right\}.$$

and the modified distance correlation statistic, $\mathcal{R}^*$, as

$$\mathcal{R}_n^*(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^*(\mathbf{X}, \mathbf{X}) \mathcal{V}_n^*(\mathbf{Y}, \mathbf{Y})}} \tag{2.16}$$

if $\mathcal{V}_n^*(\mathbf{X}, \mathbf{X}) \mathcal{V}_n^*(\mathbf{Y}, \mathbf{Y}) > 0$ otherwise $\mathcal{R}_n^*(\mathbf{X}, \mathbf{Y}) = 0$. The image of $\mathcal{R}_n^*$ is $[-1, 1]$. With this, we will also define the t-statistic

$$\mathcal{T}_n = \sqrt{\nu - 1} \frac{\mathcal{R}_n^*}{\sqrt{1 - (\mathcal{R}_n^*)^2}}, \tag{2.17}$$

where $\nu = \frac{n(n-3)}{2}$. We can then state the main theorem of Székely and Rizzo's paper as

Theorem 2. If the coordinates of $\mathbf{X}$ and $\mathbf{Y}$ are Independent and Identically Distributed (IID) with positive finite variance, for fixed sample size $n \geq 4$ the following hold.

i. Under independence of $\mathbf{X}$ and $\mathbf{Y}$,

$$P(\mathcal{T} < t) \underset{p,q \to \infty}{\to} P(t_{\nu-1} < t),$$

where $t_{\nu-1}$ is a Student t distributed random variable with $\nu - 1$ degrees of freedom.

ii. Let $c_\alpha = t_{\nu-1}^{-1}(1 - \alpha)$ denote the $(1 - \alpha)$ quantile of a Student t distribution with $\nu - 1$ degrees of freedom. The t-test of independence at significance level α rejects the independence hypothesis whenever $\mathcal{T}_n > c_\alpha$ is unbiased.

From this we also get the corollary that:

Corollary 2.1. Under independence of $\mathbf{X}$ and $\mathbf{Y}$, if the coordinates of $\mathbf{X}$ and $\mathbf{Y}$ are IID with positive finite variance, then the limit distribution of $(1 + \mathcal{R}_n^*)/2$ is a symmetric beta distribution with shape parameter $(\nu - 1)/2$. It follows that, in high dimension the large sample distribution of $\sqrt{\nu - 1}\mathcal{R}_n^*$ is approximately standard normal.

Thus, we can loosely interpret near-zero realizations of the modified correlation statistic as it being likely that the variables are independent, while realizations further away from zero would indicate dependence.

2.6.6 Probability Integral Theorem

Theorem 3 (Probability Integral Theorem)**.** Let X be a continuous random variable with CDF $F_X(x)$. Then the random variable $U = F_X(X)$ follows a uniform distribution on $[0, 1]$.

Proof. This proof follows the exposition in [31]. Let $U = F_X(X)$, then U is in $[0, 1]$. For $u \in [0, 1]$, consider:

$$F_U(u) = P(U \leq u) \tag{2.18}$$
$$= P(F_X(X) \leq u) \tag{2.19}$$

Defining the inverse CDF F_X^{-1} as

$$F_X^{-1}(y) = \inf\{x : F(x) \geq y\}, \tag{2.20}$$

gives us

$$F_U(u) = P(X \leq F_X^{-1}(u)) \tag{2.21}$$
$$= P(F_X^{-1}(F_X(X)) \leq F_X^{-1}(u)) \tag{2.22}$$
$$= P(X \leq F_X^{-1}(u)) \tag{2.23}$$
$$= F_X(F_X^{-1}(u)) = u \tag{2.24}$$

□

Chapter 3

Evaluating Methods

Before discussing the decoupled copula model in Chapter 4, we will here introduce some definitions and lemmas regarding the evaluation of DMs. We present this before the decoupled copula model because the methodology of selecting parts of the model will be connected to the theory introduced here. The theory presented here will also be used in the results Chapter 5, when comparing various different methods.

3.1 Decision Makers

We begin by establishing the fundamental concepts that underlie our evaluation framework. At the core is the concept of a DM, which formalizes how we represent beliefs about uncertain quantities.

Definition 2. A **question** Q is a continuous random variable with CDF F_Q.

Definition 3. A decision maker for a question Q is a distribution F that aims to approximate the distribution F_Q.

With this definition, all experts are DMs in the sense that each expert belief F^e aims to approximate F_Q. Similarly, any SEJ model that aggregates these beliefs is also a DM for the same question. This perspective allows us to evaluate both individual expert beliefs and aggregation methods using the same theoretical framework.

Definition 4. A decision maker F for a question Q, is **perfect** if $F = F_Q$. If the DM is an expert's belief function, $F = F^e$, we call that expert perfect.

While this defines DMs for individual questions, we are sometimes interested in the methods producing a series of DMs over multiple questions. This is particularly relevant if we only have a single realization of a question, which gives little statistical material to evaluate the performance of the DM. To analyze such systems, we introduce the concept of distribution and question generators.

Definition 5. A distribution generator is a random variable $\mathcal{F}$ that has CDF functions as realizations. With $\mathcal{F}(q)$ we denote the random variable that evaluates the CDF at q. That is, if $\mathcal{F}$ has a realization F, then $\mathcal{F}(q)$ has the realization $F(q)$. Furthermore, with $\mathcal{F}^{-1}(p)$ we refer to the random variable that evaluates the quantile function at probability p, such that if $\mathcal{F}$ has realization F, then $\mathcal{F}^{-1}(p)$ has realization $F^{-1}(p)$.

Definition 6. A question generator, $\mathcal{F}_{\mathcal{Q}}$, is a distribution generator that produces distributions corresponding to observable random variables. With $\mathcal{Q}$ we denote the random quantity that the $\mathcal{F}_{\mathcal{Q}}$ produces the distribution of. That is, a realization of a question generator is a CDF F_Q that describes the distribution of a random variable Q, and the realization q of Q is also the realization of $\mathcal{Q}$.

Example (Question Generator): Let a question generator be defined as

$$\mathcal{F}_{\mathcal{Q}} = \begin{cases} F_{Q_1} \sim \text{Unif}(0, 0.5) & \text{with probability } 0.5 \\ F_{Q_2} \sim \text{Unif}(0.5, 1) & \text{with probability } 0.5 \end{cases}. \tag{3.1}$$

corresponding to two random variables Q_1 and Q_2 that are both uniformly distributed on different supports. In this case $\mathcal{Q}$ would equal Q_1 with 0.5 probability, and similarly for Q_2. This will lead to the distribution of $\mathcal{Q}$ being different from both Q_1 and Q_2. Looking at its distribution, we can first write

$$\begin{aligned} P(\mathcal{Q} \leq q) = P(\mathcal{Q} \leq q \mid \mathcal{F}_{\mathcal{Q}} = F_{Q_1}) P(\mathcal{F}_{\mathcal{Q}} = F_{Q_1}) + \\ P(\mathcal{Q} \leq q \mid \mathcal{F}_{\mathcal{Q}} = F_{Q_2}) P(\mathcal{F}_{\mathcal{Q}} = F_{Q_2}) = \frac{1}{2}\left(F_{Q_1}(q) + F_{Q_2}(q)\right) \end{aligned} \tag{3.2}$$

which with uniform CDFs:

$$F_{Q_1}(q) = 2q\mathbb{1}(q \in [0, 0.5)) + \mathbb{1}(q \geq 0.5) \tag{3.3}$$

$$F_{Q_2}(q) = (2q - 1)\mathbb{1}(q \in [0.5, 1)) + \mathbb{1}(q \geq 1) \tag{3.4}$$

for indicator functions $\mathbb{1}(\cdot)$ yields

$$P(\mathcal{Q} \leq q) = q\mathbb{1}(q \in [0, 1)) + \mathbb{1}(q \geq 1). \tag{3.5}$$

This shows that $\mathcal{Q} \sim \text{Unif}(0, 1)$ while $Q_1 \sim \text{Unif}(0, 0.5)$ and $Q_2 \sim \text{Unif}(0.5, 1)$.

Definition 7. A DM method, $\mathcal{F}$, is a distribution generator that produces distributions that aim to estimate the distributions of a question generator, $\mathcal{F}_{\mathcal{Q}}$.

We introduce this distinction between DM methods and question generators because the distributions generated by DM methods while aiming for, do not necessarily need to produce probabilities that are connected to any measurable quantities. We denote with $\mathcal{F}_Q$ the question generator that represents potential questions of interest and $\mathcal{F}$ the DM method that generates DMs for these questions.

Definition 8. A DM method, $\mathcal{F}$, is **perfect** if $\mathcal{F}$ and $\mathcal{F}_Q$ are equal. That is, they always produce the exact same CDF distributions as realizations.

3.2 Error Measures

When evaluating DMs in practice, we often face the challenge of having only a single realization from each of many heterogeneous questions. In ground truth simulations where the target density is known or sufficient observed realizations are available, we could employ established metrics such as the Kullback-Leibler divergence or the

Wasserstein distance to measure approximation quality directly. However, in many forecasting contexts, we must develop evaluation metrics suitable for this limited observation setting.

In practice, we work with a finite set of N questions that represent realizations from a question generator $\mathcal{F}_{\mathcal{Q}}$. That is, we have questions $Q_1, \ldots, Q_N$ with target distributions $F_{Q_1}, \ldots, F_{Q_N}$ that are realizations of $\mathcal{F}_{\mathcal{Q}}$, and corresponding observed outcomes $q_1, \ldots, q_N$ that are realizations of $\mathcal{Q}$. While the question generator framework allows us to theoretically discuss the space of all questions that could be asked to a DM method, for any given empirical study we observe these concepts realized to some finite set of specific questions and their outcomes.

The questions generally correspond to different target distributions and are not identically distributed, reflecting the diverse nature of real-world forecasting scenarios. For each question i, we have DM distributions F_i^j produced by method j, representing that method's predicted distribution for the question.

3.2.1 Point Estimate Error Metrics

To evaluate DM performance, we can employ several error metrics that capture different aspects of distributional accuracy. These metrics focus on how well the DMs' point estimates align with the realized values.

The linear error for question i and DM method j is defined as the absolute deviation between the predicted median and the realized value:

$$|\text{Median}(F_i^j) - q_i|$$

Similarly, the squared error measures the deviation using the predicted mean:

$$(\mathbb{E}[F_i^j] - q_i)^2$$

where $\mathbb{E}[F_i^j]$ represents the expected value under the DM's distribution.

The theoretical foundation for these metrics stems from statistics. The median minimizes expected linear loss, while the mean minimizes expected squared loss [31]. This provides a principled basis for evaluation, if DM method j_1 consistently produces distributions whose medians align with the target distribution's median better than method j_2, then method j_1 will demonstrate lower expected linear loss. By computing mean linear or squared errors across multiple questions, we can assess which DM methods most effectively capture the central tendencies of their target distributions.

It is important to note that even a perfect DM method producing the correct target distribution can have non-zero expected error since realized values are random draws rather than deterministic central parameters.

3.2.2 Scale-Invariant Error Metrics

Because questions sometimes vary considerably in scale, the linear error metrics might be misleading when comparing performance across different question types. Relative error is a common alternative metric that provides better scale invariance:

$$\frac{|\text{Median}(F_i^j) - q_i|}{q_i}$$

While relative error does not have a well-known theoretic minimizer like the linear and squared error metrics, we use the median in this formulation for consistency with the linear error approach.

3.2.3 Aggregate Error Metrics

To summarize these evaluation approaches across multiple questions, we define several aggregate metrics. For DM method j evaluated over N questions, we define the Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE) as:

$$\text{MAE}^{\text{Median},j} = \frac{1}{N}\sum_{i=1}^{N} |\text{Median}(F_i^j) - q_i| \tag{3.6}$$

$$\text{MAPE}^{\text{Median},j} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\text{Median}(F_i^j) - q_i|}{q_i} \tag{3.7}$$

$$\text{RMSE}^{\text{Mean},j} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (E[F_i^j] - q_i)^2} \tag{3.8}$$

For the first two metrics, we also define their median analogs to provide robust alternatives less sensitive to outliers:

$$\text{MedAE}^{\text{Median},j} = \text{Median}_{i=1,\ldots,N}\left\{\left|\text{Median}(F_i^j) - q_i\right|\right\} \tag{3.9}$$

$$\text{MedAPE}^{\text{Median},j} = \text{Median}_{i=1,\ldots,N}\left\{\frac{\left|\text{Median}(F_i^j) - q_i\right|}{q_i}\right\} \tag{3.10}$$

Despite their utility, these point-estimate evaluations capture only limited aspects of distributional quality. They assess a DM method's ability to locate the center of the target distribution but do not capture the appropriateness of the distribution's shape, spread, or higher-order moments. To address part of these limitations, calibration measures will complement these point-estimate evaluations to also capture over- and under-confidence.

3.3 Calibration Metrics

Calibration metrics exist in various forecasting contexts and tend to capture whether claimed DM probabilities correspond to frequencies of observable quantities [1], [32]. In the CM, for example, this is done by examining specific inter-quantile ranges and measuring the deviation between claimed probabilities and observed frequencies. Here we introduce novel definitions to define calibration for all quantiles and for continuous random variables.

Definition 9. A DM, F, is **calibrated** if $P(Q \leq F^{-1}(p)) = p$ for all $p \in [0, 1]$.

For individual DMs, calibration is equivalent to being perfect, as shown in the following lemma.

Lemma 1. A DM is perfect if and only if it is calibrated.

Proof. ($\rightarrow$) : If the DM is perfect we have $F = F_Q$ and thus

$$P(Q \leq F^{-1}(p)) = P(Q \leq F_Q^{-1}(p)) = F_Q(F_Q^{-1}(p)) = p.$$

($\leftarrow$) :If the DM is calibrated let $U = F(Q)$ and we have

$$P(Q \leq F^{-1}(p)) = P(F(Q) \leq p) = P(U \leq p) = p,$$

thus the variable U is uniformly distributed.Then from the inverse probability integral Theorem 3, Q is distributed according to F, which means that $F = F_Q$. □

While this definition of calibration is equivalent to perfection for DMs, this is not the case for DM methods with the following calibration definitions.

Definition 10. A DM method, $\mathcal{F}$, is **calibrated for the p-quantile** if

$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) = p.$$

In particular, a DM method is calibrated for the median if it is calibrated for the 0.5-quantile.

Definition 11. A DM method, $\mathcal{F}$, is **calibrated** if all its p-quantiles are calibrated.

This last definition captures that for any probability p, the DM method produces a random quantile $\mathcal{F}^{-1}(p)$, such that the random question to be asked $\mathcal{Q}$ is below this quantile with probability p. The first definition of a specific calibrated p-quantile can be of particular interest when performing predictions using that quantile; for example, if one uses the median for predictions, then the calibration of the 0.5-quantile may be of extra interest.

The distinction between calibration and perfection for DM methods is made clear with the following lemma.

Lemma 2. If a DM method is perfect then it is also calibrated. If a DM method is calibrated, it is not necessarily perfect.

Proof. For the first statement we have from perfection that $\mathcal{F} = \mathcal{F}_{\mathcal{Q}}$ which leads to

$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) = P(\mathcal{Q} \leq \mathcal{F}_{\mathcal{Q}}^{-1}(p)) = P(\mathcal{F}_Q(\mathcal{Q}) \leq p)$$

Now for any realization of the question generator, we have from the probability integral theorem that $U = \mathcal{F}_Q(\mathcal{Q})$ is uniformly distributed. Thus

$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) = P(U \leq p) = p.$$

We show the second statement with a counter-example.

Let $\mathcal{F}_{\mathcal{Q}}(q)$ be the same question generator as in the example of Eq. (3.1). Let $\mathcal{F}$ be a DM method that is a constant uniform distribution

$$\mathcal{F} \sim \text{Unif}(0, 1).$$

This DM method is clearly not perfect with $P(\mathcal{F} = \mathcal{F}_Q) = 0$. From the example before we showed that $\mathcal{Q} \sim \text{Unif}(0, 1)$ which yields

$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) = P(\mathcal{Q} \leq p) = p. \tag{3.11}$$

Thus showing that this DM method is calibrated while not being perfect. □

This lemma shows both the advantages and limitations of looking at a DM method's calibration. While calibration is not sufficient for perfection, it is a necessary condition for perfect DM methods. It is also an interpretable measure in the sense that it measures how well the quantiles of a DM method match the probabilities of observing realizations below these quantiles for a certain class of questions.

To make testing for calibration easier, we introduce the following lemma:

Lemma 3. If a DM method $\mathcal{F}$ is calibrated, then $\mathcal{F}(\mathcal{Q})$ is uniformly distributed on $(0, 1)$.

Proof. Let U be the random variable $\mathcal{F}(\mathcal{Q})$. From the definition of calibration, we then get

$$P(U \leq u) = P(\mathcal{F}(\mathcal{Q}) \leq u) = P(\mathcal{Q} \leq \mathcal{F}^{-1}(u)) = u; \quad 0 \leq u \leq 1$$

□

With this lemma as motivation, we introduce calibration testing procedures. Even though calibration is not a sufficient criterion for perfect DM methods, it is still useful to measure since it is a required condition for perfect DM methods, which is what we ultimately want.

3.3.1 Practical Calibration Testing

To evaluate the full distributional properties of DM methods beyond central tendency measures, we introduce two descriptive metrics that measure how well a DM method satisfies the calibration criterion.

For a finite set of questions with realizations $q_1, \ldots, q_N$ and a DM method that has produced DMs $F_1, \ldots, F_N$ for these questions, we define *calibration quantities* u_i as $u_i = F_i(q_i)$ for $i = 1, \ldots, N$. From Lemma 3, we know that u_i are realizations from a uniform random variable if the DM method is calibrated. We therefore define two L^p metrics that measure the deviation between the Empirical Cumulative Distribution Function (ECDF), $\hat{F}$, of the u_i realizations and the uniform CDF:

$$L^1_{\text{Unif}} = \int_0^1 \left|\hat{F}(u) - u\right| du \tag{3.12}$$

$$L^\infty_{\text{Unif}} = \max_{u \in [0,1]} |\hat{F}(u) - u| \tag{3.13}$$

where we calculate the ECDF $\hat{F}(u)$ as

$$\hat{F}(u) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(u_i \leq u).$$

The L^1_{Unif} metric can be interpreted as the average deviation of claimed p-quantiles between the empirical probability and the p probability they represent. L^∞_{Unif}, on the other hand, represents the maximum deviation between the empirical probability and the claimed probability of any p-quantile. The L^1_{Unif} metric captures information about the entire shape of predicted distributions, including their spread and higher-order moments. In both cases, lower values of these metrics are evidence for better calibrated DM methods.

The use of the ECDF and the uniform CDF also lend themselves well for visual investigation by plotting the two distributions side by side. This will be done in the result chapter when comparing different DMs such as in Fig. 5.6.

3.3.2 Overconfidence and Underconfidence

We conclude this chapter with some qualitative calibration definitions that will aid us in describing different types of calibration behavior.

Definition 12. A DM method, $\mathcal{F}$, **overestimates** the p-quantile of a question generator, $\mathcal{F}_{\mathcal{Q}}$, if

$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) > p.$$

and **underestimates** the p-quantile if

$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) < p.$$

This wording is motivated by the fact that if a DM method overestimates a certain p-quantile, then the method tends to estimate the p-quantile as larger than the real question p-quantiles. This in turn leads to the probability of a question being smaller than estimate quantiles being larger than optimal.

Definition 13. A DM method $\mathcal{F}$, is **overconfident** for a question generator $\mathcal{F}_{\mathcal{Q}}$ if it is calibrated for the median but overestimates all p-quantiles for $p < 0.5$ and underestimates all p-quantiles for $p > 0.5$. That is, if

$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) > p \qquad \text{for } p < 0.5 \tag{3.14}$$
$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(0.5)) = 0.5 \qquad \text{for } p = 0.5 \tag{3.15}$$
$$P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) < p \qquad \text{for } p > 0.5 \tag{3.16}$$

Similarly, a method is **underconfident** if it is calibrated for the median but underestimates all p-quantiles for $p < 0.5$ and overestimates all p-quantiles for $p > 0.5$.

The overconfident behavior, where lower quantiles are overestimated while upper quantiles are underestimated, translates to the method generally producing distributions that are too narrow. Conversely, an underconfident DM method generally produces distributions that are too wide. A selection of archetypal calibration scenarios is shown in Fig. 3.1.

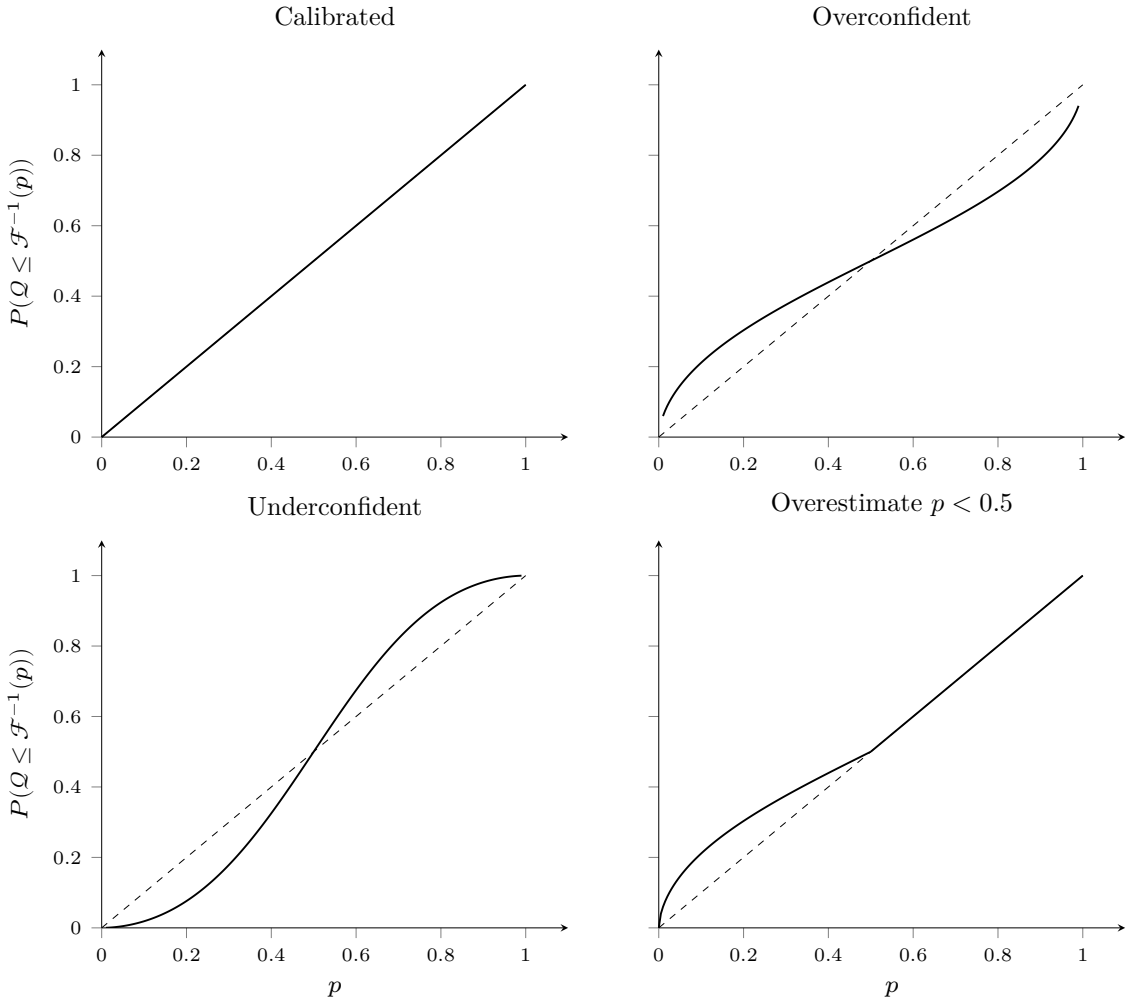


Figure 3.1: Illustration of some different archetypal DM method calibration behaviors. The dashed line represents perfect calibration where $P(\mathcal{Q} \leq \mathcal{F}^{-1}(p)) = p$.

Chapter 4

The Decoupled Copula Model

The JC model, described in Section 2.3.2, captures possible dependencies between experts via the median error random variable. However, there appears to be no particular argument for why the median error would be the most appropriate scalar to capture this dependence, nor any obvious advantage of a linear error metric over alternatives such as relative error. For both linear and relative metrics, we would expect realizations to be focused around zero when the expert's belief is close to the target distribution. However, this zero-focused property is not generally required, and we will see an example of a non-zero-focused metric later in this chapter. This generalized choice of metric we call the *decoupled random variable* and it is induced by a *decoupling function*. Much like how the median error random variable from the JC model was induced by the linear difference between the median and the quantity of interest.

In addition to providing a more flexible metric to model expert dependencies, we also study methods for estimating the distribution of the decoupled random variable, which leads us to a framework where individual expert biases can be detected and adjusted for. The chapter is structured to first establish the theoretical foundation of our model, then explore the range of parameter choices available for practical implementation, and finally describe the model's relation to existing models.

4.1 The Decoupled Copula Model

Let Q, f^e and cdf F^e be defined as in Section 2.3.2. However, in order to allow modeling dependencies between experts, we will see cdf F^e (and indirectly also f^e) as an outcome of a random variable (random function) $\mathcal{F}^e$. More explicitly, if Ω is the sample space, $\mathcal{F}^e$ maps to a continuous CDF, $\mathcal{F}^e : \Omega \to C(\mathbb{R})$. This way, experts being dependent on each other corresponds to their random variables, let us say $\mathcal{F}^{e_1}$ and $\mathcal{F}^{e_2}$, being dependent. We denote the joint random variable of these functions as $\boldsymbol{\mathcal{F}} = (\mathcal{F}^1, \dots, \mathcal{F}^E)$ having a realization $\mathbf{F} = (F^1, \dots, F^E)$.

Given these distributions, we are ultimately interested in the density $f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q)$ that describes the aggregated belief in Q given the beliefs. To later be able to use test questions to estimate this density, however, we will transform the beliefs and Q into a random variable that potentially can be invariant of question scale and type. That random variable is the decoupled random variable (matrix), $\mathbf{Z}$, and is induced

by a decoupling function, $\phi(\cdot,\cdot)$, through

$$\mathbf{Z} = \begin{bmatrix} Z_1^1 & \dots & Z_D^1 \\ \vdots & \ddots & \vdots \\ Z_1^E & \dots & Z_D^E \end{bmatrix} = \phi(Q, \boldsymbol{\mathcal{F}}) = \begin{bmatrix} \phi_1(Q, \mathcal{F}^1) & \dots & \phi_D(Q, \mathcal{F}^1) \\ \vdots & \ddots & \vdots \\ \phi_1(Q, \mathcal{F}^E) & \dots & \phi_D(Q, \mathcal{F}^E) \end{bmatrix}. \tag{4.1}$$

The decoupling function will also be referred to as simply the *decoupler*.

The number of columns, D, represents D transformed features per expert distribution that are extracted by ϕ_d for $d = 1, \dots, D$ and we call ϕ_d the component functions. Examples of features could be the mean, variance, or specific quantiles of the belief. We emphasize that we will use the same notation to denote the decoupling function ϕ that is applied to random variables $\phi(Q, \boldsymbol{\mathcal{F}})$ and to non-random arguments $\phi(q, \mathbf{F})$. Furthermore, it will sometimes be notationally convenient to place the second argument as a subscript writing $\phi_{\mathbf{F}}(q)$ and $\phi_{d,F^e}(q)$ respectively. We assume that each component function $\phi_{d,F^e}(q)$ is invertible and differentiable with respect to q, allowing us to define the inverse with respect to q as $\phi^{-1}_{d,F^e}(z)$.

Example (Decoupling Functions): A list of decoupling functions is presented in Section 4.3.1 but two examples are the linear three-quantile decoupler and the CDF decoupler defined as

$$\phi_{\text{Lin.3Q}}(q, \mathbf{F}) = \begin{bmatrix} q - m^1_{5\%} & q - m^1_{50\%} & q - m^1_{95\%} \\ \vdots & \vdots & \vdots \\ q - m^E_{5\%} & q - m^E_{50\%} & q - m^E_{95\%} \end{bmatrix}, \quad \phi_{\text{CDF}}(q, \mathbf{F}) = \begin{bmatrix} F^1(q) \\ \vdots \\ F^E(q) \end{bmatrix}$$

respectively, where $m^e_{p\%}$ is the p-percentile of expert e and $F^e(q)$ is the CDF of expert e evaluated at q. With these choices, we see how the column dimension is dependent on the choice of decoupler while the number of rows does not. These decouplers are further discussed in Section 4.3.1.

We restrict the domain of the q argument in ϕ to the support of Q and then refer to the image of $\phi_{\mathbf{F}}$ as $\Gamma_{\mathbf{F}}$

$$\Gamma_{\mathbf{F}} = \{\phi(q, \mathbf{F}) : q \in \text{support}\,(Q)\}.$$

With this, we can, in addition to each component being invertible, require that the overall ϕ function be invertible between support(Q) and $\Gamma_{\mathbf{F}}$. We denote this inverse function as $\phi^{-1}_{\mathbf{F}}(\mathbf{z})$.

Note that this $\Gamma_{\mathbf{F}}$ is a 1-dimensional curve existing in a $D \times E$-dimensional space. We denote a segment of $\Gamma_{\mathbf{F}}$ that is parameterized by a segment of q values by

$$\Gamma_{\mathbf{F}}[a, b] = \{\phi(q, \mathbf{F}) : q \in [a, b]\}$$

where $a < b$ and $a, b \in \text{support}\,(Q)$. We will generally assume compact support of Q for simplicity and because it is realistic for practical SEJ studies.

We are now equipped with the foundational notation and transformation setup needed to relate expert beliefs to the decoupled representation. The goal is to relate the unconditional density of $\mathbf{Z}$ to the density of Q conditional on $\boldsymbol{\mathcal{F}} = \mathbf{F}$. We are particularly interested in the unconditional density of $\mathbf{Z}$ because it is generally more tractable to estimate from test questions than the density of $\mathbf{Z}$ conditional on $\boldsymbol{\mathcal{F}} = \mathbf{F}$.

Definition 14. We call a function $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}} : \Gamma_{\mathbf{F}} \to \mathbb{R}$ a density of $\mathbf{Z}$ conditional on $\boldsymbol{\mathcal{F}} = \mathbf{F}$ if

$$P(\mathbf{Z} \in \Gamma_{\mathbf{F}}[a,b] \mid \boldsymbol{\mathcal{F}} = \mathbf{F}) = \int_{\Gamma_{\mathbf{F}}[a,b]} f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\mathbf{z}) d\mathbf{z}. \tag{4.2}$$

where the integral is a line integral.

Although not a density with respect to the Lebesgue measure, $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}$ would, in a measure-theoretic framework, be a density with respect to the 1D-Hausdorff measure which in this setting would measure the length of the curve $\Gamma_{\mathbf{F}}[a,b]$ in the $\mathbb{R}^{DE}$ space $\mathbf{Z}$ is in. We will not discuss this theoretical measure theory further, but it is with respect to this background that we will still refer to $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}$ as a conditional density.

Lemma 4. A density, $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}$, of $\mathbf{Z}$ conditional on $\boldsymbol{\mathcal{F}} = \mathbf{F}$ is given by

$$f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\mathbf{z}) = f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\phi_{\mathbf{F}}^{-1}(\mathbf{z})) \frac{1}{\left\|\phi'_{\mathbf{F}}(\phi_{\mathbf{F}}^{-1}(\mathbf{z}))\right\|}, \quad \forall \mathbf{z} \in \Gamma_{\mathbf{F}} \tag{4.3}$$

where $\phi'_{\mathbf{F}}(q)$ is the matrix with elements $\frac{\partial}{\partial q}\phi_d(q, F^e)$ and $\|\phi'_{\mathbf{F}}(q)\|$ is the Euclidian norm over a vector containing the matrix elements of $\phi'_{\mathbf{F}}(q)$.

Proof. Note first that Eq. (4.3) can be written as $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}$ fulfilling

$$f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}\left(\phi_{\mathbf{F}}(q)\right) \|\phi'_{\mathbf{F}}(q)\| = f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q) \tag{4.4}$$

by choosing $\mathbf{z}$ as $\phi_{\mathbf{F}}(q)$. Considering the left-hand side (LHS) of Eq. (4.2) we have

$$\text{LHS} = P(\mathbf{Z} \in \Gamma_{\mathbf{F}}[a,b] \mid \boldsymbol{\mathcal{F}} = \mathbf{F}) = P(Q \in [a,b] \mid \boldsymbol{\mathcal{F}} = \mathbf{F}) = \int_a^b f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q) dq.$$

Then from the definition of line integrals, the right-hand side (RHS) becomes

$$\text{RHS} = \int_{\Gamma_{\mathbf{F}}[a,b]} f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\mathbf{z}) d\mathbf{z} = \int_a^b f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}\left(\phi_{\mathbf{F}}(q)\right) \|\phi'_{\mathbf{F}}(q)\| dq$$

Now from Eq. (4.4) we see that also the RHS is equal to

$$\text{RHS} = \int_a^b f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q) dq$$

which concludes that the the density of Eq. (4.3) fulfills the definition. □

With the connection between the two conditional densities made, we will now relate the conditional density of the decoupled random variable to the unconditional density. We do so with the following assumption.

Assumption 1 (Independence assumption)**.** The *independence assumption* is said to hold if the conditional density $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}$ is proportional to the unconditional density $f_{\mathbf{Z}}$ for all points in $\Gamma_{\mathbf{F}}$. That is,

$$f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\mathbf{z}) \propto f_{\mathbf{Z}}(\mathbf{z}), \quad \forall \mathbf{z} \in \Gamma_{\mathbf{F}}. \tag{4.5}$$

Note that this is not the same as the random variables $\mathbf{Z}$ and $\boldsymbol{\mathcal{F}}$ being independent. In particular, the support of $\mathbf{Z}$ depends on the value of $\boldsymbol{\mathcal{F}}$, so standard independence does not hold. However, due to the similar density relationship, we retain the term 'independence assumption'.

The independence assumption assumes that the conditional densities follow, up to a constant, the same density as the unconditional density. The truthfulness of this assumption is difficult to test directly since we in general only observe a single set of beliefs $\mathbf{F}$ once, and it is more chosen as arguably the most parsimonious connection between the two densities. Other connections are discussed in the future research section of Chapter 6. While we can not easily test this assumption in isolation we can test it indirectly by seeing if the derived properties of the final model match empirical observations.

Having the conditional density relation and the independence assumption, we arrive at the core result: a tractable expression for the density $f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}$ in terms of copula components and marginal densities in the transformed space of $\mathbf{Z}$.

Theorem 4 (The Decoupled Copula Model)**.** Under the independence Assumption 1, we can write the conditional density $f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}$, as

$$f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q) \propto c_{\mathbf{Z}}\left(F_{Z_1^1}\left(\phi_1(q,F^1)\right),\dots,F_{Z_D^E}\left(\phi_D(q,F^E)\right)\right) \prod_{\substack{d=1,\dots,D\\ e=1,\dots,E}} f_{Z_d^e}(\phi_d(q,F^e))\,\|\phi'_{\mathbf{F}}(q)\| \quad (4.6)$$

where $c_{\mathbf{Z}}(u_1,\dots,u_{DE})$ is the density of the copula for $\mathbf{Z}$.

Proof. From Lemma 4 we have that $f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q) = f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\phi_{\mathbf{F}}(q))\|\phi'_{\mathbf{F}}(q)\|$ which together with the independence assumption yields

$$f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q) \propto f_{\mathbf{Z}}(\phi_{\mathbf{F}}(q))\|\phi'_{\mathbf{F}}(q)\|.$$

Being flexible with the matrix notation and seeing $\mathbf{Z}$ as a vector with DE values, we get from the copula Corollary 1.1 that we can write $f_{\mathbf{Z}}$ as

$$f_{\mathbf{Z}}(z_1^1,\dots,z_D^E) = c_{\mathbf{Z}}\left(F_{Z_1^1}\left(z_1^1\right),\dots,F_{Z_D^E}\left(z_D^E\right)\right)\prod_{\substack{d=1,\dots,D\\ e=1,\dots,E}} f_{Z_d^e}(z_d^e)$$

which yields the desired expression

$$f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}(q) \propto c_{\mathbf{Z}}\left(\{F_{Z_d^e}(\phi_d^e(q,\mathbf{F}))\}_d^e\right)\prod_{\substack{i=1,\dots,D\\ e=1,\dots,E}} f_{Z_d^e}(\phi_d^e(q,\mathbf{F}))\|\phi'_{\mathbf{F}}(q)\|$$

□

The practical point of the transformation of $f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}$ into densities of $\mathbf{Z}$, is that we choose the decoupling function ϕ such that the decoupled space $\mathbf{Z}$ is invariant to the question asked. In doing this, we can use historical questions that the experts have answered, and to which we know the answers, to help us estimate the $\mathbf{Z}$ space. Having estimated $c_{\mathbf{Z}}$ and $F_{Z_d^E}, f_{Z_d^e}$ we can use these densities in Eq. (4.6) with $\mathbf{F}$ from a new question, to yield the $f_{Q|\boldsymbol{\mathcal{F}}=\mathcal{F}}$ to this new question we do not know the answer to.

In order to estimate the copula and marginal distributions of $\mathbf{Z}$ from previous observations, however, we need the distribution to be invariant to the question asked. That is, while the distributions of Q and $\boldsymbol{\mathcal{F}}$ clearly differ between questions (otherwise the questions would be essentially identical), we will here assume that with the transformation to $\mathbf{Z}$, the connection is 'decoupled'.

Assumption 2 (Question invariance assumption)**.** The *question invariance assumption* is said to hold if there exists a common distribution $F_{\mathbf{Z}}$ such that

$$\mathbf{Z}_i = \phi(Q_i, \boldsymbol{\mathcal{F}}_i) \sim F_{\mathbf{Z}} \quad \forall i = 1, \dots, N \tag{4.7}$$

for N questions $Q_1, \dots, Q_N$ and respective N expert belief realizations $\boldsymbol{\mathcal{F}}_1, \dots, \boldsymbol{\mathcal{F}}_N$.

With this, we have an extra constraint on ϕ that should be chosen such that the question invariance assumption also holds. Equation (4.6) together with Assumption 1 and Assumption 2 is the decoupled copula model.

While the core theoretical framework of the model now is established, multiple implementation questions remain:

- How do we choose the decoupling function ϕ?
- How do we estimate the marginal densities $f_{Z_d^e}$ of $\mathbf{Z}$?
- How do we estimate the copula $c_{\mathbf{Z}}$ of $\mathbf{Z}$?

These questions form the core of making the decoupled copula model practically applicable. The choice of decoupling function must balance theoretical soundness with empirical validity, ensuring that the transformed variables satisfy our assumptions while being estimable from limited historical data. These questions will be addressed in Section 4.3 but first we will look at a a method of strengthening the question independence assumption by rejecting experts prior to applying the model.

4.2 Expert Rejection as Preprocessing

Implicit in all expert judgment studies, is the selection of experts to elicit. At one end of the spectra, there is 'wisdom of the crowd' where experts can be random members of the public [33] and on the other end they can be carefully selected experts with specific domain expertise [34]. In addition to the selection of experts to elicit, we can also further narrow down which experts' assessments to use, post-elicitation. Among existing expert judgment models, this is for example done in the CM where an expert can be rejected[1] if their assessments on the test questions are not deemed sufficiently calibrated [35]. In contrast to the CM that rejects based on calibration, however, we will reject experts based on the question invariance assumption. Preferably, we would reject experts based upon both the independence and the question invariance assumption, but because the independence assumption is difficult to test in practice we limit ourselves to testing the question invariance assumption. This will be done through the following two lemmas.

[1]The rejected expert's assessments are, however, still used in the computation of the support of the question of interest.

Lemma 5. If the question invariance assumption holds, then $\mathbf{Z}$ is independent of Q.

Proof. Since the assumption should hold for any Q_i with arbitrary distribution, it can also be phrased as

$$\mathbf{Z} = \phi(Q, \mathcal{F}) \sim F_{\mathbf{Z}}, \quad \forall F_Q.$$

Generally we can write the pdf, $f_{\mathbf{Z}}$, of $\mathbf{Z}$ in terms of the conditional distribution as

$$f_{\mathbf{Z}}(\mathbf{z}) = \int f_{\mathbf{Z}|Q=s}(\mathbf{z}) f_Q(s) ds, \quad \forall \mathbf{z}$$

With this holding for all distributions, we inspect the case where f_Q is the Dirac delta distribution at a point q, $f_Q(s) = \delta_q(s)$. The previous integral then becomes

$$f_{\mathbf{Z}}(\mathbf{z}) = f_{\mathbf{Z}|Q=q}(\mathbf{z}), \quad \forall \mathbf{z}, q$$

which shows the desired independence. □

Lemma 6. If the question invariance assumption holds, then the marginal distributions, $\mathbf{Z}^e = (Z_1^e, \dots, Z_D^e)$, of $\mathbf{Z}$ are independent of Q.

Proof. From Lemma 5, we know that $\mathbf{Z}$ is independent of Q. Given any expert e, let $\mathbf{Z}^c$ be the complement of random variables $\mathbf{Z}^e$, $\mathbf{Z}^c = \mathbf{Z} \setminus \mathbf{Z}^e$. Then we can write the marginal distribution of $\mathbf{Z}^e$ as

$$f_{\mathbf{Z}^e|Q=q}(\mathbf{z}^e) = \int f_{\mathbf{Z}|Q=q}(\mathbf{z}) d\mathbf{z}^c = \int f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}^c = f_{\mathbf{Z}^e}(\mathbf{z}^e) \tag{4.8}$$

□

Because of how the question invariance assumption implies that the decoupled expert marginals $\mathbf{Z}^e$ are independent of Q, we have that their independence to Q also is a requirement for the question invariance assumption. With this in mind, we propose a rejection procedure to remove experts whose marginals $\mathbf{Z}^e$ do not seem independent of Q, with the aim for the whole join distribution of $\mathbf{Z}$ to better satisfy the question invariance assumption.

The expert rejection procedure involves the following steps:

1. We use the distance correlation test in Section 2.6.5 to test the independence of $\mathbf{Z}^e$ and Q.
2. If the test fails for some significance level α_{Rej}, we exclude their beliefs from the $\mathbf{F}$ realization and perform the chosen aggregation method on the remaining experts.
3. If all experts are rejected, we consider only the expert with the highest p-value from the distance correlation test.

This constitutes a pre-processing step that can be performed before the aggregation process. In this work, we use the historically common threshold $\alpha_{Rej} = 0.05$, but this could be adjusted to the context of the study. For example, it could be adaptive to the number of test questions asked, or one could choose the threshold to optimize some performance metric measured on a set of test questions in a similar way to how the optimized significance threshold variant of the CM does it in Section 2.3.1.

The impact of this expert rejection procedure on the question invariance assumption is measured empirically in Section 5.1.1.

4.3 Parameters of the Decoupled Copula Model

The implementation of the decoupled copula model requires specifying three key components: the decoupling function ϕ, the marginal densities $f_{Z_d^e}$, and the copula $c_{\mathbf{Z}}$. Each choice involves multiple degrees of freedom and affects how well the model satisfies the independence and question invariance assumptions. We first examine suitable choices for the decoupling function, evaluating different transformations against theoretical requirements and empirical evidence. We then explore methods for estimating the marginal distributions of the decoupled variables, comparing approaches that balance theoretical soundness with practical robustness in small-sample settings. Finally, we investigate copula estimation strategies, ranging from independence assumptions to more flexible parametric families that can capture expert dependencies while remaining tractable for implementation.

4.3.1 Choices of Decoupling Function ϕ

While the theory in Section 4.1 includes the belief functions $\mathbf{F}$, many SEJ studies, such as all 49 studies in our dataset [14], focus on the quantiles of the experts' beliefs. When choosing the decoupling function, we need to follow these practical constraints imposed by the quantile-based elicitation format. While many decoupling functions could be imagined, we identify in this section two classes of decouplers: the CDF decoupler and several error-based decouplers.

The CDF Decoupler

In general, we define the CDF decoupler as

$$\phi_{\text{CDF}}(q, \mathbf{F}) = \begin{bmatrix} F^1(q) \\ \vdots \\ F^E(q) \end{bmatrix} \in \mathbb{R}^{E\times 1}. \tag{4.9}$$

In cases we only have access to quantile assessments, we first estimate the distributions F^e as described in Section 2.2, and then use these estimates $\hat{F}^e$, instead of F^e in the formula above.

Error-Based Decouplers

For error-based decouplers, we first define error functions: linear error, scaled linear error, and relative error, for a scalar property m of F^e as

$$\epsilon_{\text{Lin.}}(q, m) = q - m \tag{4.10}$$

$$\epsilon_{\text{Sc.Lin.}}(q, m) = \frac{q - m}{U^{e*} - L^{e*}} \tag{4.11}$$

$$\epsilon_{\text{Rel.}}(q, m) = \frac{q - m}{|m| + \varepsilon} \tag{4.12}$$

where ε is a small positive constant to avoid division by zero and ensure invertibility with respect to q for all choices of m. For this work, we set $\varepsilon = 10^{-3}$. We use the expert-specific support parameters L^{e*} and U^{e*} because, as explained in Section 4.3.2, we want the denominator to be proportional to the standard deviation and we argue

that the global support parameters L^* and U^* are more of a measure of the experts uncertainty than the standard deviation of the quantity.

If $\mathbf{m} \in \mathbb{R}^{E\times D}$ is a matrix with elements m_d^e with D properties of each experts' beliefs, such as quantiles or means, we can use the above error functions to define decoupling functions of the form:

$$\phi(q, \mathbf{F}) = \begin{bmatrix} \epsilon(q, m_1^1) & \dots & \epsilon(q, m_D^1) \\ \vdots & \ddots & \vdots \\ \epsilon(q, m_1^E) & \dots & \epsilon(q, m_D^E) \end{bmatrix} \in \mathbb{R}^{E\times D} \tag{4.13}$$

The choices for properties $\mathbf{m}$ that we investigate are:

$$\mathbf{m} = \begin{bmatrix} \mu^1_{\text{G.Sup.}} \\ \vdots \\ \mu^E_{\text{G.Sup.}} \end{bmatrix} \in \mathbb{R}^{E\times 1}$$

where $\mu^e_{\text{G.Sup.}}$ is the global support mean of expert e, calculated from $\hat{F}^e$ using global support parameters as defined in Section 2.2.

$$\mathbf{m} = \begin{bmatrix} \mu^1_{\text{E.Sup.}} \\ \vdots \\ \mu^E_{\text{E.Sup.}} \end{bmatrix} \in \mathbb{R}^{E\times 1}$$

where $\mu^e_{\text{E.Sup.}}$ is the expert support mean of expert e, calculated from $\hat{F}^e_{\text{E.Sup.}}$ using expert-specific support parameters as defined in Section 2.2.

$$\mathbf{m} = \begin{bmatrix} \text{Median}^1 \\ \vdots \\ \text{Median}^E \end{bmatrix} \in \mathbb{R}^{E\times 1}$$

where Median^e is the median of expert e, calculated as $(F^e)^{-1}(0.5)$.

$$\mathbf{m} = \begin{bmatrix} (F^1)^{-1}(5\%) & (F^1)^{-1}(50\%) & (F^1)^{-1}(95\%) \\ \vdots & \vdots & \vdots \\ (F^E)^{-1}(5\%) & (F^E)^{-1}(50\%) & (F^E)^{-1}(95\%) \end{bmatrix} \in \mathbb{R}^{E\times 3}$$

where the matrix contains the 5%, 50%, and 95% quantiles for each expert e, directly corresponding to the elicited quantile information from our dataset.

In the case of $\mathbf{m}$ existing of quantiles, we can calculate the means through

$$\mu^e = \int q\hat{f}^e(q)dq = \frac{1}{2}\sum_{d=1}^{D+1}(p_d - p_{d-1})(m_d^e - m_{d-1}^e)$$

where $m_0^e = L^*$ and $m_{D+1}^e = U^*$ (or L^{e*} and U^{e*} for the expert support variant) for all e, and p_d is the probability for quantiles m_d^e. L^* and U^* are the lower and upper support limits as defined in Section 2.2. A derivation for this result can be seen in Appendix A.2.

These property choices, combined with the three error functions, give rise to the following error-based decoupling functions:

1. $\phi_{\text{Lin.}\mu_{\text{G.Sup.}}}$: Linear error with global support means

2. $\phi_{\text{Lin.}\mu_{\text{E.Sup.}}}$: Linear error with expert support means
3. $\phi_{\text{Lin.Median}}$: Linear error with medians
4. $\phi_{\text{Lin.Q3}}$: Linear error with three quantiles
5. $\phi_{\text{Sc.Lin.}\mu_{\text{G.Sup.}}}$: Scaled linear error with global support means
6. $\phi_{\text{Sc.Lin.}\mu_{\text{E.Sup.}}}$: Scaled linear error with expert support means
7. $\phi_{\text{Sc.Lin.Median}}$: Scaled linear error with medians
8. $\phi_{\text{Sc.Lin.Q3}}$: Scaled linear error with three quantiles
9. $\phi_{\text{Rel.}\mu_{\text{G.Sup.}}}$: Relative error with global support means
10. $\phi_{\text{Rel.}\mu_{\text{E.Sup.}}}$: Relative error with expert support means
11. $\phi_{\text{Rel.Median}}$: Relative error with medians
12. $\phi_{\text{Rel.Q3}}$: Relative error with three quantiles

4.3.2 Theoretical arguments for the decoupling functions

We examine the theoretical properties of each decoupling function under the assumption of perfect experts to understand which transformations are most likely to satisfy the question invariance assumption.

CDF Decoupler: The CDF decoupler perfectly fulfills the question invariance assumption if experts are perfect and our estimation of the F^e functions is correct. When an expert is perfect, i.e., $F^e = F_Q$, then $Z_1^e = F_Q(Q)$ and Z_1^e is uniformly distributed by the probability integral transform theorem Theorem 3. This uniform distribution is completely independent of the specific question, making the CDF decoupler theoretically optimal in a sense for question invariance. Note that the dependence between experts becomes trivial in the case of all experts being perfect (they are all essentially the same expert), thus it is sufficient to look at the marginals.

Linear Error Decoupler: For a perfect expert, the mean of the linear error is zero when one of the mean estimations is chosen and correctly aligns with the mean of F^e. However, the variance remains dependent on the question:

$$Z_d^e = Q - \mu^e \implies E[Z_d^e] = E[Q] - \mu^e = 0 \quad \text{and} \quad \text{Var}(Z_d^e) = \text{Var}(Q) \tag{4.14}$$

Since the variance depends on the specific question Q, this violates the question invariance assumption.

Scaled Linear Error Decoupler: The mean of the scaled linear error is zero for a perfect expert. Additionally, if the expert support range is proportional to the standard deviation of F^e, then the variance becomes question-invariant. Assuming $\sigma_Q \propto U^{e*} - L^{e*}$ and $\mu^e = \mu_Q$:

$$Z_d^e = \frac{Q - \mu^e}{U^{e*} - L^{e*}} \implies E[Z_d^e] = \frac{E[Q] - \mu^e}{U^{e*} - L^{e*}} = 0 \quad \text{and} \quad \text{Var}(Z_d^e) = \frac{\text{Var}(Q)}{(U^{e*} - L^{e*})^2} \propto 1 \tag{4.15}$$

This proportionality assumption makes the scaled linear decoupler more promising for question invariance than the unscaled version.

Relative Error Decoupler: Similar to the scaled linear case, the mean of the relative error is zero for a perfect expert. If we assume that the mean of Q is proportional to its standard deviation, then the variance also becomes proportional to a constant, achieving question invariance in both mean and variance.

Performance Predictions: Based on these theoretical arguments, we hypothesize the following performance ranking for satisfying the question invariance assumption:

1. *CDF decoupler*: Best performance, as it can make all moments of the Z_d^e marginals independent of the question

2. *Scaled linear decoupler*: Second best, as the proportionality assumption between support range and standard deviation seems more reasonable than mean-based assumptions

3. *Relative error decoupler*: Third, requiring the additional assumption of proportionality between mean and standard deviation

4. *Linear decoupler*: Worst among error-based methods, as variance remains question-dependent

Property Choice Considerations: Since the median is not a linear function of the distribution, linear error decouplers using medians will not have the property that the expected error is zero for perfect experts. We therefore expect mean-based properties to outperform median-based properties.

Expert Rejection Effects: We expect that performing expert rejection preprocessing as described in Section 4.2 will improve independence by removing marginals that are clearly dependent on the question, thereby enhancing the performance of all decoupling functions.

Dimension of Output: There are arguably fewer arguments to expect that the 3 quantiles error decouplers should be question invariant. They do however have an information advantage in the sense that $I(Q;\boldsymbol{\mathcal{F}}) \geq I(Q;\mathbf{M}_{\text{3Quantiles}}) \geq I(Q;\mathbf{M}_{\text{Median}})$ where I is the mutual information and $\mathbf{M}$ is the random variable of $\mathbf{m}$. This follows from the data processing inequality [36] and because how $\mathbf{M}_{\text{Median}}$ is a function of $\mathbf{M}_{3Quantiles}$, that in turn is a function of $\boldsymbol{\mathcal{F}}$. Intuitively, this can be described as no information can be gained, only potentially lost, by transforming a random variable. The mutual information is relevant because it describes the amount of information obtained about Q when observing, for instance, $\mathbf{M}_{\text{Median}}$. From the inequality above, we can thus hypothesize that we can infer more about the distribution of Q if we use all quantiles instead of only the median. See [36] for a formal definition of mutual information and related results.

These theoretical arguments provide guidance for selecting appropriate decoupling functions. The empirical evaluation of these different decoupling functions and their performance in satisfying the question invariance assumption is presented in Chapter 5.

4.3.3 Composition with sigmoid function

In terms of the independence towards Q, performing an invertible transformation on $\mathbf{Z}$ does not alter this property, see Appendix A.3 for proof. In terms of practically estimating the joint density of $\mathbf{Z}$ from observations, however, transformations can facilitate the process by making the support bounded. To simplify the density estimation part in Section 4.3.4 we introduce composition with sigmoid functions for our error-based decouplers that are not naturally bounded.

For any invertible function $\sigma : \mathbb{R} \to (0, 1)$ we can define the sigmoid composed decoupler function as

$$\phi_\sigma(q, \mathbf{F}) = \begin{bmatrix} \sigma(\phi_1(q, F^1)) & \dots & \sigma(\phi_{\tilde{D}}(q, F^1)) \\ \vdots & \ddots & \vdots \\ \sigma(\phi_1(q, F^E)) & \dots & \sigma(\phi_{\tilde{D}}(q, F^E)) \end{bmatrix},$$

and its derivative wrt q at row e column d becomes

$$\sigma'(\phi_d(q, F^e)) \frac{\partial}{\partial q} \phi_d(q, F^e).$$

For this work, we use the logistic function.

$$\sigma(x; k) = \frac{1}{1 + e^{-kx}}.$$

with derivative

$$\sigma'(x; k) = \frac{ke^{-kx}}{(1 + e^{-kx})^2}.$$

as sigmoid function.

With the support of the CDF decoupler function already being bounded, we apply the sigmoid transformation only to the previously mentioned linear and relative error decoupler functions. The parameter k is chosen to ensure that the range of interest is captured without numerical precision issues.

For error-based decouplers where ϕ_d is an affine function of q (constant derivative) and is composed with the logistic function, the mean and variance of the transformed random variables can be calculated analytically. From Appendix A.2, for a piecewise linear distribution with density represented as quantile interpolation points, the expected value of the sigmoid-transformed variable is:

$$\int \sigma(\phi_d(q, F^e)) f^e(q) dq = \sum_{d=1}^{n} \frac{p_i - p_{d-1}}{(m_i^e - m_{d-1}^e) C_d k} \log \frac{1 + e^{k\phi_d(m_i^e, F^e)}}{1 + e^{k\phi_d(m_{d-1}^e, F^e)}}$$

where C_d is the constant derivative of ϕ_d with respect to q. and the primitive function for the sigmoid composition (from Appendix A.1) enables this analytical computation of moments.

4.3.4 Estimation of Margins

With the decoupler function analyzed in Section 4.3.1, we focus here on estimating the marginal densities $f_{Z_d^e}$ and $F_{Z_d^e}$ of Eq. (4.6). Of particular importance for all these estimations is that we want them to be robust even with few samples, $N \approx 10$, as discussed in Section 2.1.

For this purpose, we will investigate three approaches: MLE, empirical Bayes Maximum a Posteriori (MAP), and a non-data-driven Perfect Expert (PE) prior approach. In all approaches, we assume that Z_d^e can be parameterized as a four-parameter beta distribution. We also define the *exact perfect expert marginals* for future theoretical discussions.

The four-parameter beta distribution is defined by shape parameters $a > 0$, $b > 0$, and support $[c, d]$ where $c < d$. It has density

$$f_{\text{Beta}}(x; a, b, c, d) = \frac{1}{B(a,b)} \frac{1}{d-c} \left(\frac{x-c}{d-c}\right)^{a-1} \left(\frac{d-x}{d-c}\right)^{b-1} \tag{4.16}$$

for $x \in [c, d]$ and $f_{\text{Beta}}(x; a, b, c, d) = 0$ otherwise, where $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$ is the beta function. We write the assumption that Z_d^e is beta distributed as

$$Z_d^e \sim \text{Beta}(a_d^e, b_d^e, c_d^e, d_d^e).$$

or

$$f_{Z_d^e}(z) = f_{\text{Beta}}(z; a_d^e, b_d^e, c_d^e, d_d^e)$$

for parameters $(a_d^e, b_d^e, c_d^e, d_d^e)$. The support parameters c_d^e and d_d^e are determined by taking the support of Q and mapping it to the space of Z_d^e. Let $[L^*, U^*]$ be the support of Q as calculated in Section 2.2, then we define

$$c_d^e = \inf\{\phi_d(q, F^e) : \forall q \in [L^*, U^*]\}$$

and

$$d_d^e = \sup\{\phi_d(q, F^e) : \forall q \in [L^*, U^*]\}$$

Since these support parameters are identical across all estimation methods for a_d^e and b_d^e, we will drop the explicit notation and write $f_{\text{Beta}}(a_d^e, b_d^e)$ for brevity. The three density estimation methods are described below.

MLE estimation

For the MLE approach, we define realizations $z_{d,i}^e$ of the random variable Z_d^e for $i = 1, \ldots, N$ test questions. Note that it is possible to use other test questions as realizations due to the question invariance assumption. The MLE method selects a_d^e, b_d^e as elements from the set

$$\arg\max_{a_d^e, b_d^e} \sum_{i=1}^{N} \log\left(f_{\text{Beta}}(x; a_d^e, b_d^e)\right)$$

MAP estimation

The MAP approach employs a hierarchical setup according to

$$Z_d^e \mid A_d^e = a_d^e, B_d^e = b_d^e \sim Beta(a_d^e, b_d^e, c_d^e, d_d^e)$$

where A_d^e and B_d^e are modeled as independent and marginally log-normal distributed:

$$A_d^e \sim \text{LogNormal}(\mu_{A_d^e}, \sigma_{A_d^e}); \quad B_d^e \sim \text{LogNormal}(\mu_{B_d^e}, \sigma_{B_d^e})$$

where the log-normal density is

$$f_{\text{LogNormal}}(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right)$$

for $x > 0$, where $\mu \in \mathbb{R}$ is the location parameter and $\sigma > 0$ is the scale parameter. The MAP estimation can then be formulated as choosing a_d^e, b_d^e from the set:

$$\arg\max_{a_d^e, b_d^e} \log p(a_d^e, b_d^e \mid z_{d,1}^e \ldots z_{d,N}^e) \propto \log f_{\text{LogNormal}}\left(a_d^e; \mu_{A_d^e}, \sigma_{A_d^e}\right) + \log f_{\text{LogNormal}}\left(b_d^e; \mu_{B_d^e}, \sigma_{B_d^e}\right) + \sum_{i=1}^{N} \log f_{\text{Beta}}(z_{d,i}^e; a_d^e, b_d^e) \quad (4.17)$$

For our purposes, it is more convenient to specify the prior in terms of its mode and actual variance rather than using the standard parameterization. Note that the parameters μ and σ^2 in the standard parameterization are not the mean and variance of the log-normal distribution itself, but rather the mean and variance of the underlying normal distribution of the logarithm of the random variable. We, therefore, establish a change of variables that allows us to specify the desired mode M and variance σ^2_{Prior} of the log-normal distribution, and then derive the corresponding standard parameters μ and σ.

For a log-normal distribution with parameters μ and σ, the mode M, and variance σ^2_{prior}, are given by:

$$M = \exp(\mu - \sigma^2),$$
$$\sigma^2_{\text{prior}} = \exp(2\mu + \sigma^2)(\exp(\sigma^2) - 1).$$

This can be derived by normal probability and algebra relations or one can consult for instance [37]. These relationships lead to a change of variables that allows us to specify the desired mode M and variance σ^2_{Prior} of the log-normal distribution and then derive the corresponding standard parameters μ and σ. Letting $M_{A_d^e}$ and $M_{B_d^e}$ be the mode of the distribution of A_d^e and B_d^e respectively, and σ^2_{prior} be the variance of both distributions, we have that the log-normal distribution parameters are determined through:

$$\sigma^2_{A_d^e} = \log(u_{A_d^e}); \quad \sigma^2_{B_d^e} = \log(u_{B_d^e})$$

$$\mu_{A_d^e} = \log(M_{A_d^e}) + \sigma^2_{A_d^e}; \quad \mu_{B_d^e} = \log(M_{B_d^e}) + \sigma^2_{B_d^e}$$

where $u_{A_d^e}$ and $u_{B_d^e}$ are solutions to the respective nonlinear equations:

$$u^4_{A_d^e} - u^3_{A_d^e} - \frac{\sigma^2_{\text{Prior}}}{M^2_{A_d^e}} = 0; \quad u^4_{B_d^e} - u^3_{B_d^e} - \frac{\sigma^2_{\text{Prior}}}{M^2_{B_d^e}} = 0$$

These nonlinear equations must be solved numerically using root-finding algorithms, as they do not admit closed-form solutions. With $\frac{\sigma^2_{\text{Prior}}}{M^2_{B_d^e}} > 0$, the equation is solved for some $u > 1$. Letting $h(u) = u^4 - u^3$, we have that $h(u)$ is strictly increasing for $u \geq 1$, which makes the equation tractable to solve using, for instance, bisection methods. The solution ensures that both log-normal priors have the specified modes $M_{A_d^e}$ and $M_{B_d^e}$ while maintaining the common prior variance σ^2_{Prior}. An illustration of a selection of log-normal densities using this parameterization is shown in Fig. 4.1.

We choose the mode over the mean or median because when fixing the mean or median and increasing σ_{prior}, part of the density concentrates close to zero, which leads to the selection of smaller parameter values as σ_{prior} increases and does not converge to the MLE estimation, which we are aiming for. This is not the case for the mode parameterization, where the density becomes more flat overall as σ_{prior} increases, as illustrated in the figure.

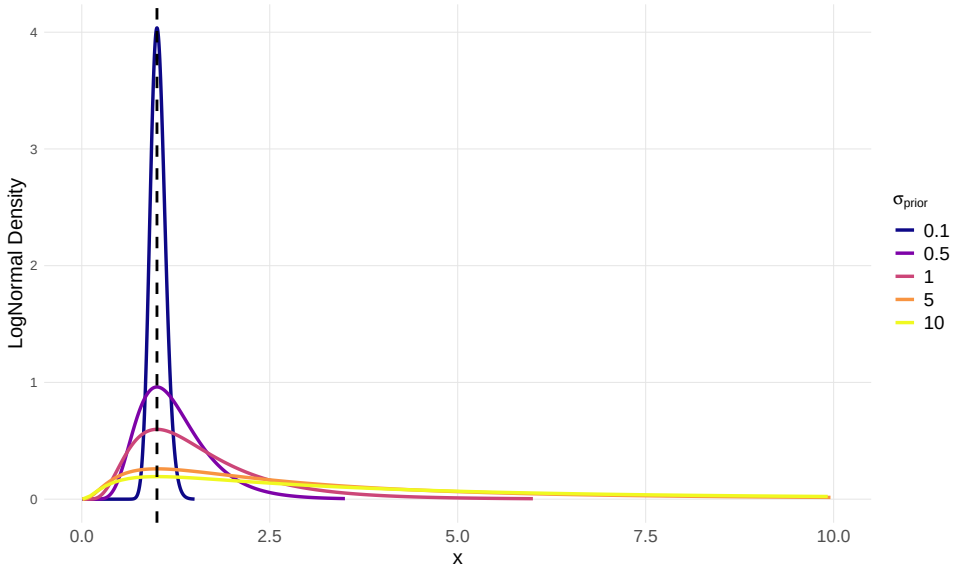


Figure 4.1: Illustration of log-normal densities with specified prior standard deviations σ_{prior} and with mode equal to 1 (dashed line).

With the parameterization of the priors complete, we are left to choose the parameters. We will leave σ_{prior} to be chosen empirically as a hyperparameter specifying the uncertainty of the prior density. For the mode, however, we will choose it somewhat based on the following principle:

Design Principle 1. The prior should be chosen such that, without evidence to the contrary, the distribution of Z_d^e should be that of a perfect expert.

This principle comes from the argument that if we have no other information about an expert, we might as well trust that they are correct. While the distribution of Z_d^e depends on the generally unknown distributions of Q and $\mathcal{F}^e$, this is not the case when an expert is perfect. If an expert is perfect, both the distribution of Q and $\mathcal{F}^e$ become known and identical. The distribution of Q becomes equal to that of the perfect expert, and $\mathcal{F}^e$ is constant at that distribution. This means that we can calculate the distribution of Z_d^e analytically under the perfect expert assumption.

For notational ease we denote the random variable induced by ϕ under the perfect expert assumption as $\tilde{\mathbf{Z}}$ and we call the mean and variance of $\tilde{Z}_d^e$, $\mu_{\tilde{Z}_d^e}$ and $\sigma^2_{\tilde{Z}_d^e}$, respectively. For our operationalization of Design Principle 1, we choose the prior modes $M_{A_d^e}$ and $M_{B_d^e}$ such that when conditioning on these modes, the mean of Z_d^e equals that of $\tilde{Z}_d^e$:

$$E[\tilde{Z}_d^e | A_d^e = M_{A_d^e}, B_d^e = M_{B_d^e}] = \frac{M_{A_d^e}}{M_{A_d^e} + M_{B_d^e}} := \mu_{\tilde{Z}_d^e}$$

$$\mathrm{Var}(\tilde{Z}_d^e | A_d^e = M_{A_d^e}, B_d^e = M_{B_d^e}) = \frac{M_{A_d^e} M_{B_d^e}}{(M_{A_d^e} + M_{B_d^e})^2 (M_{A_d^e} + M_{B_d^e} + 1)} := \sigma^2_{\tilde{Z}_d^e}$$

Solving this system yields:

$$M_{A_d^e} = \mu_{\tilde{Z}_d^e} \left(\frac{\mu_{\tilde{Z}_d^e}(1 - \mu_{\tilde{Z}_d^e})}{\sigma^2_{\tilde{Z}_d^e}} - 1 \right)$$

$$M_{B_d^e} = (1 - \mu_{\tilde{Z}_d^e}) \left(\frac{\mu_{\tilde{Z}_d^e}(1 - \mu_{\tilde{Z}_d^e})}{\sigma^2_{\tilde{Z}_d^e}} - 1 \right)$$

The final problem left is to determine $\mu_{\tilde{Z}_d^e}$ and $\sigma^2_{\tilde{Z}_d^e}$. We do this by different procedures for the CDF decoupler and the error-based decouplers. As discussed in Section 4.3.2, the CDF decoupler function produces uniform Z_d^e variables under the perfect expert assumption. Thus $\mu_{\tilde{Z}_d^e} = 1/2$ and $\sigma^2_{\tilde{Z}_d^e} = 1/12$, which are standard results for continuously uniform random variables.

For the case of sigmoid-composed error-based decouplers, $\phi_d = \sigma \circ \epsilon_d$, we generally would compute the mean and variance with

$$\mu_{\tilde{Z}_d^e} = E[\tilde{Z}_d^e] = E[\sigma(\epsilon((Q, F_Q))] = \int \sigma(\epsilon_d(q, F_Q)) f_Q(q) dq.$$

For the case of linearly interpolated beliefs f^e (and thus also f_Q), we can derive, using the same notation as in Section 2.2, that:

$$\mu_{\tilde{Z}_d^e} = \sum_{d=1}^{D+1} \frac{p_d - p_{d-1}}{m_d^e - m_{d-1}^e} \left(G_{\sigma(\epsilon_d)}(m_d^e) - G_{\sigma(\epsilon_d)}(m_{d-1}^e) \right)$$

$$\sigma_{\tilde{Z}_d^e} = E[(\tilde{Z}_d^e)^2] - \mu^2_{\tilde{Z}_d^e}$$

with

$$E[(\tilde{Z}_d^e)^2] = \sum_{d=1}^{D+1} \frac{p_d - p_{d-1}}{m_d^e - m_{d-1}^e} \left(G_{\sigma^2(\epsilon_d)}(m_d^e) - G_{\sigma^2(\epsilon_d)}(m_{d-1}^e) \right)$$

where $G_{\sigma(\epsilon_d)}$ and $G_{\sigma^2(\epsilon_d)}$ are primitive functions with respect to q of $\sigma(\epsilon_d(q, F^e))$ and $\sigma^2(\epsilon_d(q, F^e))$, respectively. This is derived in Appendix A.2. With the choice of the logistic sigmoid function, we have primitives calculated in Appendix A.1 as:

$$G_{\sigma(\epsilon_d)}(q, F^e) = \frac{\log(1 + e^{k\epsilon_d(q, F^e)})}{C_d^e k} \tag{4.18}$$

and

$$G_{\sigma^2(\epsilon_d)}(x) = \frac{1}{C_d^e k (1 + e^{k\epsilon_d(x)})} + \frac{\log(1 + e^{k\epsilon_d(x)})}{C_d^e k}. \tag{4.19}$$

The complete MAP estimation procedure can be summarized in the following steps:

1. **Calculate perfect expert moments:** Compute the target moments $\mu_{\tilde{Z}_d^e}$ and $\sigma_{\tilde{Z}_d^e}$ under the perfect expert assumption using the appropriate formulas for the chosen decoupling function (uniform distribution for CDF decoupler, or integration for sigmoid-composed error-based decouplers).

2. **Determine prior modes:** Use the perfect expert moments to calculate the prior modes $M_{A_d^e}$ and $M_{B_d^e}$ via the system of equations derived from the beta distribution moment matching.

3. **Convert to log-normal parameters:** For a specific prior variance σ^2_{Prior}, solve the nonlinear equations numerically to obtain the log-normal distribution parameters $\mu_{A_d^e}$, $\sigma_{A_d^e}$, $\mu_{B_d^e}$, and $\sigma_{B_d^e}$.

4. **Solve MAP optimization:** Maximize the posterior distribution by solving the optimization problem of a_d^e, b_d^e, that combines the log-normal priors with the beta likelihood from the observed data $z_{d,1}^e, \ldots, z_{d,N}^e$. Doing this, Z_d^e is then now modeled to have distribution $\text{Beta}(a_d^e, b_d^e, c_d^e, d_d^e)$.

For implementation of the MAP optimization problem, we used the R `rstan` package, which provides an interface to the probabilistic programming language `stan`[2].

Out of Bounds Values

Both the MLE and MAP estimation methods require observed realizations of the decoupled variables $z_{d,i}^e$ to lie within the theoretical support $[c_d^e, d_d^e]$ of the extended beta distribution. However, in practice, some observed values may fall outside this support due to the question invariance assumption not holding. If this happens we exclude those observations from the optimization algorithm.

In the extreme case where all observed realizations for a particular Z_d^e fall outside the theoretical support, the MLE and MAP methods cannot proceed with standard parameter estimation. Similarly, numerical optimization procedures may occasionally fail to converge due to poor initialization or challenging likelihood surfaces. The number of estimation failures for the upcoming empirical test is seen in Table A.2.

The Perfect Expert prior approach is unaffected by these issues since it does not rely on observed data, making it a robust fallback option when data-driven methods encounter difficulties.

Perfect Expert Prior

The PE prior approach represents a purely theory-driven method among our three estimation strategies. Rather than relying on historical data through MLE or incorporating limited prior information via MAP estimation, this approach directly implements the theoretical expectations derived from our perfect expert assumption.

In this method, we set the marginal distributions of Z_d^e to exactly match the theoretical moments of $\tilde{Z}_d^e$ calculated under the perfect expert assumption, as defined in the previous section. Specifically, we parameterize the extended beta distribution using the theoretically derived shape parameters:

$$Z_d^e \sim \text{Beta}(\mu_{A_d^e}, \mu_{B_d^e}, c_d^e, d_d^e)$$

where $\mu_{A_d^e}$ and $\mu_{B_d^e}$ are determined from the perfect expert moments $\mu_{\tilde{Z}_d^e}$ and $\sigma_{\tilde{Z}_d^e}$ using the system of equations presented earlier.

[2] https://mc-stan.org/

This approach is fundamentally non-data-driven, making no empirical inferences from observed realizations. Instead, it represents how decoupled variables would behave when experts provide assessments that perfectly align with the true underlying distributions, under the restriction to the beta distribution.

While this approach may seem overly optimistic in assuming perfect expert behavior, it serves as an important theoretical benchmark and could be used in isolation in scenarios where historical data is not available.

Exact Perfect Expert Marginal

The PE prior approach described above restricts the marginal distributions to the beta family for ease of comparison with the MAP and MLE methods. However, an alternative formulation allows for the exact distribution induced by transforming the expert beliefs without parametric constraints. In this exact approach, we directly use the distribution $Z_d^e = \phi_d(Q, F^e)$, which yields the density

$$f_{Z_d^e}(z) = f^e(\phi^{-1}_{d,F^e}(z)) \left| \frac{d}{dz} \phi^{-1}_{d,F^e}(z) \right|. \tag{4.20}$$

This formulation follows directly from the change of variables formula in probability theory, providing the exact distribution of the decoupled variables without requiring any parametric approximation.

While we do not implement this exact approach in our empirical evaluation, it serves a theoretical purpose in analyzing the model's connections to existing methods, as discussed in Section 4.4. The method can be viewed as implementing Design Principle 1 in its purest form, setting the prior distribution to reflect our theoretical expectations without the distributional restrictions imposed by the regular PE prior.

With the three practical marginal estimation approaches (MLE, MAP, and PE prior) established, we have completed the theoretical framework for estimating the marginal densities $f_{Z_d^e}$ in the decoupled copula model. The empirical comparison of these methods is presented in Chapter 5.

4.3.5 Estimation of copula

With strategies for estimating the distributions of Z_d^e laid out, we will here discuss the estimation of the copula density $c_{\mathbf{Z}}$. Following the inference functions for margins (IFM) approach described in Section 2.6.4, we estimate marginal and copula parameters sequentially.

For the copula estimation, we evaluate several different strategies. First, we consider the independence copula, which assumes no dependence between experts' errors and serves as our baseline approach. We also examine vine copulas with an independence or single parameter estimation per bivariate copula using Dißmann's algorithm. Additionally, we implement multivariate Frank copulas with MLE and an empirical Bayes Gaussian copulas approach with LKJ priors on the correlation matrix. Finally, we apply a connection threshold dimension reduction technique to both vine and Gaussian copula methods to improve computational tractability and numerical stability.

For copula estimation, we consider a study with test questions $Q_1, \ldots, Q_N$ having realizations $q_1, \ldots, q_N$ and corresponding expert beliefs $\mathbf{F}_1, \ldots, \mathbf{F}_N$. The estimation

procedure involves two key transformations. First, we compute the decoupled variables $\mathbf{z}_i = \phi(q_i, \mathbf{F}_i)$ with elements $z^e_{d,i} = \phi_d(q_i, F^e_i)$. Second, we transform these to uniform marginals by defining $\mathbf{u}_i$ with elements $u^e_{d,i} = F_{Z^e_{d,i}}(z^e_{d,i})$ for $i = 1, \dots N$, where $F_{Z^e_{d,i}}$ represents the estimated marginal distribution of the decoupled variable $Z^e_{d,i}$. The copula estimation methods then use these uniform samples $\mathbf{u}_i$ to estimate the copula density $c_{\mathbf{z}}$.

Independence Copula

For the independence copula, we assume independence between experts and simply set the copula density to that of the independence copula

$$c_{\mathbf{z}}(\mathbf{u}) = 1.$$

Estimation through vine copulas

Moving beyond the simple independence assumption, we employ vine copula estimation using Dißmann's sequential approach as outlined in Section 2.6.3. This algorithm involves a large set of potential parameters, details of which can be seen in Appendix A.5. Because the bivariate independence copula is part of the potential copulas the method selects from, this method represents a superset of the independence copula method.

This approach may provide flexibility in modeling complex dependence structures while maintaining computational tractability through the sequential decomposition of the multivariate copula into bivariate components.

Multivariate Frank Copula

As an alternative to the flexible but complex vine approach, we consider the multivariate Frank copula. The JC model described in Section 2.3.2 employed a multivariate Frank copula with manually selected parameters by the research team. Following [38], we implement the same Frank copula family but estimate the parameter using MLE from the observed data.

The multivariate Frank copula with parameter θ has the form:

$$C_{\text{Frank}}(\mathbf{u}; \theta) = -\frac{1}{\theta} \log \left(1 + \frac{\prod_{i=1}^{DE} (e^{-\theta u_i} - 1)}{(e^{-\theta} - 1)^{DE-1}} \right) \tag{4.21}$$

for $\mathbf{u} = (u_1, \dots, u_{DE}) \in [0,1]^{DE}$ and $\theta \neq 0$. For $\theta = 0$ we define it equal to the independence copula as that is the limit as $\theta \to 0$ [22]. The expression for the corresponding copula density, $c_{\text{Frank}}(\mathbf{u}; \theta)$, is involved for arbitrary dimension but a method of evaluating it can be referenced in [39].

Empirically, we observed better generalizability by applying regularization in the form of constraining the parameter θ to values that correspond to Kendall's tau in the range $[-0.9, 0.9]$. This constraint prevents extreme dependencies that can lead to numerical difficulties during optimization while maintaining a reasonable range of dependence structures.

The MLE proceeds by maximizing the log-likelihood:

$$\sum_{i=1}^{N} \log c_{\text{Frank}}(\mathbf{u}_i; \theta) \tag{4.22}$$

subject to the constraint that the corresponding Kendall's tau remains within the specified bounds. We employ the bound-constrained optimization algorithm by Bard et al. [40] to solve this constrained optimization problem.

MAP with Gaussian Copula

We model the joint dependence structure among the DE transformed variables using a Gaussian copula framework. The Gaussian copula provides an approach for capturing linear and near-linear dependencies.

The multivariate Gaussian copula with correlation matrix $\mathbf{R}$ has density:

$$c_{\text{Gaussian}}(\mathbf{u}; \mathbf{R}) = \frac{1}{\sqrt{\det(\mathbf{R})}} \exp\left(\frac{1}{2}\Phi^{-1}(\mathbf{u})^T(\mathbf{R}^{-1} - \mathbf{I})\Phi^{-1}(\mathbf{u})\right)$$

where $\Phi^{-1}(\mathbf{u}) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_{DE}))^T$ represents the vector of inverse standard normal CDFs applied element-wise to $\mathbf{u}$, and $\mathbf{I}$ is the identity matrix of dimension $DE \times DE$. For properties of the Gaussian Copula model, see for instance [41].

To implement a Bayesian approach, we place an LKJ (Lewandowski-Kurowicka-Joe) prior on the correlation matrix $\mathbf{R}$:

$$f(\mathbf{R}; \eta) \propto \det(\mathbf{R})^{\eta-1}$$

This prior possesses two, for us, notable properties. First, when $\eta = 1$, it provides a uniform distribution over the space of valid correlation matrices. Second, the off-diagonal elements of $\mathbf{R}$ follow a beta distribution on the interval $(-1, 1)$ [42]:

$$\text{Beta}\left(\eta + \frac{DE-2}{2}, \eta + \frac{DE-2}{2}, -1, 1\right)$$

It is worth noting that uniform distribution over correlation matrices does not imply uniform marginal distributions for individual correlation coefficients. This comes from not all configurations of matrix elements yielding valid correlation matrices that are positive definite and have ones on the diagonal.

The parameter η serves as a concentration parameter: as η increases above 1, the beta distribution becomes more concentrated around zero, effectively placing stronger prior belief on independence between variables. This allows us to control the degree of prior skepticism about dependencies in our model. We interpret $\eta = 1$ as a relatively uninformative prior, while larger values express stronger prior beliefs favoring independence.

The posterior distribution for the correlation matrix is:

$$\log f(\mathbf{R} \mid \mathbf{u}_1, \ldots, \mathbf{u}_N) \propto \sum_{i=1}^{N} \log c_{\text{Gaussian}}(\mathbf{u}_i; \mathbf{R}) + (\eta - 1) \log \det(\mathbf{R})$$

We employ MAP estimation to obtain point estimates of $\mathbf{R}$ by maximizing this posterior. This becomes equivalent to MLE when $\eta = 1$ because of the flat prior.

Connection Threshold as Dimension Reduction

The number of parameters required to model with the Gaussian copula scaled quadratically with dimension size. With limited sample size this can make numerical

optimization unstable, particularly when dealing with many experts. To address this challenge, we introduce a dimension reduction heuristic that models dependencies only between groups of highly correlated experts. This approach is inspired by Diß-mann's algorithm, which constructs vine copula structures by identifying variables with high Kendall tau correlations.

The procedure begins with a set of random variables $X_1, \dots, X_n$ and a threshold value $\tau_{\text{threshold}} \geq 0$. We calculate the Kendall tau correlation $\tau_{i,j}$ between each pair of variables X_i and X_j for all i and j. We then define the indicator function $\delta_{i,j} = \mathbb{1}(|\tau_{i,j}| \geq \tau_{\text{threshold}})$ and construct the adjacency matrix

$$A = \begin{bmatrix} 0 & \delta_{1,2} & \cdots & \delta_{1,n} \\ \delta_{2,1} & 0 & \cdots & \delta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{n,1} & \delta_{n,2} & \cdots & 0 \end{bmatrix}$$

We apply graph clustering to the adjacency matrix A to identify connected components, which represent groups of seemingly highly dependent experts. Let $G = \{G_1, G_2, \dots, G_k\}$ be the partition of $\{1, 2, \dots, n\}$ into k groups, where each G_i represents a connected component. For each group G_i, we define $\mathbf{X}_{G_i} = (X_j : j \in G_i)$ as the subvector of experts in group i.

The key assumption underlying this dimension reduction technique is that groups are independent of each other, allowing us to factorize a joint copula C as

$$C(X_1, \dots, X_n) = \prod_{i=1}^{k} C_{G_i}(\mathbf{X}_{G_i}),$$

where C_{G_i} is the copula restricted to the variables in group G_i. This factorization reduces the complexity of the estimation problem by replacing a single high-dimensional copula with multiple lower-dimensional copulas.

When implementing this connection threshold method in conjunction with the previously described copula estimation techniques, we first partition the Z_d^e variables according to the clustering procedure outlined above. Each group-specific copula C_{G_i} is then fitted separately using one of the established copula estimation methods.

With the copula estimation strategies established, including independence, vine, Frank, and Gaussian copulas, along with the connection threshold dimension reduction technique, we have completed the theoretical framework for modeling dependencies in the decoupled copula model. The empirical comparison of these copula estimation approaches is presented in Chapter 5.

4.3.6 Recovering from Numerical Failure

The marginal and copula estimation methods presented in the previous sections rely on numerical optimization procedures that can occasionally fail to converge. In practical applications focused on a single study, practitioners could invest more effort in tuning optimization parameters, adjusting starting values to achieve convergence. However, our evaluation framework requires running thousands of LOOCV evaluations across multiple studies and parameter combinations, making manual intervention for each failed optimization infeasible.

To address this in the upcoming results chapter, we implement a hierarchical fallback strategy as follows:

- **Marginal estimation failure**: When MLE or MAP estimation fails to converge for a particular Z_d^e, we substitute the PE prior approach, which requires no numerical optimization.

- **Copula estimation failure**: When vine, Frank, or Gaussian copula estimation fails to converge, we fall back to the independence copula which also requires no numerical optimization.

These substitutions are required for fewer than 1% of samples for any given optimization method tested on our dataset. Because of the low substitution rate, we will assume that this fallback strategy does not materially affect the overall performance comparisons and conclusions drawn from these.

4.3.7 Sampling of Decoupled Copula Model

Because the independence assumption only requires proportionality between the densities of $\mathbf{Z}$ and $\mathbf{Z} \mid \boldsymbol{\mathcal{F}}$, the density $f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}}$ of the decoupled copula model in Eq. (4.6) is unnormalized. Having discussed the selection and estimation of all components of the copula model, we now address the sampling procedure necessary to draw inferences from this unnormalized posterior distribution. Since direct sampling is not feasible, we employ Markov Chain Monte Carlo (MCMC) methods to generate samples from the target distribution.

We employ the extended differential evolution Markov chain (DE-MC) method presented in [43] and implemented in the R `BayesianTools` package [44]. This method is well-suited for expert aggregation problems as it efficiently handles multimodal posterior distributions by running multiple chains in parallel and exploiting information from past states to generate informed proposal jumps.

For initialization, we sample a starting population from the experts' interpolated beliefs $\widehat{f}^e$, running $50E$ chains where E is the number of experts. The starting points are generated by sampling 50 points from each expert's distribution $\widehat{F}^e$. This initialization strategy reflects the rationale that the aggregated belief likely exhibits modalities near those of individual expert beliefs, allowing the chains to efficiently explore the relevant regions of the parameter space. The specific parameters used for sampling with the `BayesianTools` package are detailed in Appendix A.4.

4.4 Connection to Existing Models

While structurally different from linear pooling methods, the decoupled copula model generalizes both the DP model and bears similarity to the Jouini-Clemen model under certain parameter choices. These connections demonstrate how our framework encompasses existing approaches as special cases while providing additional flexibility.

4.4.1 Connection to the Density Product Model

The decoupled copula model reduces to the DP model through specific parameter choices. To establish this connection, we set $D = 1$ and choose the identity decoupler

$\phi_d(q, F) = q$. Under this configuration, $\frac{d}{dz}\phi^{-1}_{d,F}(z) = 1$ and $\|\phi'_{\mathbf{F}}(q)\| = 1$. Applying these simplifications to Eq. (4.6) yields:

$$f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}} \propto \prod_{e=1}^{E} f_{Z_d^e}(q)$$

When combined with the exact perfect expert marginal estimation method, we have $f_{Z_d^e}(q) = f^e(q)$, which gives us:

$$f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}} \propto \prod_{e=1}^{E} f^e(q) \tag{4.23}$$

This is precisely the DP model, demonstrating that our framework generalizes this approach.

4.4.2 Connection to the Jouini-Clemen Model

The relationship to the JC model is established through a different set of parameter choices. We employ the exact perfect expert marginal estimation method, select the Frank copula for dependency modeling, and use the linear median error decoupler $\phi_d(q, F^e) = q - m^{e*}$, where m^{e*} is the median of F^e. Under these conditions, we again have $\|\phi'_{\mathbf{F}}(q)\| = 1$ and $\frac{d}{dz}\phi^{-1}_{d,F}(z) = 1$.

Substituting these choices into Eq. (4.6) produces:

$$f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}} \propto c_{\mathbf{Z},\text{Frank}}(F_{Z^e}(q - m^{1*}), \dots, F_{Z^e}(q - m^{E*})) \prod_{e=1}^{E} f_{Z^e}(q - m^{e*}) \tag{4.24}$$

With the exact perfect expert marginal estimation, $f_{Z^e}(z) = f^e(z + m^{e*})$ and $F_{Z^e}(z) = F^e(z + m^{e*})$, which simplifies the expression to:

$$f_{Q|\boldsymbol{\mathcal{F}}=\mathbf{F}} \propto c_{\mathbf{Z},\text{Frank}}(F^1(q), \dots, F^E(q)) \prod_{e=1}^{E} f^e(q) \tag{4.25}$$

This formulation closely resembles the JC model presented in Eq. (2.12), with the difference that the copula arguments are $F^e(q)$ rather than $1 - F^e(q)$. This discrepancy appears to stem from the choice of decomposition target in the original JC derivation. If the copula decomposition had been applied to the joint error distribution $\boldsymbol{\mathcal{E}}$ instead of the joint median distribution $\mathbf{M}$, the result would match Eq. (4.25) exactly. This alternative derivation appears to be referenced in the prose of [2], suggesting a potential inconsistency in their original mathematical derivation.

4.5 Summary of the method

The decoupled copula model provides a framework for aggregating expert judgments by modeling dependencies in a transformed space. We refer to the copula method as the copula model, Eq. (4.6), in addition to the procedures of estimating the marginal and copula densities. The following enumerated list provides a summary of the method:

1. **Belief Elicitation and Decoupling Function**

   (a) Obtain expert belief distributions F_i^e for experts $e = 1, \ldots, E$ across historical test questions $i = 1, \ldots, N$ with known realizations q_i. One approach to obtain the belief distributions is to collect expert quantile assessments such as $\{m^e_{5\%}, m^e_{50\%}, m^e_{95\%}\}$ and estimate the distributions using piecewise linear interpolation with support parameters as defined in Section 2.2.

   (b) Choose a decoupling function such as one from the list in Section 4.3.1.

   (c) Compute decoupled variables: $\mathbf{z}_i = \phi(q_i, \mathbf{F}_i)$ for all historical questions $i = 1, \ldots, N$ where $\mathbf{F}_i = (F_i^1, \ldots, F_i^E)$.

   (d) Optionally apply expert rejection preprocessing using distance correlation test to check independence between $\mathbf{Z}^e$ and Q for each expert. This is done by using the realizations $\mathbf{z}_i$ and q_i for $i = 1, \ldots, N$. Reject experts with p-values below $\alpha_{Rej} = 0.05$, retaining at least one expert.

   (e) Calculate support bounds: $c_d^e = \inf\{\phi_d(q, F^e) : q \in [L^*, U^*]\}$ and $d_d^e = \sup\{\phi_d(q, F^e) : q \in [L^*, U^*]\}$. For the CDF decoupler, this is always $[0, 1]$.

2. **Marginal Density Estimation**

   (a) Select one of the marginal estimation methods from Section 4.3.4: MLE, MAP, or PE prior.

   (b) Fit or select parameters of the extended beta distributions

   $$Z_d^e \sim \text{Beta}(a_d^e, b_d^e, c_d^e, d_d^e)$$

   for each expert e and property d, depending on the estimation method.

   (c) For MAP estimation: calculate perfect expert moments, set prior modes based on theoretical expectations, solve numerically for log-normal parameters, and maximize the posterior distribution.

   (d) Handle estimation failures by falling back to PE prior approach.

3. **Copula Estimation**

   (a) Transform decoupled variables to uniform marginals: $u_{d,i}^e = F_{Z_d^e}(z_{d,i}^e)$ for expert e, property d, and test question i using the estimated marginal CDFs.

   (b) Select one of the copula families from Section 4.3.5: independence, vine, Frank, or Gaussian copula.

   (c) If using connection threshold dimension reduction: calculate Kendall tau correlations, construct adjacency matrix with threshold $\tau_{\text{threshold}}$, apply graph clustering, and factorize the copula as $C(\mathbf{u}) = \prod_{i=1}^k C_{G_i}(\mathbf{u}_{G_i})$.

   (d) Estimate copula parameters using the chosen method (MLE, MAP, or another approach specific to the copula family).

   (e) Handle estimation failures by falling back to independence copula.

4. **MCMC Sampling for New Question**

   (a) For a new question with expert beliefs $\mathbf{F}_{\text{new}}$, construct the unnormalized posterior density using Eq. (4.6), using the estimated densities of $\mathbf{Z}$ from the previous steps.

   (b) Initialize sampler using starting points sampled from individual expert beliefs $\hat{F}^e_{\text{new}}$.

   (c) Compute point estimates and uncertainty quantification from MCMC samples.

Having established the theoretical foundation and implementation framework of the decoupled copula model, we now turn to its empirical evaluation in Chapter 5 and subsequently discuss the implications and limitations of our approach in Chapter 6.

Chapter 5

Results

The previous chapter established the theoretical framework of the decoupled copula model, presenting the mathematical foundation and various methodological choices for each component. This chapter provides the empirical evaluation and performance analysis of the proposed approach. The analysis proceeds in two main stages: first, we evaluate and select optimal parameter configurations for the decoupled copula model through component-level analysis and system-level comparison. Second, we benchmark the selected configuration against established SEJ methods to assess the practical value of the proposed approach.

The evaluation employs the LOOCV framework introduced in Section 2.5 and performance metrics introduced in Chapter 3 to give measurable comparisons of the different parameter choices and will also aid in benchmarking the decoupled copula model against existing SEJ models.

5.1 Choosing Parameters of the Copula Model

This section presents the empirical evaluation of the decoupled copula model components and parameter selection. We begin by evaluating individual components, decoupling functions, marginal estimation methods, and copula estimation approaches, before comparing different parameter configurations of the complete system.

5.1.1 Empirical Investigation of Decoupling Functions

For empirical evaluation of the decoupling functions, we will test the question invariance assumption by looking at the dependence between $\mathbf{Z}$ and Q as this a required (but not sufficient) criterion as stated by Lemma 5. Empirically, we will test this by independence testing on the dataset from Section 2.4. Let the independence null hypothesis H_0 be defined as

$$H_0 : \quad \mathbf{Z} = \phi(Q; \mathbf{F}) \text{ is independent from } Q. \tag{5.1}$$

If a study has n calibration questions, we will use the test statistic, $\mathcal{T}_n$, as introduced in Section 2.6.5, and reject H_0 when $\mathcal{T}_n > c_\alpha = t_{\nu-1}^{-1}(1-\alpha)$ where c_α is the $1-\alpha$ quantile of a Student t distribution with $\nu - 1$ degrees of freedom where $\nu = \frac{n(n-3)}{2}$.

Applying this independence test to all studies in the dataset yields the results shown in Table 5.1 without expert rejection preprocessing, while Table 5.2 shows

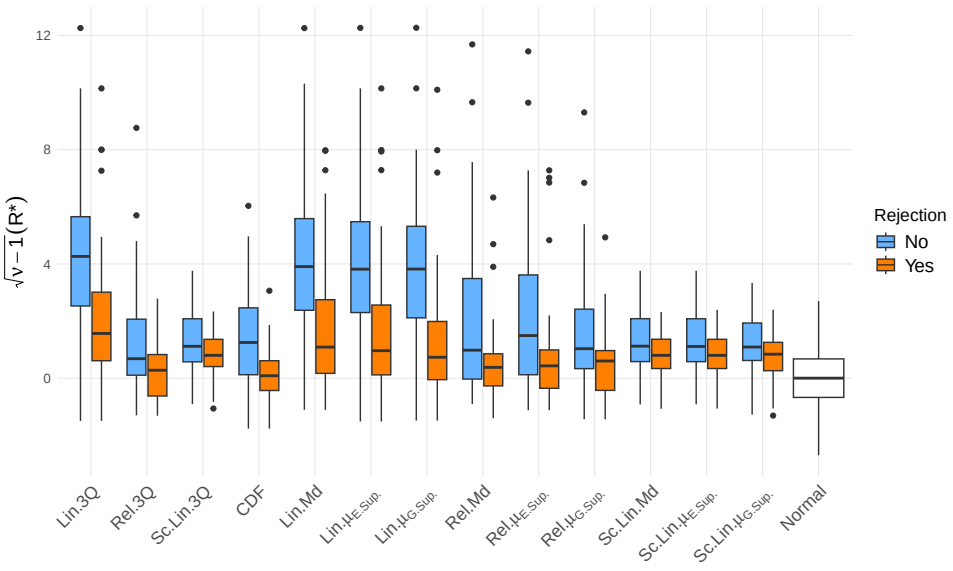


Figure 5.1: Box plot over $\sqrt{\nu-1}\mathcal{R}^*$ for different decoupler functions. Also includes a box plot of a standard normal distribution.

the results with the preprocessing. An important limitation in interpreting these results is that we cannot directly compare decoupler functions with different output dimensions ($D = 1$ vs $D = 3$), as the statistical power of the test varies with dimension. However, meaningful comparisons can be made between decouplers of the same dimension.

The relative performance rankings among decouplers depend notably on whether expert rejection is employed. With expert rejection preprocessing, the CDF decoupler, relative median, relative mean with global support, and relative 3-quantile decouplers demonstrate the lowest rejection rates. Without expert rejection, the scaled linear error decouplers perform best among the $D = 1$ dimension decouplers, while the relative 3-quantile decoupler maintains its strong performance in the $D = 3$ category.

The problem of comparing decouplers with different dimensions is also illustrated in tables Tables 5.1 and 5.2: we would theoretically expect the relative 3-quantile decoupler to exhibit stronger dependence with Q than the relative median decoupler, based on the mutual information arguments presented earlier. However, our empirical results show the opposite pattern. This apparent contradiction likely results from the distance correlation test having reduced statistical power in higher dimensions.

In addition to null hypothesis testing, we have from the theory of distance correlation that the rescaled statistic $\sqrt{\nu-1}\mathcal{R}^*$ defined in Section 2.6.5, converges asymptotically to a normal distribution under the independence assumption. Figure 5.1 provides a visual investigation of the empirical distribution of $\sqrt{\nu-1}\mathcal{R}^*$ to compare against this asymptotic normality result. The figure demonstrates how expert rejection preprocessing appears beneficial for question invariance across all decoupler types, as well as showing how the CDF decoupler with expert rejection seems to have quartiles most similar to that of the standard normal.

The quantitative evidence for this improvement is apparent when comparing Table 5.1 and Table 5.2. The proportion of studies where each decoupler is rejected decreases systematically across all decoupler types when expert rejection is

Table 5.1: Result of null hypothesis independence test without expert rejection. The 'Highest p-value' column is computed separately for $D = 1$ and $D = 3$. A total of 47 studies were analyzed.

Decoupler	**D**	**H_0 Rejected (% of studies)**	**Highest p-value (% of studies)**
Lin.3Q	3	85%	13%
Rel.3Q	3	30%	57%
Sc.Lin.3Q	3	36%	30%
CDF	1	40%	15%
Lin.Md	1	85%	4%
Lin.$\mu_{\text{E.Sup.}}$	1	81%	2%
Lin.$\mu_{\text{G.Sup.}}$	1	81%	2%
Rel.Md	1	43%	21%
Rel.$\mu_{\text{E.Sup.}}$	1	45%	9%
Rel.$\mu_{\text{G.Sup.}}$	1	40%	15%
Sc.Lin.Md	1	34%	4%
Sc.Lin.$\mu_{\text{E.Sup.}}$	1	34%	4%
Sc.Lin.$\mu_{\text{G.Sup.}}$	1	30%	23%

applied. This suggests that the preprocessing step successfully removes experts whose marginal distributions violate the question invariance assumption.

Based on these empirical findings, we focus subsequent analysis on the CDF decoupler and the relative median decoupler due to their strong performance in the expert rejection scenario. We exclude $D = 3$ decouplers from further consideration for two reasons: clarity of exposition and the lack of compelling theoretical motivation for why the relative error between realizations and specific percentiles (5% or 95%) should constitute question-invariant quantities across studies.

Having identified the promising decoupling functions, we now turn to evaluating the marginal estimation methods that will model the distributions of the decoupled variables Z_d^e.

5.1.2 Empirical Results of Marginal Estimation Methods

To evaluate the performance of the three marginal estimation methods (MLE, MAP, and PE prior), we assess them using the calibration and point estimates presented in Chapter 3. Within the framework for evaluating DMs, each marginal estimation method can be viewed as a DM aiming to estimate Z_d^e. Using the LOOCV procedure described in Section 2.5, we fit each method on training questions and evaluate them on held-out test questions.

Table 5.3 presents performance measures for the relative median decoupler with $k = 0.05$ and the CDF decoupler across all estimation strategies and with σ_{prior} values of $0.1, 0.25, 0.5$ and 0.75. Among the relative median decouplers, we focus on $k = 0.05$ as it generally demonstrates better performance than other k values; complete results for all k values can be found in Table A.3. A visual inspection of the calibration is shown in Figs. 5.2 and 5.3 for the CDF and the relative decoupler respectively. The plots show the empirical CDFs of the values $u_{d,s,t}^e = F_{Z_d^e}(z_{d,s,t}^{e*})$

Table 5.2: Result of null hypothesis independence test with expert rejection preprocessing. The 'Highest p-value' among studies is computed separately for $D = 1$ and $D = 3$. 'Mean Experts Excluded' refers to the average number of rejected experts per study. A total of 47 studies were analyzed.

Decoupler	**D**	**H_0 Rejected (% of studies)**	**Highest p-value (% of studies)**	**Mean Experts Excluded**
Lin.3Q	3	49%	17%	6.9
Rel.3Q	3	4%	64%	2.2
Sc.Lin.3Q	3	15%	19%	2.4
CDF	1	11%	28%	2.0
Lin.Md	1	38%	9%	5.1
Lin.$\mu_{\text{E.Sup.}}$	1	36%	4%	5.1
Lin.$\mu_{\text{G.Sup.}}$	1	34%	6%	4.9
Rel.Md	1	11%	19%	2.3
Rel.$\mu_{\text{E.Sup.}}$	1	15%	9%	2.4
Rel.$\mu_{\text{G.Sup.}}$	1	11%	15%	2.4
Sc.Lin.Md	1	15%	2%	2.2
Sc.Lin.$\mu_{\text{E.Sup.}}$	1	15%	0%	2.1
Sc.Lin.$\mu_{\text{G.Sup.}}$	1	13%	9%	2.1

where $F_{Z_d^e}$ is the estimated Z_d^e distribution of a method, and $z_{d,s,t}^{e*}$ is the test sample from the LOOCV procedure (Section 2.5) for study s, expert e, and property d. Because we are using expert rejection preprocessing the number of experts in a study changes per decoupler, and thus also the exact number of $u_{d,s,t}^e$ samples. For the used dataset specific counts can be seen in Table A.2 but it ranges between around 4200 and 4600 samples.

The empirical results reveal several patterns. The relative median decoupler appears to outperform the CDF decoupler across all metrics for all estimation methods. This could either be because the relative decoupler fulfills the model assumptions better, or because the beta distribution family used for Z_d^e is more suitable to capture the densities induced by the relative decoupler. We also see that the $\text{MAE}_{\text{Median}}$ and $RMSEmean$ metrics show relatively stable performance (varying by approximately $\pm 3\%$ for the CDF and $\pm 6\%$ for the relative decoupler) between different estimation methods within the same decoupler. They do differ between decouplers but because the support of the Z_d^e is dependent on the choice of decoupler, this metric is difficult to compare between different decoupling functions.

An interesting inverse relationship emerges between calibration performance and the degree of data-driven estimation. The relative decoupler tends to achieve better calibration with less data-driven methods (performing best with PE prior and worst with MLE), while the CDF decoupler exhibits the opposite trend. This pattern may indicate that while the CDF PE Prior approach has an exact analytical solution under the perfect expert assumption (uniformly distributed), this theoretical ideal does not appear to reflect actual expert behavior patterns.

With marginal estimation methods evaluated, we next examine the copula estimation approaches that will model dependencies between the decoupled variables.

Table 5.3: Marginal estimation performance comparison for CDF decoupler and Rel.Md. decoupler with $k = 0.05$. When a method specifies only the value of σ_{prior} then the marginal MAP estimation method has been used with that prior standard deviation. The table is sorted in ascending L^1_{Unif} order.

Decoupler	**Method**	$\mathbf{L^1_{Unif}}$	$\mathbf{L^\infty_{Unif}}$	$\mathbf{MAE_{Median}}$	$\mathbf{RMSE_{Mean}}$
Rel.Md.	PE prior	0.035	0.065	0.018	0.053
Rel.Md.	$\sigma_{\text{prior}} = 0.1$	0.035	0.065	0.017	0.052
Rel.Md.	$\sigma_{\text{prior}} = 0.25$	0.038	0.086	0.017	0.052
Rel.Md.	$\sigma_{\text{prior}} = 0.5$	0.041	0.097	0.018	0.052
Rel.Md.	$\sigma_{\text{prior}} = 0.75$	0.043	0.101	0.018	0.052
Rel.Md.	MLE	0.046	0.105	0.018	0.052
CDF	MLE	0.049	0.096	0.346	0.383
CDF	$\sigma_{\text{prior}} = 0.75$	0.062	0.119	0.341	0.378
CDF	$\sigma_{\text{prior}} = 0.5$	0.067	0.131	0.339	0.377
CDF	$\sigma_{\text{prior}} = 0.25$	0.076	0.164	0.337	0.373
CDF	$\sigma_{\text{prior}} = 0.1$	0.084	0.205	0.335	0.371
CDF	PE prior	0.087	0.229	0.336	0.371

5.1.3 Empirical Comparison of Copula Estimation

To evaluate the copula estimation methods presented in the previous sections, we conduct an empirical comparison using the LOOCV framework described in Section 2.5. For the MAP method, we test η values of 1, 10, and 50, while for both MAP and Vine methods, we employ connection thresholds $\tau_{\text{threshold}}$ of 0, 0.5, and 0.7. Due to the multidimensional nature of the copula estimation problem, we cannot apply the same metrics used for decoupler and marginal estimation evaluation. Instead, we perform a likelihood-based comparison using the independence method, that is constant one, as our baseline.

It turns out that the numerical stability of these methods varies considerably, with MAP methods experiencing substantial failure rates when used without connection threshold dimension reduction. Detailed failure rates for each method are presented in Table A.1. To ensure reliable comparisons, we focus exclusively on methods with successful convergence rates higher than 99% of the times it was applied. Among the included models, the MAP method with $\eta = 1$ and $\tau_{\text{threshold}} = 0.7$ exhibits the lowest convergence rate at 99.3%. The selected high-convergence methods are compared in Fig. 5.4.

Figure 5.4 shows the likelihood comparison across different copula estimation methods, illustrating the relative performance of each approach in terms of model fit quality.

To quantify the performance relative to the independence copula baseline, Table 5.4 presents the percentage of LOOCV evaluations where each method achieved likelihood values above or below 1 (the constant likelihood of the independence copula). The results reveal interesting patterns in method performance. The Frank copula demonstrates the most extreme behavior, achieving both the highest frequency of higher likelihood relative to the independence baseline and the highest frequency of lower likelihood. Notably, only the MAP estimation methods consistently performed better more often than they underperformed.

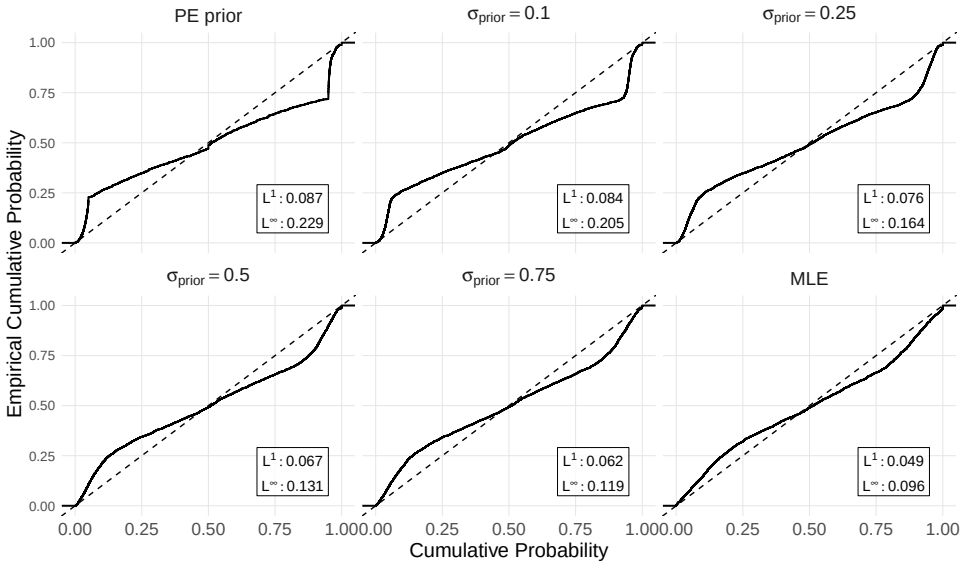


Figure 5.2: ECDFs of calibration quantities for different marginal estimations procedures for the CDF decoupler. The dashed diagonal line represents perfect calibration. The ECDFs are estimated from 4468 samples each.

Table 5.4: Percentage of LOOCV samples with likelihood values above and below the independence copula baseline (likelihood = 1). The horizontal line separates methods that performed better than the independence method more often than they underperformed.

Method	% likelihood > 1	% likelihood < 1
MAP:$\eta(50)$:$\tau_{\text{threshold}}(0.7)$	38.7	24.0
MAP:$\eta(10)$:$\tau_{\text{threshold}}(0.7)$	38.6	24.4
MAP:$\eta(1)$:$\tau_{\text{threshold}}(0.7)$	35.7	27.2
Vine:$\tau_{\text{threshold}}(0.7)$	24.7	26.1
Frank	44.4	55.6
Vine	34.5	49.3
Vine:$\tau_{\text{threshold}}(0.5)$	34.4	49.3

Since the Frank and Vine estimation methods represent supersets of the independence copula, their generally lower likelihood values compared to the independence baseline can be interpreted as evidence of overfitting to the training data. However, these same methods also exhibit the highest proportion of likelihood values exceeding 5.

Overall, no method, arguably, demonstrates a clear and significant advantage over the independence copula across all evaluation criteria. For the subsequent benchmarking analysis, we select the MAP method with $\eta = 10$ and $\tau_{\text{threshold}} = 0.7$, along with the independence copula method. Although the MAP method with $\eta = 50$ and $\tau_{\text{threshold}} = 0.7$ performed marginally better according to Table 5.4, its behavior is more similar to the independence copula. Therefore, we choose the $\eta = 10$ variant to introduce greater methodological diversity in our evaluation.

With the individual components evaluated, we now assess how different combinations of these components perform when integrated into the complete decoupled

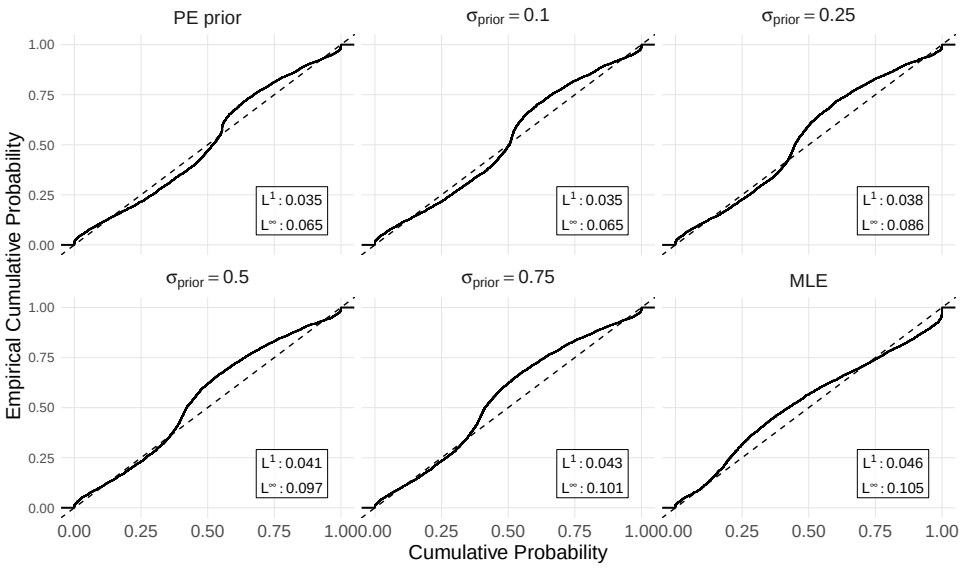


Figure 5.3: ECDFs of calibration quantities for different marginal estimations procedures for the relative median decoupler with $k = 0.05$. The dashed diagonal line represents perfect calibration. The ECDFs are estimated from 4468 samples each

copula model.

5.1.4 Parameter Configuration Comparison

Based on the component-level analysis in the previous subsections, we now evaluate different parameter configurations of the complete decoupled copula model. The empirical analysis suggested that a sigmoid-composed relative median error decoupler with $k = 0.05$, combined with PE prior marginal estimation and Gaussian MAP copula estimation (connection threshold 0.7, $\eta = 10$), may provide favorable performance characteristics. To evaluate whether those performance characteristics translate to practical DM performance, we assess this configuration alongside a broader range of parameter combinations.

Specifically, we compare the relative median decoupler against the CDF decoupler. For marginal estimation, we evaluate the complete range of methods examined in Section 4.3.4: PE prior, MAP estimation with various σ_{prior} values (0.1, 0.25, 0.5, 0.75), and MLE estimation. For copula estimation, we compare the Gaussian MAP method (connection threshold 0.7, $\eta = 10$) with the independence copula baseline.

The evaluation employs the same LOOCV framework used in previous component analyses, applying the performance metrics L^1_{Unif}, L^∞_{Unif}, $\text{MedAE}_{\text{median}}$, and $\text{MedSE}_{\text{Mean}}$ introduced in Chapter 3. Results across different parameter configurations are presented in Fig. 5.5.

The results reveal several patterns. For all metrics, the relative decoupler tends to outperform the CDF decoupler across most parameter configurations. The CDF decoupler appears to exhibit more continuous performance changes along the marginal estimation method spectrum, transitioning smoothly from less data-driven approaches (PE prior) to more data-driven methods (MLE). This pattern also holds for the relative decoupler between the PE prior and the largest σ_{prior} approach, but the relative decoupler exhibits a notable discontinuity for the calibration metrics

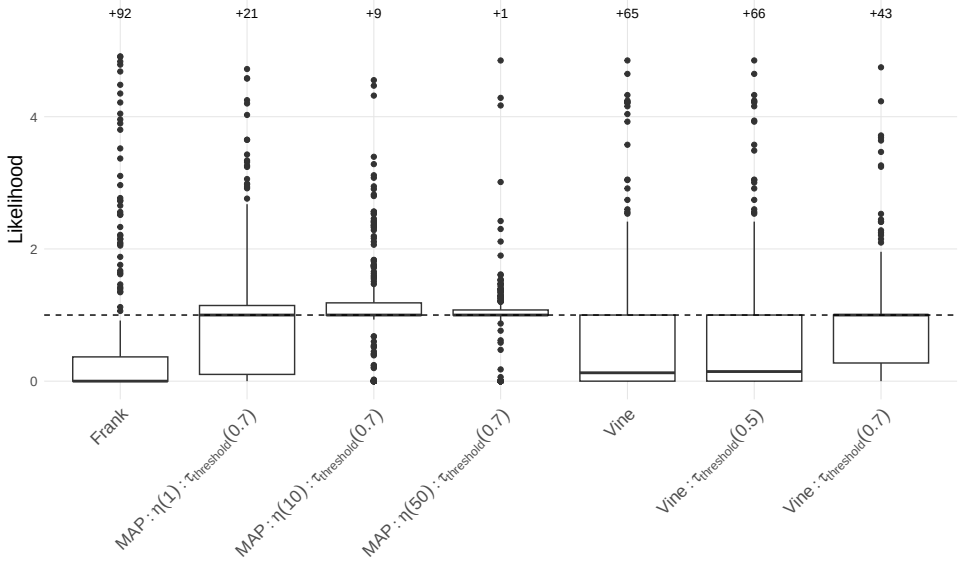


Figure 5.4: Likelihood comparison across different copula estimation methods. The boxplots show the distribution for likelihoods less than or equal to 5. The numbers on the top of the plot indicate how many likelihoods were greater than 5 for that copula estimation method. $y = 1$ is dashed in red to indicate the likelihood of the independence copula.

between the $\sigma_{\text{prior}} = 0.75$ and MLE estimation.

Regarding the median and mean performance metrics, the relative decoupler shows less sensitivity to the choice of marginal estimation approach compared to the CDF decoupler.

The comparison between copula estimation methods indicates similar performance levels for both the Gaussian MAP approach and the independence copula, suggesting that the additional complexity of modeling expert dependencies may not provide substantial performance improvements for this dataset.

While the PE prior performed the best for the relative decoupler in terms of the calibration metrics when measured against predicting Z_d^e as was seen in Table 5.3, here the PE prior underperforms all other marginal estimation methods except MLE in terms of calibration.

5.2 Benchmarking Against Existing Methods

Based on the parameter comparison results in the previous section, we select the configuration that balances performance across multiple metrics for comparison with existing methods. Specifically, we evaluate the sigmoid-composed relative median decoupler with $k = 0.05$, combined with MAP marginal estimation using $\sigma_{\text{prior}} = 0.5$, Gaussian copula estimation, expert rejection preprocessing with $\alpha_{\text{Rej}} = 5\%$, and dimension reduction with $\tau_{\text{threshold}} = 0.7$. This configuration demonstrated consistently good performance, relative to the other configurations, across the different calibration and point estimate metrics. In this section, this configuration is referred to as the Decoupled Copula model.

We benchmark this configuration against the established methods introduced in Section 2.3: the Density Product model, the Equal Weights model, the Jouini-Clemen

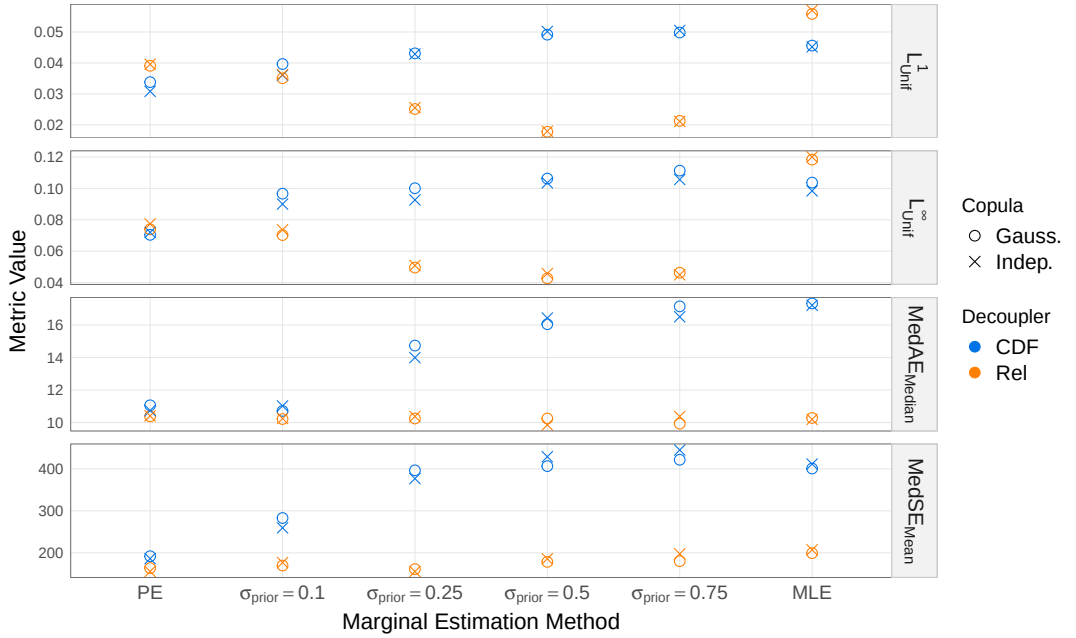


Figure 5.5: Point estimate performance comparison across different decoupled copula method configurations. The figure shows four performance metrics (L^1_{Unif}, L^∞_{Unif}, $\text{MedAE}_{\text{median}}$, and $\text{MedSE}_{\text{Mean}}$) for combinations of two decoupling functions (CDF and relative median with $k = 0.05$), five marginal estimation methods (PE prior, MAP with $\sigma_{\text{prior}} \in \{0.1, 0.25, 0.5, 0.75\}$, and MLE), and two copula estimation approaches (Gaussian MAP and Independence). All configurations use $\tau_{\text{threshold}} = 0.7$ dimension reduction and expert rejection preprocessing with $\alpha_{\text{Rej}} = 5\%$.

model, the Optimized Classical Model, and the Uniform model.

The calibration analysis through empirical cumulative distribution functions is presented in Fig. 5.6. The Equal Weights model shows good calibration for the median but tends to underestimate lower quantiles while overestimating the upper ones, indicating underconfident behavior. Similarly, the Jouini-Clemen model has a calibrated median but exhibits a significantly larger underconfident behavior. The Globally Optimized Classical Model exhibits similar calibration patterns to Equal Weights but with slightly more median overestimation, with approximately 56% of observations falling below the claimed median.

The Decoupled Copula model demonstrates slight overestimation of lower tail quantiles but appears well-calibrated for the median and quantiles above the median. In contrast, the Uniform model shows substantial overestimation beginning from the first quartile. Notably, the Decoupled Copula, Equal Weights, and Jouini-Clemen models all exhibit calibrated means, though the Jouini-Clemen model shows poorer overall calibration metrics compared to the closely related Density Product model.

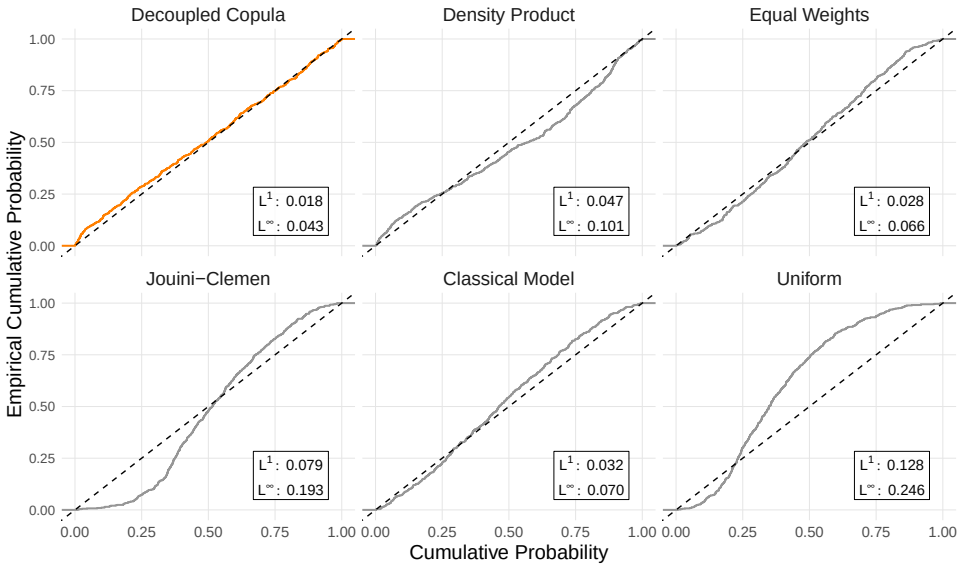
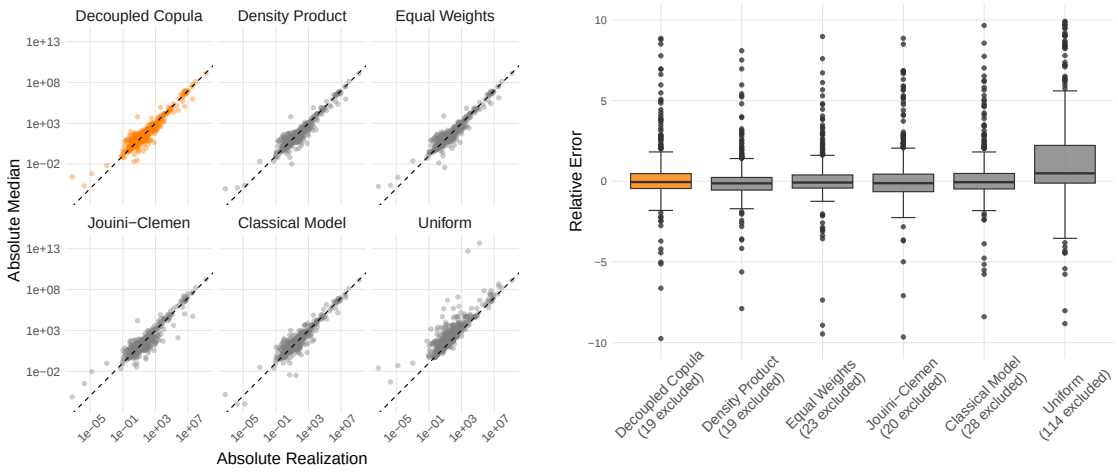


Figure 5.6: ECDFs from the calibration quantities from a subset of copula method configurations comparing calibration performance. The dashed line shows the optimal uniform CDF that represents perfect calibration. The ECDFs are estimated from 548 samples each.



(a) Absolute values comparison

(b) Relative errors boxplot comparison

Figure 5.7: Performance comparison across different method configurations showing (a) absolute median against absolute realizations comparison and (b) relative error distribution. Both plots exclude 1 question, out of the 548 in the dataset, that had a realization of 0. In the boxplot, the axis is limited to show relative errors with absolute values smaller or equal to 10. Under each method name, it is reported how many relative errors had an absolute value larger than 10.

Chapter 6

Discussion

This work introduced the decoupled copula model as a novel framework for aggregating expert judgments that can adjust for systematic biases in expert assessments and capture inter-expert dependencies through flexible statistical measures beyond linear error structures. The approach provides explicit numerical procedures for parameter estimation and includes theoretical criteria for evaluating dependency measures. The research also contributed novel calibration definitions and criteria for DM evaluation and includes empirical comparisons against existing SEJ models. This chapter discusses the implications of our findings, examines the limitations of the proposed approach, and suggests directions for future research.

6.1 Main Contributions

6.1.1 Theoretical Framework

The decoupled copula model contributes to the structured expert judgment literature by addressing key limitations of existing methods through several theoretical advances. The framework provides a systematic approach for correcting individual expert biases by transforming assessments into a space where systematic patterns can be identified and adjusted through historical data analysis.

The framework generalizes the density product approach and includes a special case that is closely related to the JC model for specific parameter choices. Unlike previous Bayesian approaches that are constrained to particular dependency measures such as linear errors, the proposed framework offers flexibility in the choice of dependency measures through decoupling functions, while providing both theoretical criteria and empirical tests for evaluating these choices.

The concept of question invariance represents a key theoretical contribution, establishing a criterion for evaluating whether transformed expert assessments maintain consistent distributional properties across different questions. The framework also enables the modeling of expert dependencies when they exist, though this capability occupies a more secondary role given the empirical findings.

In addition to the decoupled copula model, this work also proposes novel definitions and calibration criteria for evaluating DM methods when having access to only single realizations of heterogeneous random variables. While the classical model also has a calibration measure, ours is not limited to specific quantile ranges, and the L^1 and L^∞ measures provide inherently descriptive metrics that are not sensitive

to sample sizes such as the p-value based calibration score of the classical model. These calibration measures add more nuance in comparing DM methods to assess uncertainty quantification better than only looking at the predictive performance of the median or mean.

6.1.2 Empirical Evidence

The benchmarking analysis indicates that the decoupled copula model can achieve mean and median predictive performance levels similar to that of established methods while outperforming these methods with respect to calibration. When configured appropriately, the method demonstrated calibration properties that outperform equal weights aggregation and the classical model, while offering greater theoretical flexibility in modeling expert relationships.

While we kept the selection of possible parameter values coarse to avoid severe overfitting, it is worth noting that the benchmarked configuration was chosen because of its performance on the dataset and that this might not translate perfectly to a new dataset. Our goal with the benchmarking section, however, was not to be a comprehensive performance comparison. Rather, the empirical analysis provides initial evidence that the theoretical framework can translate into practical aggregation performance, suggesting that the added complexity of the model may be justified and worth improving on.

An important empirical finding was that modeling expert dependencies appeared to provide only marginal performance increases. This was observed both in direct copula fit performance measures and in the predictive performance of the final DM, suggesting that the benefits of sophisticated dependency modeling may be more limited than initially anticipated. This limited impact could occur for two reasons: either dependencies between experts do not significantly matter for changing predictions, or such dependencies exist but are elusive to capture with the current estimation methods and available data.

6.2 Insights Into the Decoupled Copula Model

6.2.1 Decoupling Function Performance

The analysis of different decoupling functions revealed a few important patterns for the future development of the model. The CDF decoupler possessed both stronger theoretical arguments for question invariance and empirically demonstrated better independence properties in testing. However, this theoretical and empirical advantage in achieving question invariance did not translate into superior marginal estimation performance or better end-target predictive performance compared to the relative median decoupler.

While the exact reason for this disconnect remains an open question, one potential explanation could be that the beta distribution parameterization for the marginals was more fitting for capturing distributions induced by the relative decoupler than the CDF decoupler.

6.2.2 Marginal Estimation Inconsistencies

An unexpected pattern emerged in the relationship between component-level and system-level performance. For the relative median decoupler, the perfect expert prior approach demonstrated superior calibration performance when evaluated on marginal estimation tasks, yet performed worse than MAP methods when evaluated on final DM calibration performance. Inversely, for the CDF decoupler the MLE marginal estimation approach demonstrated better calibration performance on the marginal estimation tasks while the PE prior gave the best calibration on final DM performance. This inconsistency suggests that better calibration of individual components does not necessarily translate to better overall system calibration and that the interactions between different model components may be more complex than anticipated.

6.2.3 The Value of Bayesian Modeling

While dependency modeling showed limited empirical benefits, the comparison between different marginal and copula estimation approaches revealed the importance of Bayesian modeling compared to purely data-driven approaches. The MAP marginal estimation methods, which combined theoretical priors with empirical evidence, generally outperformed both purely theoretical (PE prior) and purely data-driven (MLE) alternatives in final system performance. Similarly, for the copula estimation benchmarks, the Gaussian MAP methods outperformed the data-driven Frank and vine estimation methods in terms of providing higher likelihoods for unseen samples.

6.3 Limitations and Methodological Considerations

6.3.1 Distributional Modeling Constraints

The current implementation of the decoupled copula model assumes continuous distributions for the transformed variables Z_d^e, which may not adequately capture all aspects of expert behavior. The CDF decoupler, in particular, can produce boundary values of 0 and 1 when observations fall outside expert-assessed probability ranges, creating mixed discrete-continuous distributions that the current framework does not explicitly model.

A more comprehensive approach might explicitly model discrete probability masses at boundary values while maintaining continuous distributions for interior values. Even for relative error decouplers, that do not induce mixed discrete-continuous distributions, it might be advantageous to limit the relative errors to a fixed range and treat observations outside of this range as discrete events. This could be interpreted from a perspective of expert behavior: for small errors, experts may exhibit systematic biases that can be observed and adjusted for across questions, while for large errors, experts may not exhibit systematic beliefs or such patterns may be significantly harder to capture. We suspect this is the most promising future research direction for this model.

6.3.2 Parameter Convergence Inconsistencies

The analysis revealed inconsistencies in the behavior of certain modeling components across different decoupling functions. The way σ_{prior} was defined was intended to ensure that as it increased, the MAP marginal estimations would converge towards MLE. This convergence pattern was observed for the CDF decoupler but not for the relative decoupler, as can be seen visually in the marginal empirical CDF plots and in the trends shown in the point metrics analysis. This inconsistency suggests that parameter selection may require more decoupler-specific guidance than initially anticipated to be interpretable across different decoupler functions.

6.3.3 MCMC Induced Variance

The use of MCMC sampling for all benchmarking methods, even those with closed analytical formulas, was chosen to better compare the core modeling differences rather than their practical implementation advantages. While this approach ensures a fair comparison of the foundational modeling, it does not reflect optimal implementation strategies for practical applications. In particular, it misfavors the linear pooling methods that have easy-to-compute closed-form expressions for their DM densities and CDFs.

6.4 Future Research Directions

Based on the insights gained from this work, several promising directions emerge for extending and improving the decoupled copula framework. These research avenues address both technical limitations identified in our analysis and broader opportunities for methodological advancement.

1. **Mixed Discrete-Continuous Extensions.** As discussed in the limitations, developing explicit mixed discrete-continuous modeling for boundary values and extreme errors represents a promising direction. This could be particularly beneficial for the relative decoupler and might help explain some of the performance differences observed between decoupling functions.

2. **Refined Independence Assumption.** It might be fruitful to explore the possible connections between $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\varepsilon)$ and $f_{\mathbf{Z}}(\varepsilon)$ further than just assuming proportionality. For example, one could explore projecting the joint $\mathbf{Z}$ space onto the $\Gamma_{\mathbf{F}}$ line as potentially being a more robust estimation of $f_{\mathbf{Z}|\boldsymbol{\mathcal{F}}=\mathbf{F}}(\varepsilon)$. Pragmatically, at least, this kind of estimation would have the benefit of producing a density with equality, and not proportionality, which might improve ease of inference.

3. **Question-Level Dependencies.** While this work focused on dependencies between experts, dependencies between questions remain an open research area. Expert judgment studies often involve related questions that may exhibit systematic patterns in expert performance or bias. Extending the framework to capture these question-level dependencies could potentially improve aggregation performance, particularly in studies with many related questions.

4. **Alternative Approaches to Uncertainty Quantification.** If expert dependencies exist but are difficult to capture reliably, an alternative approach would be to simulate different potential dependency structures and investigate the spread of outcomes. This could provide a wider range of uncertainty estimates, which might be desirable in risk-averse decision-making situations where acknowledging model uncertainty about expert relationships is important.

5. **Context-Specific Applications.** Future research could investigate whether expert dependencies are more pronounced in specific domains or contexts. Understanding when dependency modeling provides greater benefits could inform more targeted applications of the framework, particularly in fields where experts share common training backgrounds, information sources, or analytical frameworks.

6.5 Conclusion

The decoupled copula model represents an advance in structured expert judgment theory by providing a flexible Bayesian framework for bias correction and dependency modeling. The method achieved competitive predictive performance and superior calibration against the benchmarked models. The findings revealed that dependency modeling provided limited benefits. At the same time, the marginal Bayesian combination of theoretical priors with empirical evidence seemed particularly valuable relative to the MLE and PE prior methods. Future developments in mixed discrete-continuous modeling may further enhance the framework's practical applicability.

Bibliography

[1] R. Cooke, *Experts in uncertainty: opinion and subjective probability in science* (Environmental ethics and science policy series), Nachdr. New York: Oxford Univ. Press, 1991, 321 pp., ISBN: 978-0-19-506465-0.

[2] M. N. Jouini and R. T. Clemen, "Copula models for aggregating expert opinions," *Operations Research*, vol. 44, no. 3, pp. 444–457, Jun. 1996, ISSN: 0030-364X, 1526-5463. DOI: 10.1287/opre.44.3.444.

[3] N. Dalkey and O. Helmer, "An experimental application of the delphi method to the use of experts," *Management Science*, vol. 9, no. 3, pp. 458–467, 1963. DOI: 10.1287/mnsc.9.3.458.

[4] R. L. Winkler, "Combining probability distributions from dependent information sources," *Management Science*, vol. 27, no. 4, pp. 479–488, Apr. 1981, ISSN: 0025-1909. DOI: 10.1287/mnsc.27.4.479.

[5] A. Babuscia and K.-M. Cheung, "An approach to perform expert elicitation for engineering design risk analysis: Methodology and experimental results," *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 177, no. 2, pp. 475–497, Feb. 1, 2014, ISSN: 0964-1998. DOI: 10.1111/rssa.12028.

[6] A. M. Hanea, G. F. Nane, T. Bedford, and S. French, Eds., *Expert Judgement in Risk and Decision Analysis* (International Series in Operations Research & Management Science), en. Cham: Springer International Publishing, 2021, vol. 293, ISBN: 978-3-030-46473-8 978-3-030-46474-5. DOI: 10.1007/978-3-030-46474-5.

[7] C. Werner, T. Bedford, R. M. Cooke, A. M. Hanea, and O. Morales-Nápoles, "Expert judgement for dependence in probabilistic modelling: A systematic literature review and future research directions," en, *European Journal of Operational Research*, vol. 258, no. 3, pp. 801–819, May 2017, ISSN: 03772217. DOI: 10.1016/j.ejor.2016.10.018.

[8] A. Tversky, "Assessing uncertainty," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 148–159, 1974, ISSN: 2517-6161. DOI: 10.1111/j.2517-6161.1974.tb00996.x.

[9] M. Zellner, A. E. Abbas, D. V. Budescu, and A. Galstyan, "A survey of human judgement and quantitative forecasting methods," en, *Royal Society Open Science*, vol. 8, no. 2, rsos.201187, 201187, Feb. 2021, ISSN: 2054-5703. DOI: 10.1098/rsos.201187.

[10] T. McAndrew, N. Wattanachit, G. C. Gibson, and N. G. Reich, "Aggregating predictions from experts: A review of statistical methods, experiments, and applications," *Wiley interdisciplinary reviews. Computational statistics*, vol. 13, no. 2, e1514, 2021, ISSN: 1939-5108. DOI: 10.1002/wics.1514.

[11] D. Hartley and S. French, "A bayesian method for calibration and aggregation of expert judgement," *International Journal of Approximate Reasoning*, vol. 130, pp. 192–225, Mar. 1, 2021, ISSN: 0888-613X. DOI: 10.1016/j.ijar.2020.12.007.

[12] X. Wang, R. J. Hyndman, F. Li, and Y. Kang, "Forecast combinations: An over 50-year review," *International Journal of Forecasting*, vol. 39, no. 4, pp. 1518–1547, Oct. 2023, ISSN: 01692070. DOI: 10.1016/j.ijforecast.2022.11.005.

[13] F. Bolger and G. Rowe, "The aggregation of expert judgment: Do good things come to those who weight?" *Risk Analysis*, vol. 35, no. 1, pp. 5–11, 2015, ISSN: 1539-6924. DOI: 10.1111/risa.12272.

[14] R. M. Cooke, D. Marti, and T. Mazzuchi, "Expert forecasting with and without uncertainty quantification and weighting: What do the data say?" *International Journal of Forecasting*, vol. 37, no. 1, pp. 378–387, Jan. 1, 2021, ISSN: 0169-2070. DOI: 10.1016/j.ijforecast.2020.06.007.

[15] L. J. Savage, *The foundations of statistics*. New York: Dover Publications, 1972, ISBN: 978-0-486-62349-8 978-0-486-13710-0.

[16] A. H. Bajgiran, M. Mardikoraem, and E. S. Soofi, "Maximum entropy distributions with quantile information," *European Journal of Operational Research*, vol. 290, no. 1, pp. 196–209, Apr. 2021, ISSN: 03772217. DOI: 10.1016/j.ejor.2020.07.052.

[17] A. Hicks, J. Barclay, P. Simmons, and S. Loughlin, "An interdisciplinary approach to volcanic risk reduction under conditions of uncertainty: A case study of tristan da cunha," *Natural Hazards and Earth System Sciences*, vol. 14, no. 7, pp. 1871–1887, Jul. 28, 2014, ISSN: 1561-8633. DOI: 10.5194/nhess-14-1871-2014.

[18] A. Tadini, M. Bisson, A. Neri, R. Cioni, A. Bevilacqua, and W. P. Aspinall, "Assessing future vent opening locations at the somma-vesuvio volcanic complex: 1. a new information geodatabase with uncertainty characterizations," *Journal of Geophysical Research: Solid Earth*, vol. 122, no. 6, pp. 4336–4356, 2017, ISSN: 2169-9356. DOI: 10.1002/2016JB013858.

[19] A. Colson, R. M. Cooke, and R. Lutter, *How does breastfeeding affect IQ? applying the classical model of structured expert judgment*, Rochester, NY, Jul. 6, 2016. DOI: 10.2139/ssrn.2849605.

[20] E. Scourse, W. Aspinall, and N. Chapman, "Using expert elicitation to characterise long-term tectonic risks to radioactive waste repositories in japan," *Journal of Risk Research*, vol. 18, no. 3, pp. 364–377, Mar. 16, 2015, ISSN: 1366-9877, 1466-4461. DOI: 10.1080/13669877.2014.971334.

[21] T. Oraby, M. G. Tyshenko, M. Westphal, *et al.*, "Using expert judgments to improve chronic wasting disease risk management in canada," *Journal of Toxicology and Environmental Health, Part A*, vol. 79, no. 16, pp. 713–728, Sep. 1, 2016, ISSN: 1528-7394, 1087-2620. DOI: 10.1080/15287394.2016.1174005.

[22] P. Jaworski, F. Durante, W. K. Härdle, and T. Rychlik, Eds., *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25-26 September 2009*, vol. 198, Lecture Notes in Statistics, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, ISBN: 978-3-642-12465-5. DOI: 10.1007/978-3-642-12465-5.

[23] M. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Annales de l'ISUP*, vol. VIII, no. 3, pp. 229–231, 1959. [Online]. Available: https://hal.science/hal-04094463.

[24] R. B. Nelsen, *An Introduction to Copulas*. New York: Springer, 2006, ISBN: 978-0-387-28659-4. DOI: 10.1007/0-387-28678-0.

[25] T. Bedford and R. M. Cooke, "Probability density decomposition for conditionally dependent random variables modeled by vines," *Annals of Mathematics and Artificial Intelligence*, vol. 32, no. 1, pp. 245–268, Aug. 2001, ISSN: 1012-2443, 1573-7470. DOI: 10.1023/A:1016725902970.

[26] J. Dißmann, E. C. Brechmann, C. Czado, and D. Kurowicka, "Selecting and estimating regular vine copulae and application to financial returns," *Computational Statistics & Data Analysis*, vol. 59, pp. 52–69, Mar. 1, 2013, ISSN: 0167-9473. DOI: 10.1016/j.csda.2012.08.010.

[27] T. Nagler and T. Vatter, *Rvinecopulib: High performance algorithms for vine copula modeling*, 2025. [Online]. Available: https://vinecopulib.github.io/rvinecopulib/.

[28] H. Joe, *Multivariate Models and Multivariate Dependence Concepts*. CRC Press, May 1, 1997, 422 pp., ISBN: 978-0-412-07331-1.

[29] G. J. Székely and M. L. Rizzo, "The distance correlation t -test of independence in high dimension," *Journal of Multivariate Analysis*, vol. 117, pp. 193–213, May 2013, ISSN: 0047259X. DOI: 10.1016/j.jmva.2013.02.012.

[30] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The Annals of Statistics*, vol. 35, no. 6, pp. 2769–2794, Dec. 2007, ISSN: 0090-5364, 2168-8966. DOI: 10.1214/009053607000000505.

[31] G. Casella and R. Berger, *Statistical Inference*, 2nd ed. New York: Chapman and Hall/CRC, May 2024, ISBN: 978-1-003-45628-5. DOI: 10.1201/9781003456285.

[32] D. A. Moore, S. A. Swift, A. Minster, *et al.*, "Confidence calibration in a multiyear geopolitical forecasting competition," *Management Science*, vol. 63, no. 11, pp. 3552–3565, Nov. 2017, ISSN: 0025-1909, 1526-5501. DOI: 10.1287/mnsc.2016.2525. (visited on 06/30/2025).

[33] L. Sjöberg, "Are all crowds equally wise? a comparison of political election forecasts by experts and the public," *Journal of Forecasting*, vol. 28, no. 1, pp. 1–18, Jan. 2009, ISSN: 0277-6693, 1099-131X. DOI: 10.1002/for.1083. (visited on 06/30/2025).

[34] E. Beshearse, B. B. Bruce, G. F. Nane, *et al.*, "Attribution of illnesses transmitted by food and water to comprehensive transmission pathways using structured expert judgment, united states," *Emerging Infectious Diseases*, vol. 27, no. 1, pp. 182–195, Jan. 2021, ISSN: 1080-6040, 1080-6059. DOI: 10.3201/eid2701.200316.

[35] R. Cooke, M. Mendel, and W. Thijs, "Calibration and information in expert resolution; a classical approach," *Automatica*, vol. 24, no. 1, pp. 87–93, Jan. 1, 1988, ISSN: 0005-1098. DOI: 10.1016/0005-1098(88)90011-8.

[36] O. C. Mesner and C. R. Shalizi, "Conditional Mutual Information Estimation for Mixed, Discrete and Continuous Data," *IEEE Transactions on Information Theory*, vol. 67, no. 1, pp. 464–484, Jan. 2021, ISSN: 1557-9654. DOI: 10.1109/TIT.2020.3024886.

[37] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Vol. 1*. New York: Wiley-Interscience, 1994, 761 pp., ISBN: 978-0-471-58495-7.

[38] K. J. Wilson, "An investigation of dependence in expert judgement studies with multiple experts," *International Journal of Forecasting*, vol. 33, no. 1, pp. 325–336, Jan. 2017, ISSN: 01692070. DOI: 10.1016/j.ijforecast.2015.11.014.

[39] U. Cherubini, E. Luciano, and W. Vecchiato, *Copula methods in finance* (Wiley finance series), Reprinted. Chichester Weinheim: Wiley, 2006, 293 pp., ISBN: 978-0-470-86344-2.

[40] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, Sep. 1995, ISSN: 1064-8275, 1095-7197. DOI: 10.1137/0916069.

[41] P. Xue-Kun Song, "Multivariate dispersion models generated from gaussian copula," *Scandinavian Journal of Statistics*, vol. 27, no. 2, pp. 305–320, 2000, ISSN: 1467-9469. DOI: 10.1111/1467-9469.00191.

[42] D. Lewandowski, D. Kurowicka, and H. Joe, "Generating random correlation matrices based on vines and extended onion method," *Journal of Multivariate Analysis*, vol. 100, no. 9, pp. 1989–2001, Oct. 2009, ISSN: 0047259X. DOI: 10.1016/j.jmva.2009.04.008.

[43] C. J. F. ter Braak and J. A. Vrugt, "Differential Evolution Markov Chain with snooker updater and fewer chains," en, *Statistics and Computing*, vol. 18, no. 4, pp. 435–446, Dec. 2008, ISSN: 1573-1375. DOI: 10.1007/s11222-008-9104-9.

[44] F. Hartig, F. Minunno, and S. Paul, *Bayesiantools: General-purpose mcmc and smc samplers and tools for bayesian statistics*, 2023. [Online]. Available: https://CRAN.R-project.org/package=BayesianTools.

Appendix A

Appendix

A.1 Primitive Function of Sigmoid Composed Affine Function

In case $\phi_d(q, F)$ is an affine function $\epsilon_d(q, F)$ wrt to q (constant derivative) and is composed with the logistic function

$$\sigma(x) = \frac{1}{1 + e^{-kx}}.$$

That is, if

$$\phi_d(q, \mathbf{F}) = \sigma(\epsilon_d(q, \mathbf{F})),$$

then, a primitive function of ϕ wrt q is

$$G_{\sigma(\epsilon_d)}(q, \mathbf{F}) = \frac{\log(1 + e^{k\epsilon_d(q,\mathbf{F})})}{C_d k} \tag{A.1}$$

where $C_d = \frac{d}{dq}\epsilon_d(q, \mathbf{F})$ and is constant by assumption.

Proof. We drop the subscript d for notational clarity. Taking the derivative we get:

$$\frac{d}{dq}G(q) = \frac{1}{Ck} \cdot \frac{d}{dq}\log(1 + e^{k\epsilon(q)}) = \frac{1}{Ck} \cdot \frac{e^{k\epsilon(q)} \cdot k\epsilon'(q)}{1 + e^{k\epsilon(q)}},$$

$$= \frac{\epsilon'(q)}{C_d^e} \cdot \frac{e^{k\epsilon(q)}}{1 + e^{k\epsilon(q)}} = \frac{1}{1 + e^{-k\epsilon(q)}} = \phi(q)$$

Thus, $G(q)$ is a primitive of $\phi(q)$. □

For the squared sigmoid function:

$$\sigma^2(x) = \frac{1}{(1 + e^{-kx})^2}.$$

we can similarly show that

$$G_{\sigma^2(\epsilon_d)}(x) = \frac{1}{C_d k(1 + e^{k\epsilon_d(x)})} + \frac{\log(1 + e^{k\epsilon_d(x)})}{C_d k} \tag{A.2}$$

is a primitive function of $\phi^2(q, \mathbf{F}) = \sigma^2(\epsilon(q, \mathbf{F}))$ with respect to q.

Proof. We drop the subscript d for notational clarity. Taking the derivative we get:

$$\frac{d}{dq}G_{\sigma^2(\epsilon)}(q) = \frac{d}{dq}\left[\frac{1}{Ck(1+e^{k\epsilon(q)})} + \frac{\log(1+e^{k\epsilon(q)})}{Ck}\right] \tag{A.3}$$

$$= \frac{-k\epsilon'(q)e^{k\epsilon(q)}}{Ck(1+e^{k\epsilon(q)})^2} + \frac{k\epsilon'(q)e^{k\epsilon(q)}}{Ck(1+e^{k\epsilon(q)})} \tag{A.4}$$

$$= \frac{\epsilon'(q)e^{k\epsilon(q)}}{C(1+e^{k\epsilon(q)})^2}\left[-1+(1+e^{k\epsilon(q)})\right] \tag{A.5}$$

$$= \frac{e^{k\epsilon(q)}}{(1+e^{k\epsilon(q)})^2}\cdot e^{k\epsilon(q)} \tag{A.6}$$

$$= \frac{e^{2k\epsilon(q)}}{(1+e^{k\epsilon(q)})^2} = \left(\frac{e^{k\epsilon(q)}}{1+e^{k\epsilon(q)}}\right)^2 = \sigma^2(\epsilon(q)) \tag{A.7}$$

□

A.2 Mean of Piecewise Continuous Function

If random variable X has a piecewise constant density

$$f(x) = \begin{cases} 0 & \text{if } x \le m_0^e \\ a_1, & \text{if } x \in (m_0^e, m_1^e] \\ \vdots & \vdots \\ a_{D+1}, & \text{if } x \in (m_D^e, m_{D+1}^e] \\ 0 & \text{if } x \ge m_{D+1}^e \end{cases}$$

for a strictly increasing set of cutoff points $m_0^e, \dots, m_{D+1}^e$ and positive values $a_1, \dots, a_{D+1}$. Then the mean is given by

$$E[X] = \frac{1}{2}\sum_{i=1}^{D+1} a_i(m_i^{e2} - m_{i-1}^{e2}).$$

This comes directly from integration:

$$\int_{-\infty}^{\infty} xf(x)dx = \sum_{i=1}^{D+1} a_i \int_{m_{i-1}^e}^{m_i^e} xdx = \frac{1}{2}\sum_{i=1}^{D+1} a_i(m_i^{e2} - m_{i-1}^{e2}).$$

If f is specified in terms of strictly increasing quantile values $p_0, \dots p_{D+1}$ with $p_0 = 0$, $p_{D+1} = 1$ such that its CDF F is equal to

$$F(x) = \begin{cases} 0 & \text{if } x \le m_0^e \\ p_{i-1} + (p_i - p_{i-1})\frac{x - m_{i-1}^e}{m_i^e - m_{i-1}^e} & \text{if } x \in (m_{i-1}^e, m_i^e] \\ 1 & \text{if } x \ge m_{D+1}^e \end{cases}$$

then $a_i = \frac{p_i - p_{i-1}}{m_i^e - m_{i-1}^e}$ for $i = 1, \dots, D+1$ and we can write the mean as

$$E[X] = \frac{1}{2}\sum_{i=1}^{D+1}(p_i - p_{i-1})\frac{(m_i^{e2} - m_{i-1}^{e2})}{m_i^e - m_{i-1}^e} = \frac{1}{2}\sum_{i=1}^{D+1}(p_i - p_{i-1})(m_i^e + m_{i-1}^e).$$

Furthermore $E[g(X)]$ is given by

$$E[g(X)] = \sum_{i=1}^{D+1} a_i \int_{m_{i-1}^e}^{m_i^e} g(x)dx = \sum_{i=1}^{D+1} a_i \left(G(m_i^e) - G(m_{i-1}^e)\right) \tag{A.8}$$

where $G(x)$ is a primitive function of $g(y)$. Expressed in terms of CDF interpolation points we have

$$E[g(X)] = \sum_{i=1}^{D+1} \frac{p_i - p_{i-1}}{m_i^e - m_{i-1}^e} \left(G(m_i^e) - G(m_{i-1}^e)\right)$$

If $g(x)$ is $\sigma(\epsilon(x))$ as in Appendix A.1, then we have

$$E[\sigma(\epsilon(X))] = \sum_{i=1}^{D+1} \frac{p_i - p_{i-1}}{(m_i^e - m_{i-1}^e)Ck} \log \frac{1 + e^{k\epsilon(m_i^e)}}{1 + e^{k\epsilon(m_{i-1}^e)}}$$

A.3 Independence Invariant of Invertable Transformations

Random variables $X : \Omega \to E_X$ and $Y : \Omega \to E_Y$ are independent if and only if $g(X)$ and Y are independent for any invertable function $g : E_X \to E_Y$.

Proof ($\Rightarrow$): For any $A \subseteq E_X$ and $B \subseteq E_y$ we have

$$P(g(X) \in A, Y \in B) = P(X \in g^{-1}(A), Y \in B)$$

Where $g^{-1}(A) = \{g^{-1}(a) : a \in A\}$. Using that X and Y are independet

$$P(X \in g^{-1}(A))P(Y \in B) = P(g(X) \in A)P(Y \in B).$$

The proof is analogous for the other difrection.

A.4 Sampling Parameters

In Listing 1 you can see the parameters used for the sampling of unnormalized densities. The assumed variables in the listing are `starting_population` that have 100 points sampled from each $\hat{f}^e$ density, `chain_start_values` that has 50 points sampled from each $\hat{f}^e$ density, `log_density` that is a function that evaluated the log of the target density, `L_star` that is the lower bound L^* and `U_star` that is the upper bound U^*.Note that because we are only sampling from univarate distributions, multiple parameter choices are redundant. These include for example the `pSnooker` argument that is.

A.5 Vine Copula Fitting Implementation

The relevant parameters for the vine copula estimation are provided in the Appendix A.5 using the `vinecop` function of the `rvinecopulib` R package.

Listing 2 Setup of vine copula estimation

```
rvinecopulib::vinecop(
  data,
  var_types = rep("c", NCOL(data)),
  family_set = c("onepar", "indep),
  structure = NA,
  par_method = "mle",
  nonpar_method = "constant",
  mult = 1,
  selcrit = "aic",
  weights = numeric(),
  psi0 = 0.9,
  presel = TRUE,
  allow_rotations = TRUE,
  trunc_lvl = Inf,
  tree_crit = "tau",
  threshold = 0,
)
```

A.6 Copula Estimation Failure Rates

Table A.1 presents the failure rates for different copula estimation methods across all studies in the dataset. The table shows the number of cases where estimation failed (NA Count), the total number of estimation attempts, and the corresponding failure percentage. MAP estimation methods show varying failure rates depending on the η parameter and threshold settings, while Frank copula, vine copula, and some threshold-based methods achieved perfect reliability with zero failures.

Table A.1: Copula estimation failure rates by method

Method	NA Count	Total	Percentage NA
MAP:$\eta(50)$	144	299	48.2%
MAP:$\eta(1)$	139	299	46.5%
MAP:$\eta(10)$	135	299	45.2%
MAP:$\eta(1)$:$\tau_{\text{threshold}}(0.5)$	55	299	18.4%
MAP:$\eta(10)$:$\tau_{\text{threshold}}(0.5)$	48	299	16.1%
MAP:$\eta(50)$:$\tau_{\text{threshold}}(0.5)$	45	299	15.1%
MAP:$\eta(1)$:$\tau_{\text{threshold}}(0.7)$	2	299	0.7%
MAP:$\eta(50)$:$\tau_{\text{threshold}}(0.7)$	2	299	0.7%
Frank	0	299	0.0%
MAP:$\eta(10)$:$\tau_{\text{threshold}}(0.7)$	0	299	0.0%
Vine	0	299	0.0%
Vine:$\tau_{\text{threshold}}(0.5)$	0	299	0.0%
Vine:$\tau_{\text{threshold}}(0.7)$	0	299	0.0%

A.7 Addtional Marginal Estimation Data

A.7.1 Marginal Estimation Sample Sizes

Table A.2 mainly presents the number of total test samples made using LOOCV for the the different decouplers. This differs because of the use of expert rejection preprocessing. It also shows if any of the methods had numerical procedures that did not converge. This was only true for the MLE method.

Table A.2: Marginal estimation number of LOOCV samples and number of numerical failures by decoupler and method

Decoupler	**Method**	**Failures**	**Total**
Rel.Md.$k = 0.05$	MLE	3	4317
Rel.Md.$k = 0.05$	PE prior	0	4332
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.1$	0	4332
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.25$	0	4332
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.5$	0	4332
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.75$	0	4332
Rel.Md.$k = 0.1$	MLE	4	4203
Rel.Md.$k = 0.1$	PE prior	0	4230
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.1$	0	4230
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.25$	0	4230
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.5$	0	4230
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.75$	0	4230
Rel.Md.$k = 0.5$	MLE	4	4168
Rel.Md.$k = 0.5$	PE prior	0	4204
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.1$	0	4204
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.25$	0	4204
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.5$	0	4204
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.75$	0	4204
Rel.Md.$k = 1$	MLE	4	4188
Rel.Md.$k = 1$	PE prior	0	4222
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.1$	0	4222
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.25$	0	4222
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.5$	0	4222
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.75$	0	4222
CDF	MLE	0	4468
CDF	PE prior	0	4468
CDF	$\sigma_{\text{prior}} = 0.1$	0	4468
CDF	$\sigma_{\text{prior}} = 0.25$	0	4468
CDF	$\sigma_{\text{prior}} = 0.5$	0	4468
CDF	$\sigma_{\text{prior}} = 0.75$	0	4468

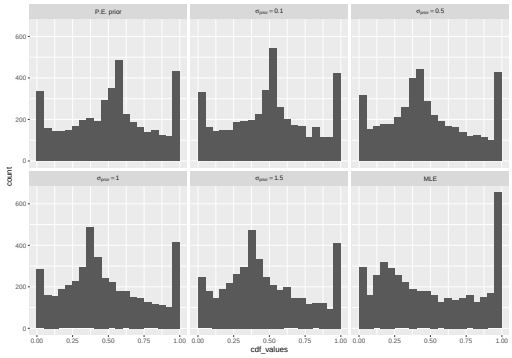
A.7.2 Complete Marginal Estimation Performance Results

Table A.3 presents the complete performance results for all marginal estimation methods tested across different decouplers and settings. The table shows L1 and L2 error rates for all combinations of decouplers (CDF, Rel.Md. with various k values) and estimation approaches (MLE, MAP with different σ_{prior} values, and PE prior). Results are sorted by L1 error in ascending order.

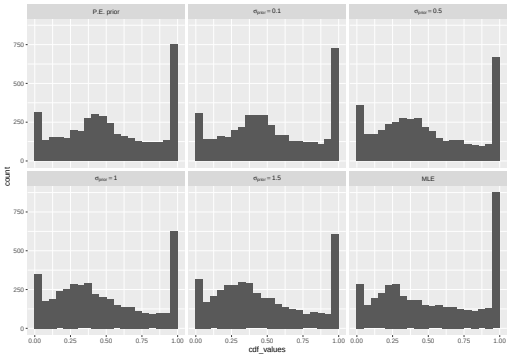
Table A.3: Complete marginal estimation performance results across all methods and settings

Decoupler	**Settings**	$\mathbf{MAE_{Median}}$	$\mathbf{RMSE_{Mean}}$	$\mathbf{L^1_{Unif}}$	$\mathbf{L^\infty_{Unif}}$
Rel.Md.$k = 0.05$	PE prior	0.018	0.053	0.035	0.065
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.1$	0.017	0.052	0.035	0.065
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.25$	0.060	0.126	0.037	0.130
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.25$	0.017	0.052	0.038	0.086
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.1$	0.059	0.127	0.039	0.132
Rel.Md.$k = 0.1$	PE prior	0.060	0.128	0.041	0.133
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.5$	0.018	0.052	0.041	0.097
Rel.Md.$k = 0.05$	$\sigma_{\text{prior}} = 0.75$	0.018	0.052	0.043	0.101
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.5$	0.061	0.125	0.045	0.125
Rel.Md.$k = 0.05$	MLE	0.018	0.052	0.046	0.105
CDF	MLE	0.346	0.383	0.049	0.096
Rel.Md.$k = 0.1$	MLE	0.062	0.126	0.050	0.169
Rel.Md.$k = 0.1$	$\sigma_{\text{prior}} = 0.75$	0.062	0.125	0.051	0.120
Rel.Md.$k = 0.5$	PE prior	0.140	0.203	0.055	0.176
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.1$	0.140	0.201	0.056	0.174
Rel.Md.$k = 0.5$	MLE	0.146	0.201	0.057	0.171
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.25$	0.140	0.199	0.061	0.168
CDF	$\sigma_{\text{prior}} = 0.75$	0.341	0.378	0.062	0.119
Rel.Md.$k = 1$	MLE	0.204	0.253	0.063	0.178
CDF	$\sigma_{\text{prior}} = 0.5$	0.339	0.377	0.067	0.131
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.5$	0.143	0.198	0.069	0.158
Rel.Md.$k = 0.5$	$\sigma_{\text{prior}} = 0.75$	0.145	0.198	0.072	0.152
Rel.Md.$k = 1$	PE prior	0.200	0.252	0.075	0.202
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.1$	0.199	0.250	0.075	0.198
CDF	$\sigma_{\text{prior}} = 0.25$	0.337	0.373	0.076	0.164
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.25$	0.199	0.247	0.078	0.190
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.5$	0.201	0.247	0.082	0.178
Rel.Md.$k = 1$	$\sigma_{\text{prior}} = 0.75$	0.204	0.248	0.082	0.170
CDF	$\sigma_{\text{prior}} = 0.1$	0.335	0.371	0.084	0.205
CDF	PE prior	0.336	0.371	0.087	0.229

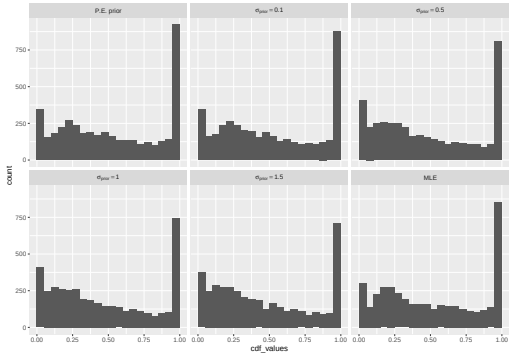
A.7.3 Marginal Density Estimation Plots



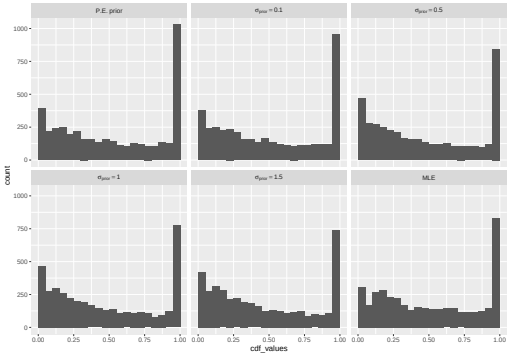
(a) k = 0.05



(b) k = 0.1

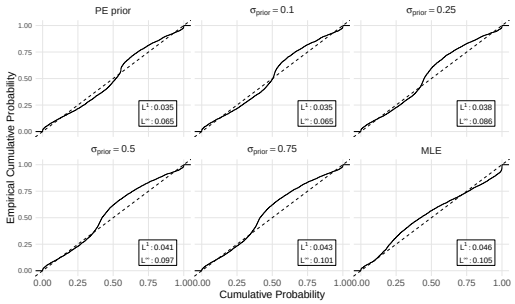


(c) k = 0.5

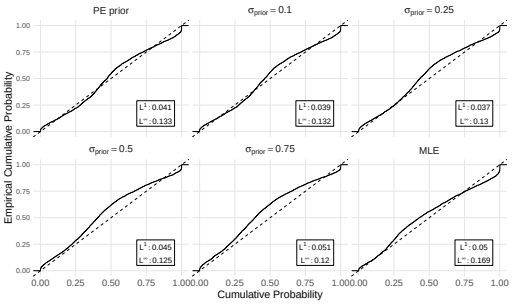


(d) k = 1

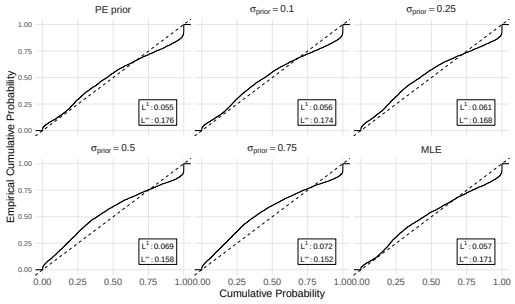
Figure A.1: Histogram comparisons of marginal density estimation methods for different sigmoid scaling k values of $\phi_{\sigma(\text{Rel.Md.})}$



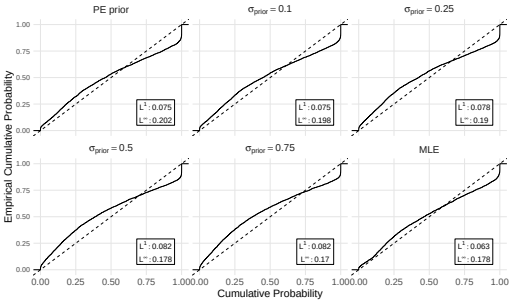
(a) k = 0.05



(b) k = 0.1



(c) k = 0.5



(d) k = 1

Figure A.2: Marginal density estimation comparisons for different sigmoid scaling k values of $\phi_{\sigma(\text{Rel.Md.})}$

Listing 1 Sampling steup using DEzs algorithm

```
starting_population # Sapmle 100 points from each belief function
chain_start_values # Sample 50 points from each belief function

bayesian_setup <- BayesianTools::createBayesianSetup(
  log_density,
  lower = L_star,
  upper = U_star
)

bayesian_settings <- list(
  iterations = num_samples,
  startValue = matrix(chain_start_values, nrow =
   ↪  length(chain_start_values), ncol = 1)
)

samples <- BayesianTools::DEzs(
  bayesianSetup,
  settings = list(
    iterations = 10000,
    Z = matrix(starting_population, length(starting_population), ncol =
     ↪  1),
    startValue = matrix(chain_start_values, nrow =
     ↪  length(chain_start_values), ncol = 1),
    pSnooker = 0.1,
    burnin = 100,
    thin = 1,
    f = 2.38,
    eps = 0,
    parallel = NULL,
    pGamma1 = 0.1,
    eps.mult = 0.2,
    eps.add = 0,
    consoleUpdates = 100,
    zUpdateFrequency = 1,
    currentChain = 1,
    blockUpdate = list(
      "none",
      k = NULL,
      h = NULL,
      pSel = NULL,
      pGroup = NULL,
      groupStart = 1000,
      groupIntervall = 1000
    ),
    message = TRUE
  )
)
```