A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Laying Algebraic-Geometric Foundation for Probabilistic Models of Natural Languages

Akihiro Maeda

Supervisor: Shohei Hidaka

Japan Advanced Institute of Science and Technology
Division of Advanced Science and Technology

[Information Science]

March 2026

Contents

Abstract

Human language exhibits a remarkable form of compositionality: complex expressions are systematically built from simpler parts, and this structure supports strong out-of-distribution generalization. Despite the empirical success of modern neural language models, no existing approach provides a principled probabilistic account of how such compositional structure is represented, computed, or learned. This dissertation addresses this gap by developing a mathematical framework that treats linguistic probability itself as an algebraic and geometric object.

The first contribution is the formulation of a *structural probability model*, grounded in algebraic statistics. Sentences are represented as joint probability tensors whose algebraic constraints correspond to geometric entities. This provides the first unified framework in which the compositional organization of language is expressed as algebraic system and appears as low-dimensional geometric structure in distributional representations.

The second contribution is the introduction of a new invariant unit of probabilistic structure, termed the *Minimum Invariant Constraint (MIC)*. An MIC is defined as the atomic algebraic component of a probability tensor, mathematically characterized by a vanishing 2×2 minor. MICs generalize the notions of independence; they are invariant under reparameterization, robust across corpora, and serve as the irreducible building blocks for complex structural patterns. This framework provides a principled explanation for the local rank-one patterns observed in PMI matrices and co-occurrence statistics.

The third contribution is the development of computational methods for discovering MICs and their compositions in empirical data. Two complementary approaches are introduced: (i) a divide–and–conquer method based on marginalization identities and cumulants, which interprets PMI heuristics as second-order cumulants; and (ii) a harmonic-analysis method using the Walsh–Hadamard transform and geometric algebra, enabling efficient detection of symmetric and low-rank structure in high-order tensors. Proof-of-concept experiments demonstrate that these methods recover invariant components that were previously inaccessible to conventional tensor decompositions.

Together, these contributions establish a new algebraic–geometric foundation for linguistic compositionality. They show that the internal organization of language emerges as invariant algebraic constraints on probability distributions, offering both theoretical insight into the structure of language and practical tools for analyzing modern language models.

Keywords: language model, algebraic statistic, invariant constraint, compositionality, vanishing binomial

Acknowledgments

I would like to express my deepest gratitude to my primary supervisor, Prof. Shohei Hidaka. For the past five years, his mentorship has been nothing short of exceptional. Through our weekly discussions, he not only provided compassionate guidance and precise advice but also respected my autonomy, allowing me to pursue research driven by my own intellectual curiosity. Whenever I faced difficulties, he guided me from an elevated perspective, offering profound insights that constantly reshaped my understanding. Without his patience and broad vision, this thesis would not have been possible.

I am also deeply indebted to Prof. Takuma Torii, who has been a constant source of support as a senior colleague. His active participation in the weekly discussions with Prof. Hidaka provided me with invaluable feedback. I am also grateful for his efforts in creating opportunities for me to present my research, which significantly contributed to my growth as a researcher.

My sincere thanks go to my secondary supervisor, Prof. Ryuhei Uehara. At times when my research seemed to reach an impasse, he provided me with opportunities to broaden my horizons. His guidance was pivotal in helping me acquire essential skills in computational theory and algorithms, which became a cornerstone of my work.

I would also like to extend my gratitude to Prof. Yohei Oseki for accepting me as a visiting student at the University of Tokyo. The three years I spent in his linguistics laboratory were incredibly stimulating, and I benefited greatly from his insightful advice and the academic environment he provided.

I am grateful to Prof. Naoya Inoue for his deep understanding of my research and his generous support regarding conference presentations. I also thank Prof. Miho Fuyama for offering me to participate in her Quantum Cognition Project, which provided me with valuable research opportunities.

This work was supported by the JSPS Research Fellowships for Young Scientists (DC2). I am thankful for the financial support that allowed me to focus on my doctoral studies.

Finally, words cannot express my gratitude to my wife, Junko. When I made the reckless decision to quit my job and pursue a career in academia, she accepted it without a single word of objection. For five years, she allowed me to prioritize my research over household responsibilities and endured the instability of this path. More than anything, I thank her for believing in me and supporting me every step of the way. This dissertation is dedicated to her.

Chapter 1

Introduction

1.1 Background and Motivation: Compositionality Matters

Language as a structured system of meaning Human language is more than a sequence of symbols—it is a system that allows the construction of infinitely many expressions from a finite set of vocabulary. This ability, known as *compositionality*, means that the meaning of a whole expression depends on the meanings of its parts and the rules used to combine them [103]. It underlies the productivity and systematicity of language, where productivity refers to the unbounded capacity to generate novel expressions, and systematicity denotes the semantic correlation between structurally related expressions: for example, one who understands "Ann loves Bob" should also understand "Bob loves Ann" [49]. From a cognitive standpoint, compositionality explains how humans generalize from limited data and adapt flexibly to new environments, achieving the kind of out-of-distribution generalization characteristic of human reasoning [13].

The challenges of language models In computational linguistics, one of the central goals is to understand and capture such human capacity by modeling language probabilistically [80]. Large language models such as GPTs [119], built upon the Transformer architecture [138]—a form of deep neural network—, have achieved remarkable performance using massive data and computation. However, their inner mechanisms remain largely opaque. These models generate fluent text but offer little principled account of how they understand and compose meanings. In contrast, linguistically motivated language models attempted to encode explicit syntactic rules, but they suffered from the curse of dimensionality because the number of required rules grows exponentially with sentence length and syntactic variation.

Motivation for a mathematical foundation Neural language models, including Transformer-based architectures, embed words into a vector space and compute meanings on these representations compositionally. Furthermore, word embeddings—most famously Word2Vec [95]—exhibit striking geometric properties. For example, similar words lie close to each

other in the vector space, words in the same semantic category form clusters, and semantic relations such as analogies (e.g., *king:queen::man:woman*) correspond to geometric regularities, often approximated by a parallelogram relation $\boldsymbol{v}(king) - \boldsymbol{v}(queen) \approx \boldsymbol{v}(man) - \boldsymbol{v}(woman)$.

These embeddings, including internal representations of large language models, are learned from corpora and believed to encode statistical properties of word co-occurrence. More precisely, word embeddings can be interpreted as capturing geometric relations that arise from probability distributions in co-occurrence data [136], which correspond mathematically to matrix factorizations of PMI (pointwise mutual information) derived from co-occurrence matrices [82].

Such observations indicate that word distributions reveal latent syntactic and semantic structures whose geometric patterns reflect underlying algebraic regularities among probabilities. This suggests that the probability distributions underlying language possess an internal organization that might explain the compositionality of language. To make this organization explicit, this dissertation develops a mathematical foundation that unifies algebra, geometry, and probability, thereby elucidating the structure inherent in language.

Statistics, Algebra and Geometry In mathematics, several cross-disciplinary fields have been developed to bridge statistics, algebra, and geometry. Algebraic geometry studies geometric objects by describing them through polynomial equations. Algebraic statistics applies tools from algebraic geometry to probability distributions, enabling the analysis of structured statistical models that arise, for example, in biological and social data. Geometric algebra (Clifford algebra) provides algebraic operations for representing and manipulating geometric configurations. Representation theory connects abstract algebraic structures—most prominently groups—with concrete linear transformations, revealing how symmetry governs computational structure. Both are actively studied for applications in geometric machine learning in computer vision. These mathematical frameworks have rarely been applied in computational linguistics, although early structural linguistics (e.g., Zellig Harris) hinted at algebraic perspectives on language. Such mathematical tools are potentially relevant to understanding and modeling the structural regularities that underlie natural language.

1.2 Research Problem and Approach

Research problem While human language is compositional, existing language models do not provide a formal mechanism by which linguistic expressions are combined through algebraic operations whose outputs remain meaningful. Transformers and other neural architectures compose sequences of tokens, but they do not implement *algebraic compositionality*, lacking the mathematical guarantee that combining linguistic units corresponds to a systematic and interpretable operation in meaning space[93]. A central challenge, therefore, is to develop a language model in which compositionality is not an emergent

byproduct of large-scale training, but an explicit structural principle.

Concurrently, distributed representations such as word embeddings exhibit clear geometric and algebraic regularities—cluster structure, linear relations, and low-rank patterns—suggesting that statistical data encode aspects of linguistic organization. However, these emergent mathematical structures do not straightforwardly align with the formal properties posited in linguistic theory or language models, thereby failing to provide a coherent explanation for these observed phenomena. A significant challenge is thus to reconcile these two structural domains: relating the algebraic–geometric patterns observed in distributional data to the formal linguistic properties that language itself seems to possess.

Research questions These gaps motivate several fundamental questions about what mechanism brings the algebraic and geometric patterns observed in modern representations and how language generative system should be modeled. This dissertation therefore centers around the following research questions:

1. **What kind of probabilistic model can capture algebraic and geometric structure in linguistic data?** This question seeks a formulation in which compositionality is encoded as a structural property of probability distributions—rather than as an emergent statistical artifact—so that the combination of linguistic units corresponds to explicit algebraic operations.

2. **How can the structural properties of language be formulated in precise mathematical terms?** The aim here is to represent observable regularities in probabilistic data as mathematically defined objects—such as invariances, low-rank constraints, or algebraic equations—that characterize the geometry of linguistic probability tensors and identify the minimal building blocks of compositional structure.

3. **Given such a formulation, can we interpret word embeddings and large language models in a genuinely compositional way?** This question investigates whether the resulting algebraic–geometric framework can account for structural phenomena observed in distributed representations—for example, the parallelogram structure in word analogies—and thereby explain why certain aspects of compositionality appear in current neural models.

Approach and framework A sentence is modeled as a joint probability distribution over its constituent words, making the combinatorial structure of language explicit at the probability level. In this view, the way words combine is reflected in how their joint distribution factorizes, a dependency structure that can be represented by a bipartite graph defined by its incidence matrix. This combinatorial structure, in turn, imposes multilinear algebraic constraints on the probability tensor associated with the sentence. These algebraic constraints induce geometric structure (such as low-dimensional varieties) within the space of all possible probability distributions. This dissertation employs tools from algebraic statistics and algebraic geometry to analyze and characterize these geometric structures.

Highlights of the main results

1. **Formulation of an Algebraic Language Model** This dissertation establishes the first rigorous theoretical framework to represent linguistic probability distributions as probability tensors and conceptualize their structure as geometric objects known as algebraic varieties, providing a unified mathematical framework bridging from sentence structures, to probability tensors, to the geometric relations in distributed word representations. The framework models language with a probability model called a Toric model (a subclass of exponential families) and then re-integrates it into an algebraic system (such as Clifford algebra with the log semiring). This formulation provides a mathematical basis connecting the compositional structure of language—expressed as algebraic operations in an underlying algebraic system—with the low-dimensional geometric patterns that emerge in vector-space representations.

2. **Discovery of Minimum Invariant Constraints (MICs)** Through a novel attempt to decompose probabilistic structures into their irreducible algebraic components, this work discovers *Minimum Invariant Constraints (MICs)*, defined as the smallest irreducible representation. MICs are the fundamental atoms of statistical regularities and act as the building blocks of invariant structures within a probability distribution, such as conditional independence. Each MIC corresponds to a zero determinant, or a vanishing 2×2 minor, of a probability tensor and encodes a generalized independence constraint. This provides a principled explanation for the local statistical patterns observed in word co-occurrence distributions as manifestations of latent algebraic structure.

3. **Invariant-based Structure Learning Methods** Moving beyond theoretical discovery, this dissertation develops novel principles to directly identify algebraic structures latent in data. These methods detect invariant structure in high-order probability tensors using algebraic signatures derived from polynomial constraints, instead of relying on the optimization of objective functions. These methods enable the discovery of MICs and larger invariant components in empirical data, and proof-of-concept experiments on toy models demonstrate their viability.

These results collectively reveal that linguistic compositionality manifests as invariant algebraic structure embedded within probability distributions, forming the conceptual and theoretical foundation for the theory and methods developed in the subsequent chapters.

1.3 Contributions and Novelty

The most original and groundbreaking contribution of this research is its proposal of a paradigm shift in the study of language. This conceptual shift reframes the probability distribution of language itself as a mathematical object possessing explicit algebraic-geometric structure.

This approach fundamentally diverges from conventional language models. Whereas approaches explicitly incorporating symbolic rules (e.g., Probabilistic Context-Free Grammars) face combinatorial explosion, and modern neural models learning statistical patterns from large-scale data grapple with a lack of interpretability (Transformer based models such as GPTs), this research redefines the problem setting itself. Rather than merely refining existing models, it establishes the inherent structure within linguistic probability distributions as the direct object of analysis. This opens a new avenue for mathematically elucidating the compositionality of language.

(1) Structural probability model The first contribution lies in the theoretical formulation of a unified structural probability model, which conceptualizes linguistic probability distributions as algebraic-geometric objects. The core of this theory is the capacity to represent and analyze the whole set of sentences as algebraic objects. Under this formulation, the joint probability tensor exhibits a complex of local regularities that corresponds to a Segre variety, or a secant variety as some mixture of Segre varieties.

Our proposed model has an impact that it establishes a unified mathematical framework that connects compositional structure in language with algebraic geometry. The framework theoretically elucidates that the structure of linguistic probability distributions is a geometric one, necessarily induced by the algebraic constraints inherent within them.

For novelty, to the best of our knowledge, no existing model for languages—including graphical models, compositional distributional semantics, or neural architectures—treats linguistic probability itself as an algebraic–geometric variety. Thus, this dissertation provides the first rigorous bridge between algebraic statistics and computational linguistics at the level of probabilistic structure.

(2) Minimum Invariant Constraint (MIC) The second contribution is the definition of the Minimum Invariant Constraint (MIC): the minimal algebraic unit of invariance in a probability tensor, mathematically represented as a vanishing 2×2 minor. MICs serve as the irreducible building blocks from which independence, conditional independence, and more complex forms of dependency patterns are composed.

The definition of MICs prompts a conceptual shift: from merely describing statistical correlations to identifying the invariant constituent units of the probabilistic structure itself. Its mathematical and theoretical value lies in detecting invariance and irreducibility, which explain persistent properties inherent to the linguistic probability structure, independent of specific datasets or parameters. This provides a unified explanation for the local low-rank patterns observed in data, such as in PMI matrices.

Our definition of the MIC as minimum irreducible representation gives a novel way to describe a statistical regularity and to generalize a notion of independence. This is also the first characterization of local rank-one structure as a direct sum of minimal idempotent projectors in Clifford algebra. No related existing work on independence—neither Bayesian networks nor Markov random fields—decomposes probabilistic structure into algebraic

irreducible components.

(3) Tensor-based invariant learning The third contribution is developing two complementary computational methods for discovering MICs and their compositions in high-order probability tensors: (i) Spatial approach: a method based on moments and cumulants that involves the direct analysis of the probability tensor. It employs statistical operations, such as marginalization (summing probabilities over a subset of variables) and cumulants (quantities capturing intrinsic dependence), to directly detect locally present 2×2 vanishing minors within the tensor. This is an intuitive and interpretable method that captures the structure in situ. (ii) Harmonic analysis approach: a method based on the Walsh-Hadamard Transform that analyzes the probability tensor from an alternative perspective. It utilizes the a kind of Fourier Transform to convert the probability tensor from the spatial domain to the frequency domain. In the frequency domain, algebraic properties such as symmetry and low-rank structure become more explicit, enabling their efficient detection and the capture of global structure.

The practical impact of these methods expected to be immense. They make the latent algebraic structures of language computationally discoverable from data. It will also reveal that PMIs, a conventional heuristics in natural language processing, is a special case of cumulants. In the future, these techniques hold the potential to become powerful tools for analyzing the internal representations of large language models and enhancing their interpretability.

As far as I know, no existing NLP or machine-learning method uses vanishing ideals or Clifford–Walsh harmonic analysis for structure discovery. The idea of using invariant constraints for structural learning, instead of optimizing cost functions for parameter learning, is novel. The proposed approach is the first to connect cumulants, toric geometry, and invariant tensor analysis within a unified algorithmic framework, enabling principled detection of compositional structure in probability data.

Summary Together, these contributions provide a new algebraic–geometric foundation for compositionality in natural languages: linguistic structure emerges as invariant algebraic constraints on probability distributions. This perspective yields both theoretical insight and practical tools for understanding the geometry of language and the internal organization of modern language models.

Structure of the Dissertation: Each chapter begins with a brief overview that clarifies its role within the overall framework.

- Chapter 2 surveys existing language models from a compositional perspective.

- Chapter 3 develops the proposed compositional probability model and its mathematical foundations.
- Chapter 4 presents tensor-based learning methods for detecting invariant structures.
- Chapter 5 applies the proposed theory to identify semantic rank-one patterns in PMI matrices that reflect invariant constraints.
- Chapter 6 concludes the dissertation and discusses remaining issues and future directions toward an algebraic language model.

Chapter 2

Survey: Language Models from Compositional View

2.1 Overview of the Chapter

Building upon the motivation and theoretical gap identified in Chapter 1, this chapter surveys how the principle of compositionality has been treated across major paradigms of language modeling. From symbolic and probabilistic traditions that emphasized explicit grammatical rules to neural approaches that achieve remarkable fluency through large-scale statistical learning, research in natural language processing has oscillated between structure and flexibility. Neither approach alone, however, explains how linguistic form and meaning interact to produce systematic generalization. This chapter therefore examines how different paradigms realize—or fail to realize—*structure-preserving mappings* between syntax and semantics, clarifying what is required for a model to connect compositionality and probability within a unified mathematical framework. It thereby prepares the ground for the theoretical development in Chapter 3.

To achieve this aim, the chapter traces the historical evolution of language modeling from rule-based and count-based systems to deep neural architectures, analyzing how each embodies or neglects compositional generalization. Representative models—including n-gram and probabilistic grammars, distributional embeddings such as LSA and Word2Vec, Bayesian and graphical models, and contemporary Transformers—are reconsidered in terms of their handling of homomorphism, systematicity, and productivity. In addition, recent evaluation methodologies for compositionality are reviewed, ranging from intrinsic probing and extrinsic generalization benchmarks to behavioral assessments of large language models. This comparative analysis reveals both the conceptual continuity and the persistent deficiencies that characterize the evolution of language modeling.

The findings converge on several key observations. Symbolic models succeed in capturing grammatical structure but remain rigid and data-hungry, whereas neural models achieve impressive flexibility and fluency but at the cost of interpretability. Despite their differences, none of the existing paradigms provides a mathematically consistent theory

that connects compositionality, probability, and learning. The survey therefore reframes the longstanding challenge of learning linguistic structure from data as a *representation-theoretic inverse problem*: given observable linguistic outputs, one must infer the latent algebraic mapping that generates them. This interpretation explains the recurring phenomena of data sparsity, over-parameterization, and instability in both symbolic and neural systems.

Recognizing this inverse-problem nature points directly toward the theoretical development that follows. The next chapter introduces the *compositional probability model*, which regularizes this ill-posed learning problem by invoking the symmetries inherent in language as mathematical constraints. The requirements extracted from the present survey—explicit structural representation, probabilistic consistency, and algebraic interpretability—serve as the design principles for the algebraic framework that forms the foundation of the dissertation's theoretical contribution.

2.2 Survey Methodology

2.2.1 Compositional Generalization in Large Language Models

Large Language Models (LLMs) have acquired language capabilities comparable to, or in some cases surpassing, those of humans [2, 129, 123]. However, no consensus has yet been reached on the fundamental question of whether LLMs use language in the same way humans do [112, 142], or if they are merely mimicking the statistical patterns present in their vast training data [12, 93]. The core of this question is whether these models have acquired human-like generalization capabilities; that is, the ability to systematically combine known knowledge to handle novel situations not seen in the training data (Out-of-Distribution; OOD [13, 125]). This capability is known as *Compositional Generalization*, and its achievement is regarded as an essential challenge in assessing the generalization power of language models [76, 72].

An LLM is, by definition, a large-scale language model, and a language model is a statistical model that predicts the occurrence of words. The questions of how *Compositionality* is formulated and learned within such probabilistic models, and how compositional generalization is subsequently acquired, were common challenges for language models even before the advent of LLMs. Motivated by this challenge, this study aims to clarify the conditions necessary for a language model to possess compositional generalization by surveying the historical progression of language models.

2.2.2 The Principle of Compositionality and Its Mathematical Foundations

Compositionality originates from Frege's Principle [50], which states that "the meaning of the whole is determined by the meanings of its parts and the rules of their combination," and was later formulated as the principle of compositionality through Montague's formal semantics [103, 109].From a cognitive science perspective, [49] pointed out that human language and thought possess *Productivity*—the ability to generate infinite sen-

tences from finite elements—and *Systematicity*—the ability to understand and produce novel expressions by recombining known constituents. They advocated that these are essential properties inherent to compositional systems.Furthermore, [117] proposed the Generative Lexicon theory, which posits that as word meanings dynamically shift based on co-occurring words (Co-compositionality), compositional properties are also context-dependent.In this study, we distinguish two aspects of compositionality: (1) the combinatorial property that "the whole (a sentence) consists of its parts (words or phrases)," and (2) the structure-preserving property that "the composition of the meanings of the parts equals the meaning of the whole." We then rigorously formulate the latter property using the mathematical concept of *Homomorphism*.A homomorphism is defined as a structure-preserving map from one algebraic system (a set equipped with structure) to another.Whereas the combinatorial property relates to a single system, a homomorphism establishes a correspondence between two systems.Compositionality in language can be formulated as the meaning assignment function μ from a syntactic (formal) algebraic system $(S, +)$ to a semantic algebraic system $(M, *)$ being a homomorphism, which must satisfy the condition $\mu(s_1 + s_2) = \mu(s_1) * \mu(s_2)\ \forall s_1, s_2 \in S$ (2.1) [66].This mathematical formulation via homomorphism provides a fundamental perspective for analyzing how language model paradigms have attempted to achieve compositionality.A crucial viewpoint is whether a language model merely possesses combinatorial properties, or if it incorporates constraints equivalent to a homomorphism.

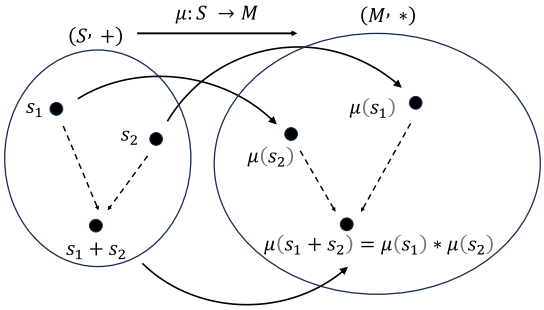


Figure 2.1: Formulating compositionality as a homomorphism. The homomorphism $\mu : S \to M$ provides a structure-preserving correspondence from the syntactic algebraic system $(S, +)$ to the semantic algebraic system $(M, *)$.

2.2.3 Approach

Therefore, this study uses the mathematical understanding of compositionality—specifically as homomorphism—as an analytical lens. We survey how compositionality has been formulated across the historical progression of language models. Our approach does not focus on quantitatively comparing the performance of individual models. Instead, its primary objective is to trace the evolution of the design philosophies behind these paradigms. By analyzing the presence, absence, or degree to which homomorphism is realized in their mathematical frameworks, we aim to elucidate how language models

have addressed the compositionality inherent in language and the challenges they have encountered.

This analysis reveals that the seemingly diverse difficulties faced by historical models—*combinatorial explosion, data sparsity, and lack of interpretability*—in fact converge on a single, common mathematical challenge. While language models attempt to learn the latent structure of language from observable data, identifying the homomorphism essential to compositionality requires establishing a correspondence between multiple algebraic systems. This task is mathematically difficult and inherently involves *ill-posedness*. By adopting the perspective of homomorphism, the root of the fundamental difficulties faced by each language model can be explained in a unified manner.

2.3 Models Based on Symbolic Representation

2.3.1 Conditional Probabilities of Symbol Sequences

The first attempts at language modeling were n-gram models, which treat words or characters as symbols and formulate sequences of n symbols using conditional probabilities [128]. The probability of the t-th symbol w_t occurring is given by Equation (2.1) as a conditional probability (a Markov chain) conditioned only on the $n-1$ preceding symbols $w_{t-n+1}, \ldots, w_{t-1}$:

$$Pr(w_t \mid w_{t-n+1}, \ldots, w_{t-1}). \tag{2.1}$$

The challenge for word n-grams is data sparsity. Because many words occur infrequently (Zipf's Law) [102], most of the n-word combinations needed for estimation rarely, if ever, appear in the corpus. This causes the maximum likelihood estimates (MLE) of their probabilities to be zero or the estimates themselves to be unreliable. As countermeasures, smoothing techniques (e.g., Kneser-Ney smoothing [73]) were proposed to compensate for unobserved data using counts from observed, smaller n-grams.Data sparsity is an essential limitation inevitably faced by any finite corpus. It arises because the observed events are only a small subset of what is possible, given the productivity of language (the ability to generate infinite novel sentences).At the root of this problem lies combinatorial explosion: as n increases, the space of possible combinations grows exponentially, leading to an explosive increase in the number of model parameters that must be estimated. This is a manifestation of the curse of dimensionality in high-dimensional spaces, which simultaneously exacerbates data sparsity and introduces issues of *computational intractability*.From the perspective of compositionality, n-gram models are limited to probabilistically capturing local combinations of symbol sequences (i.e., the Markov chain). This model does not explicitly formulate, within its structure, the constraints equivalent to a homomorphism that would map symbolic combinations to semantic compositions. As a result, it has been pointed out that word sequences stochastically generated by n-gram models using MLE from a corpus are often meaningless [102](p135). This suggests that an approach modeling only the local combinatorial properties of symbol sequences, while lacking the

constraints equivalent to a homomorphism, cannot account for meaning and is insufficient for achieving compositionality.

2.3.2 Introduction of Syntactic Structure

One model that formulates sentence probability using symbol sequences equipped with syntactic structure is the Probabilistic Context-Free Grammar (PCFG) [92].A PCFG is a probabilistic model designed to parse ambiguous sentences. It operates by pre-assigning probabilities to the syntactic rules that form a syntactic tree structure, and then selecting the combination of rules that yields the highest overall probability.Classical studies [26] calculate the probability of a sentence as the product of the probabilities of its constituent rules, using maximum likelihood estimates (MLE) based on the frequency of those rules in a corpus.Similar to n-gram models, as sentences become longer, the search space for possible tree structures grows exponentially. Consequently, PCFGs also face combinatorial explosion and data sparsity.PCFGs, by definition, assume that syntactic rules are context-free. However, in reality, syntactic choices are often context-dependent (e.g., the ambiguity in prepositional phrase attachment, as in *She saw a man with a telescope*).To address this, more refined models were proposed, such as those introducing grammatical rules or syntactic structures that reflect context-dependency (e.g., Combinatory Categorial Grammar [33], Dependency Grammar [94]), or Lexicalized PCFGs that introduce syntax rules specific to vocabulary [35].While accommodating context-dependency improved parsing performance, it also increased the total number of rules to be considered, thereby exacerbating the problems of combinatorial explosion and data sparsity.Evaluating PCFGs from the perspective of compositionality clarifies their contributions and limitations.The syntactic rules defined by a PCFG encapsulate a homomorphism between words and syntactic categories. Therefore, unlike n-gram models, PCFGs can generate (at least) syntactically correct sentences.However, this mapping does not consider semantics. As such, it cannot exclude semantically anomalous sentences like *Colorless green ideas sleep furiously.* It lacks the homomorphism between the syntactic (S) and semantic (M) systems.Furthermore, just like n-gram models, any attempt to refine the syntactic rules led to severe combinatorial explosion and data sparsity.This indicates that the symbolic modeling approach not only failed to introduce a mapping to semantics, but also that the approach of explicitly learning structure—even just at the syntactic level of compositionality—involves inherent computational difficulties.As a side note, grammar induction—the attempt to learn the grammar itself from data rather than relying on hand-crafted rules—was similarly a difficult problem due to the vastness of the search space. However, neural PCFGs (e.g., Recurrent Neural Network Grammar [43]) have been proposed based on deep learning (discussed later).

2.4 Models Based on Distributed Representation

2.4.1 Representing Inter-Symbol Relationships in a Distributed Manner

The concept of representing words in a distributed manner is long-standing. Influenced by the distributional hypothesis [60], early research in psychology and cognitive science established that vector representations derived from word co-occurrence counts encode semantic and syntactic information [79].A representative method of this count-based approach is Latent Semantic Analysis (LSA), proposed by [77]. LSA applies Singular Value Decomposition (SVD) to a word-document co-occurrence matrix to acquire dense, dimensionally-reduced distributed representations.[15] proposed a novel method for obtaining distributed representations using neural networks, demonstrating that its performance surpassed that of conventional n-gram models.Furthermore, [14] pointed out that while traditional symbolic representations are one-hot (an n-dimensional vector with only one component set to 1 and the rest 0), in distributed representations, each dimension within the vector is involved in expressing many features, and the combinations thereof exponentially increase the expressive power.The most well-known word distributed representation is Word2Vec, proposed by [96].For instance, in the model known as Skip-gram, the probability of a context word c appearing given a center word w_t is modeled using two types of distributed representations (the center word vector $\boldsymbol{v}_{w_t}$ and the context word vector $\boldsymbol{u}_c$) as shown in Equation (2.2). The representations are then learned by maximizing the log-likelihood (V denotes the vocabulary).

$$p(c \mid w_t) = \frac{\exp(\boldsymbol{u}_c \cdot \boldsymbol{v}_{w_t})}{\sum_{c' \in V} \exp(\boldsymbol{u}_{c'} \cdot \boldsymbol{v}_{w_t})}. \tag{2.2}$$

In the learned representations, it was observed that in addition to the property that semantically similar words have geometrically proximate vectors, linear algebraic relationships hold for word vectors in analogy tasks, such as $\boldsymbol{v}_{king} - \boldsymbol{v}_{queen} \approx \boldsymbol{v}_{man} - \boldsymbol{v}_{woman}$.These findings—that distributed representations capture the structural relationships of meaning with high precision—were met with surprise.In essence, this suggests that a structure-preserving correspondence—a homomorphism—exists between the linear algebraic structure of the vector space and the semantic relationships within the set of linguistic expressions. This demonstrated that a property analogous to compositionality was present at the word level.

2.4.2 Composition Operations for Word Vectors

In the same way that words combine to form phrases, and phrases combine to form sentences, Compositional Distributional Semantics (CDS) attempts to combine two word vectors to create vectors for phrases or sentences.Let W be the set of linguistic expressions (words, phrases) and V be a vector space, with a map $\phi : W \to V$ from expressions to vectors.The goal of CDS is to identify a composition function $* : V \times V \to V$ between

vectors such that, for two expressions (e.g., red, $car \in W$) and their combined expression (e.g., $red\_car \in W$), the following holds (where _ represents a syntax-driven composition operation in W) [7]:

$$\phi(red) * \phi(car) = \phi(red\_car) \tag{2.3}$$

As candidates for the composition function, early research explored operations such as addition [77], weighted addition, element-wise multiplication [100], tensor products [130], and circular convolution [113].To evaluate these models, comparisons were made between the right-hand side of Equation (2.3) (i.e., the phrase vector $\phi(red\_car)$ learned separately from the corpus) and the resulting composite vector on the left-hand side ($\phi(red) * \phi(car)$). Other evaluations compared the rank correlation between the similarities of phrase vectors computed via the left-hand side and human similarity perception.The results reported a tendency for simple addition or element-wise multiplication to outperform other, more complex operations [80].Conversely, [7] pointed out the limitations of such simple "mixture" models and advocated for a formulation more faithful to formal semantics (specifically, Montague's type-driven functional composition).This approach involves assigning types based on parts of speech—for example, using vectors for nouns and sentences, matrices for adjectives modifying nouns, and higher-order tensors for adverbs modifying adjectives. A series of models adhering to this recommendation was subsequently proposed [34, 56, 32].The body of research on CDS was a theoretical attempt to directly introduce compositionality by identifying an explicit homomorphism between the combination operations in linguistic expressions and the vector operations in distributed representations. However, this attempt was based on mathematically naive assumptions.Specifically, the problem formulation itself—which assumes the existence of a composition function $*$ satisfying Equation (2.3) and attempts to identify it empirically from limited data—likely suffered from severe data sparsity and non-identifiability (the uniqueness of the solution is not guaranteed), rendering it mathematically underdetermined.Furthermore, the introduction of higher-order tensors in accordance with sophisticated linguistic (type-driven) theories led to increased model complexity.Consequently, this approach faced the familiar problems of combinatorial explosion and data sparsity, as the number of parameters to be estimated grew exponentially.Moreover, type-driven theories were criticized for side effects that undermined the original advantages of distributed representations. For instance, because derived words with different parts of speech (e.g., *destroy* and *destruction*) are mapped to tensors of different orders, their semantic similarity is lost in the vector space.Whereas CDS empirically searched for a fixed composition function $*$, [131] attempted an approach where the composition function itself was replaced by a neural network and learned from data.This model used pre-computed parse trees and composed word vectors bottom-up along this structure using a Recursive Neural Network (Recursive NN) to obtain phrase and sentence vectors.The approach was groundbreaking in that it made the composition function learnable while explicitly utilizing symbolic syntactic constraints.However, this strong dependence on parse trees limited the model size. Furthermore, its performance

was only comparable to simple additive models [18].In practice, this outcome spurred a shift toward end-to-end learning (e.g., RNNs and later Transformers) that does not explicitly use syntactic rules. This suggests an essential limitation: learning while preserving symbolic compositionality is computationally difficult.

2.5 Graphical Models

2.5.1 Formulation as a Joint Probability of Multiple Variables

A graphical model formulates the joint probability distribution of multiple variables by using a graph, where nodes represent the random variables and edges represent the conditional dependencies between them. Graphical models are models that explicitly assume structure in the form of these inter-variable dependencies.[22] used a directed graph, the Hidden Markov Model (HMM), to propose a class-based n-gram model. They formulated the transition probabilities of hidden variables corresponding to word classes using a bigram model, assuming that words were generated from these hidden states. When they assigned class classifications to maximize average mutual information, they found that the extracted hidden states tended to group based on either syntactic or semantic properties.Furthermore, [75] proposed a graphical model for modeling the sequential structure of language, using an undirected graph known as a Conditional Random Field (CRF). This discriminative model performs sequential part-of-speech (POS) labeling on word sequences.Unlike generative models such as HMMs, CRFs have the advantage of flexibly incorporating (and overlapping) sequence-related features without needing to model the probability distribution of the observation sequence (the word sequence) itself.Moreover, by performing global normalization, they also resolved the label bias problem inherent in directed discriminative models.[101] proposed a model that learns the joint probability distribution of n-grams using a Restricted Boltzmann Machine (RBM), which was developed by Geoffrey Hinton.An RBM is a type of undirected graphical model (an energy-based model) that has connections only between the layer of observed variables and the layer of hidden variables.This research, similar to [15], demonstrated that distributed representations of words could be learned using a graphical model, serving as a bridge between distributed representations and graphical models.The greatest strength of graphical models lies in their ability to explicitly assume structure, such as variable dependencies. The studies mentioned above achieved high interpretability that aligned with linguistic intuition.However, although theoretically capable of handling complex dependencies, in practice, the problem of computational cost (combinatorial explosion) restricted their application to extremely simple structures, such as first-order Markov processes.This created a fundamental limitation: they were, in principle, unable to capture the complexities of language, such as long-distance contextual dependencies.In addition, there were other issues: the high cost of manual feature engineering, and the difficulty of learning the graph structure itself from data, which meant that the results were strongly dependent on the

designer's initial assumptions.While graphical models succeeded in modeling structural relationships between symbols (like parts of speech), they were not used to model the compositionality involving the algebraic composition of semantic representations, such as vectors.Even when framed as the problem of finding a homomorphism between two graphs—one representing syntactic structure and the other semantic structure—this is generally an NP-hard problem [53].

2.5.2 Computational Graphs with Combinatorial Structure

The Sum-Product Network (SPN) [115], which conceptualizes probabilistic models as a combination of probability mixtures (sums) and conditional independencies (products), is an architecture that bridges deep learning and graphical models. Strictly speaking, SPNs are not graphical models but rather computational graphs that can efficiently compute conditional probabilities and marginalizations for multivariate probability distributions. They represent complex probability distributions by hierarchically combining Sum nodes (mixtures) and Product nodes (decompositions). Thus, SPNs provide a computational foundation equipped with compositionality, explicitly combining the two compositional operations of mixing and decomposition; a correspondence to graphical models has also been established.

A representative study applying SPNs to natural language [27] reported that their model, which learns word distributed representations by directly taking sentences as input, predicted the probability of a word at a target position comparably to competing models. It was also reported to have generated some interpretable sentences. However, it is difficult to conclude that these limited generation results are a manifestation in language of the theoretical compositionality (the mixing and decomposition operations) inherent in SPNs. Further verification is required.

2.6 Deep Learning Models

2.6.1 End-to-End Deep Learning

Models that relied on symbolic syntactic structures or hand-crafted composition functions faced a wall: combinatorial explosion and difficulty in learning. Consequently, around 2010, the research focus shifted to end-to-end deep learning models, which do not require explicitly given structures and instead learn directly from data.

The Recurrent Neural Network (RNN) is a deep learning model with a recursive structure, specialized for handling sequential data, which reuses the results computed in the hidden state for the next time step. Seq2Seq [29, 134], an encoder-decoder model proposed for machine translation tasks using LSTM (Long Short-Term Memory) (a type of RNN), encodes an input word sequence into a single fixed-length vector (the encoder), from which the decoder recursively generates output words probabilistically. RNN/LSTM models, including Seq2Seq, were expected to acquire representations encoding sentences

or phrases through the processing of vector sequences. However, it was reported that their performance was sometimes inferior to that of sentence vectors derived from simple vector addition [5, 36, 144]. These models also faced an information bottleneck: because they compress the entire sentence into one fixed-length vector, the accuracy of the translated sentence degrades as the input sentence length increases. This problem was observed to be particularly pronounced when encountering sentences longer than those in the training corpus [6].

In contrast, ELMo [111] took a different approach using the same LSTM architecture. It trained forward (next-word prediction) and backward (previous-word prediction) language models independently and then used the internal states (contextualized word vectors) obtained during this process, dramatically improving performance on downstream tasks (such as question answering). Unlike Seq2Seq, ELMo does not compress information into a single internal representation; instead, it uses the vector corresponding to each word as its distributed representation, thereby retaining rich contextual information for every word. Nevertheless, ELMo still relied on the RNN/LSTM architecture and thus inherited its common challenges. Specifically, because sentences are processed one word at a time, dependencies between words that are far apart are lost as the sentence length increases (the long-range dependency problem). Furthermore, the sequential computation requirement makes parallelization difficult, posing a constraint on computational speed.

2.6.2 Breakthroughs via the Attention Mechanism

To overcome the long-range dependency problem of RNNs/LSTMs and to enable computational parallelization, the Transformer architecture was proposed [138]. It is equipped with a self-attention mechanism that selectively references surrounding words as a condition for predicting the next word. Using the Transformer, both BERT [40], which encodes word sequences by referencing context bidirectionally, and GPT [120], which generates words by encoding unidirectionally, achieved high, general-purpose language processing capabilities across many downstream tasks.

On benchmarks designed to measure compositional generalization, such as SCAN and COGS (discussed in 2.7.1), Transformer-based models have shown high performance under specific conditions. However, they also exhibit instability, with performance dropping sharply when conditions are slightly altered [72, 76]. Consequently, evaluations are divided as to whether Transformers have acquired true compositionality. Interpretability research [122, 48] aimed at elucidating the internal mechanisms of pre-trained language models has suggested that the Transformer's internal representations may encode hierarchical structures, such as syntactic trees, and semantic roles [31, 135]. This provides a basis for believing that the models are implicitly learning some form of combinatorial structure from the data.

The observation that performance improves according to predictable power laws as model parameter count, data size, and computational resources increase (i.e., Scaling Laws

[70]) suggests that the data sparsity challenge is mitigated by large-scale data, and that the models may be memorizing a greater number of combinatorial patterns [142]. Furthermore, detailed analysis of the Grokking phenomenon—where generalization performance rapidly improves after a prolonged period of overfitting—suggests that a phase transition may be occurring inside the model, shifting from simple pattern memorization to a more general, systematic solution [106]. This also hints at the possibility that compositional structures are implicitly acquired during training. In contrast to the scaling approach, research is also being conducted on Syntactic LLMs that provide syntactic structure as an inductive bias [118, 147], and on BabyLM, which attempts to learn from small amounts of data [141].

As a recent functional enhancement of LLMs, In-Context Learning (ICL) [23] has been observed in Transformer-based models, where they can execute new tasks merely by being shown a few examples in a prompt. This appears to be a form of dynamic compositional generalization, where the model infers implicit rules from the provided examples and applies them to new inputs [58]. Additionally, Chain-of-Thought (CoT) prompting [143], by presenting intermediate reasoning steps, improves the model's ability to decompose complex problems into simpler sub-problems, which are then recomposed to derive a solution.

LLMs based on the Transformer architecture, through the investment of massive computational resources, appear to have overcome the challenges of combinatorial explosion and data sparsity that plagued symbolic models. Behaviors like ICL and CoT give the impression, at least superficially, that LLMs have acquired powerful compositional generalization. On the other hand, this has come at the cost of the model's internal processing becoming a black box, obscuring what linguistic structures the acquired knowledge corresponds to. Although recent studies have revealed that semantic relationships between words and meaning shifts due to contextualization exhibit linear relationships within the internal representations of LLMs [62, 91], the evaluation of their compositionality remains divided, and the problem that the fundamental principles of their behavior cannot be explained persists [93].

From the standpoint of this study—that the essence of compositionality is homomorphism—this is not considered a mere technical, unsolved problem. Rather, it is viewed as a manifestation of a more fundamental difficulty, which we will later discuss: the non-identifiability of learning, whereby the internal structure cannot be uniquely determined from the observable outputs. The behavior of LLMs appears compositional, but the question of whether it is truly compositional relates to the question of whether the homomorphism between syntax and semantics has been uniquely identified through learning from data.

2.7 Evaluation Methods for Compositionality

2.7.1 Measuring Compositional Generalization (Extrinsic Evaluation)

The most representative early dataset purported to evaluate compositional generalization is SCAN [76]. It is a benchmark designed to directly measure a model's ability to generalize to new combinations not seen during training, after being trained on a specific dataset. For instance, after learning simple commands like *walk left*, the model is tested on whether it can correctly execute unseen compound commands, such as *jump around left*. COGS [72] is a grammatically and semantically more complex dataset than SCAN. It measures the ability to generalize to sentences with different syntactic structures or lexical arrangements, such as whether a model trained on active-voice sentences can correctly perform semantic parsing on passive-voice sentences. CFQ [71] is a question-answering dataset based on a large-scale knowledge base. It evaluates compositional semantic parsing capabilities by intentionally eliminating the overlap of compound question combinations between the training and test data. While all these benchmarks measure compositional generalization by strictly controlling the combinations of constituents between training and testing, they face a noted limitation: they use artificially generated data based on limited syntactic rules, which may not sufficiently reflect the diverse compositionality found in natural language [63, 64].

2.7.2 Probing Internal Representations (Intrinsic Evaluation)

Methods exist (known as probing) to investigate whether a model's internal representations encode compositional linguistic knowledge, such as syntactic information or semantic roles, by training a simple auxiliary classifier on these representations. [84] proposed a task to probe whether a model captures hierarchical syntactic structures by evaluating its ability to correctly predict subject-verb number agreement in sentences. [46] proposed diagnostic tasks to evaluate the structural semantic understanding of Transformer-based models, examining, for example, whether they understand compositional semantic operations such as those involving the negation word *not*. Using BERT, [135] investigated which internal layers of the model correspond to different stages of a traditional NLP pipeline (e.g., POS tagging, parsing), suggesting that linguistic structures are represented hierarchically within the model. These are approaches that verify the encoding of compositional knowledge through the analysis of internal representations.

2.7.3 Evaluating Model Behavior (Operational Evaluation)

This method evaluates the compositional behavior of models by making systematic modifications to their input sentences and observing the resulting changes in their output. [121] proposed a method for semi-automatically generating test cases, such as checking whether a model's predictions remain robust when proper nouns or locations in a sentence are swapped, or whether predictions are inverted when negation is added. [63] decomposed

Paradigm	Language Model	Modeling of Compositionality	Learning Challenges
Symbolic Models	n-gram, PCFG	Lacks homomorphism constraints (n-gram), or is limited to the syntactic level (PCFG).	Constrained by data sparsity
Distributed Models	Word2Vec, CDS	Suggests word-level homomorphism (analogy). Attempts to empirically identify composition functions (CDS).	Underdetermined due to non-uniqueness and combinatorial explosion
Graphical Models	HMM, CRF, RBM, SPN	Explicitly introduces dependencies. Not used for algebraic composition of meaning.	Computational intractability; finding homomorphism is NP-hard
Deep Learning	Seq2Seq, ELMo, GPT, etc.	Appears to implicitly acquire compositional behavior via large-scale data/computation, but evaluations are divided.	Lack of interpretability and non-identifiability

Table 2.1: Analysis of the formulation of compositionality (viewed as a homomorphism to semantic algebra) and its learning challenges in each language model.

compositionality into five properties: systematicity and productivity, plus substitutivity (robustness to synonym substitution), locality (whether compositional operations are local or global), and overgeneralization (whether the model follows rules or exceptions). They evaluated major language models using a dataset generated according to a PCFG. Their findings revealed that each architecture's compositionality has different strengths and weaknesses, and also clarified that none of them fully learn a truly compositional generative system (e.g., they may rely on chunk-based processing). [107, 98] evaluate the degree of compositional generalization by systematically manipulating the training data. [146] proposed an analysis method focusing on the systematicity of reasoning through Natural Language Inference (NLI) tasks. They constructed an inference dataset comprising novel combinations of semantic components (logical words and content words) to evaluate systematic generalization, reporting that LLMs lack robust generalization capabilities. These methods attempt to evaluate compositionality not as a formal property, but as a property that closely resembles the behaviors observed in human language understanding and reasoning.

2.8 Analysis and Discussion

2.8.1 Paradigms and Structural Constraints

The models surveyed in Table 2.1 can be roughly divided into two paradigms. *Theory-driven* approaches explicitly build linguistic structure into the model (e.g., PCFGs), treating compositionality as hierarchical combination. They clarified many aspects of syntactic structure, but faced severe computational limits: the number of possible expressions grows exponentially with sequence length, while real corpora remain finite and sparse. As a result, it is practically impossible to fully infer the intended structure from data, and such models struggle to account for diverse, context-dependent language use.

In contrast, *data-driven* approaches such as Transformer-based large language models do not assume an explicit syntactic backbone. Instead, a flexible architecture, trained on massive corpora with large computational budgets, implicitly acquires generative structure. This yields striking empirical performance, but introduces new issues: training cost scales sharply with model and data size, and the relation between internal representations and linguistic structure remains opaque. Consequently, whether such models genuinely realize compositional generalization is still an open problem.

To organize these difficulties, we adopt a formal-semantics perspective [67]. Let the *syntactic algebra* S be a set of linguistic expressions (words, phrases, sentences) equipped with a binary operation of concatenation that generates complex expressions. Let the *semantic algebra* M be a set of meaning representations (e.g., entities, predicates, truth values) equipped with operations for semantic composition. A meaning assignment $\mu : S \to M$ is compositionally well-behaved if it is a homomorphism of algebras: the meaning of a complex expression is determined by the meanings of its parts and the mode of combination.

A compositional language model induces a third algebraic structure: the *representation space* V, typically a vector space with operations used by distributed representations and neural networks. We assume an *encoding* map $\pi : S \to V$ from syntax to representations and a *decoding* map $\Phi : V \to M$ from representations to meanings. Ideally, both π and Φ are homomorphisms, and the following diagram commutes:

$$\begin{array}{ccc} S & \xrightarrow{\mu} & M \\ {\scriptstyle\pi}\downarrow & \nearrow_{\Phi} & \\ V & & \end{array} \tag{2.4}$$

A language model is then compositional precisely when it realizes such a representation space V and homomorphisms π and Φ.

Learning a compositional model from data can thus be viewed as an *inverse problem*: given only a sparse subset $D \subset S$ of actually observed expressions and their empirical distributions, we seek to infer a suitable V together with maps π and Φ that make Equation (2.4) approximately hold. This is intrinsically hard because the hypothesis space induced by the combinatorial richness of S is extremely high-dimensional, while the available corpus is necessarily sparse. The tension between this high-dimensional hypothesis space and data sparsity is exactly the manifestation of the curse of dimensionality that underlies the limitations of both theory-driven and data-driven paradigms.

2.8.2 Essential Difficulty of the Inverse Problem: Ill-posedness

Inverse problems typically suffer from *ill-posedness*, in the sense that at least one of Hadamard's three conditions for a well-posed problem—existence, uniqueness, and stability of solutions—fails [44]. Language, with its inherently combinatorial structure, forces any inverse formulation of language modeling to confront the curse of dimensionality. The hypothesis

space grows exponentially, while available corpora remain sparse; consequently, many distinct internal structures can explain the same limited observations. This destroys uniqueness, yields non-identifiability of latent structure, and makes the learning problem fundamentally ill-posed. Moreover, learning requires approximating the two homomorphisms in the commutative diagram $\mu = \Phi \circ \pi$ simultaneously, which further exacerbates instability and non-uniqueness.

Treating language modeling as an ill-posed inverse problem is nevertheless fruitful, because it invites mathematical tools designed to overcome such difficulties. The root cause of ill-posedness is excessive degrees of freedom; thus the standard approach is *regularization*, i.e., imposing additional structural constraints to select a unique and stable solution. This acts as a form of preprocessing that suppresses spurious variation in the data by projecting it onto a subspace consistent with hypothesized structure (e.g., a symmetry). For language, exploiting algebraic symmetries inherent in its combinatorial structure provides a principled basis for such regularization and offers a path toward tractable and interpretable learning.

2.8.3 Formulating Language Modeling as an Inverse Problem in Representation Theory

What constitutes an appropriate regularization for the inverse problem of language modeling? A useful clue comes from the approximate homomorphic behavior observed in word vectors (Section 2.4.1). Distributed word representations reflect semantic relations: synonyms cluster, and analogy pairs such as *king:queen :: man:woman* form parallelograms. Recent work [136] suggests that these geometric patterns arise not from the learning algorithm but from structural regularities inherent in word co-occurrence statistics. Thus, the mapping $\Phi : V \to M$ from the representation space to the semantic algebra may already approximate a homomorphism. Underlying this phenomenon are semantic *symmetries*, such as gender or social-status distinctions [90]. For example, antonyms like *hot/cold* may appear with similar probabilities in certain templates (*"Today's weather is hot/cold"*), indicating an exchangeability that the semantic algebra M should capture. These observations suggest that M carries not only compositional operations but also transformations such as permutations, naturally modeled by group actions.

Representation theory provides the mathematical framework for analyzing such symmetry. A *representation* of a group G on a vector space V is a homomorphism $\rho : G \to GL(V)$, allowing abstract symmetries to manifest as linear transformations. This connects group-theoretic structure to geometric behavior in V: rotations, reflections, or permutations appear as linear maps. If meanings in M admit a group action $G \times M \to M$, then a representation ρ induces the equivariance condition

$$\Phi(\rho(g)v) = g \cdot \Phi(v), \qquad g \in G,\ v \in V,$$

formalizing how semantic relations become geometric relations in vector space. Even

analogy parallelograms can be expressed uniformly by incorporating a projective extension of $GL(V)$.

Seen through this lens, representation theory offers a principled route to regularization. Group symmetries impose structural constraints on V by organizing data into *orbits*, low-dimensional manifolds formed by points related under group actions. Treating points within the same orbit as equivalent and extracting features invariant under G dramatically reduces the effective hypothesis space, mitigating ill-posedness. Symmetry-induced constraints restrict the model to low-dimensional submanifolds within V, avoiding exhaustive search over an exponentially large combinatorial space and enabling stable learning from sparse data.

Thus, symmetry-based regularization amounts to introducing a strong inductive bias: assuming that language conforms to a particular algebraic structure, and choosing an appropriate group G, determines which invariances should guide learning. This viewpoint reframes language modeling as an inverse problem in representation theory, where resolving ambiguity and instability relies on uncovering the symmetries latent in linguistic data.

Chapter 3

Theory: Compositional Probability Model

3.1 Overview of the Chapter

This chapter develops the theoretical foundation of the compositional probability model, which unifies algebraic statistics, geometry, and probability into a single formal framework. Building upon the issues identified in Chapter 2, it reformulates linguistic compositionality as an invariant property within a structured probability space. By viewing probability distributions not as isolated numerical tables but as algebraic objects constrained by geometric relations, the chapter bridges symbolic, statistical, and geometric perspectives on language.

The model begins by representing language as a probability tensor, whose entries encode joint occurrences of linguistic variables. The structure of this tensor is governed by algebraic constraints that is vanishing binomials such as 2×2-minors (zero determinants), which specify when subsets of variables behave independently. Within this framework, a new concept—the *Minimum Invariant Constraint* (MIC)—is introduced as the atomic unit of structure: a local rank-one component that remains invariant under symmetry operations on the probability tensor. Each MIC corresponds to a point on a Segre variety, a low dimensional geometrical object in a high dimensional probability space, and, equivalently, to a minimal idempotent in Clifford algebra, revealing a deep correspondence between algebraic independence and geometric simplicity.

Through these constructions, the chapter demonstrates that independence, compositionality, and invariance are mathematically equivalent notions when viewed through the lens of algebraic geometry. Probabilistic, algebraic, and geometric perspectives coincide in the relations

$$\text{independence} \iff \text{factorization} \iff \text{rank-one structure} \iff \text{vanishing minor},$$

showing that the same constraint can be expressed as a probabilistic relation, an alge-

braic identity, or a geometric subvariety. This equivalence establishes a one-to-one correspondence between linguistic structure and algebraic invariants of probability, offering a principled way to define compositionality as invariance in the space of distributions.

The theoretical results obtained here provide the mathematical basis for the learning algorithms presented in Chapter 4, where these invariant structures are identified empirically in data. They also explain the algebraic regularities—such as low-rank and structured patterns in PMI matrices—that will be analyzed in Chapter 5. In this way, the present chapter serves as the conceptual and mathematical core of the dissertation, establishing the axioms and derivations upon which the subsequent computational and empirical studies are built.

3.2 Model

In this section, we develop a mathematical framework for constructing a probabilistic model of natural language based on algebraic and geometric principles. Our objective is twofold. First, we aim to capture the *compositional nature* of language—how the probability of a whole expression (e.g., a sentence) systematically relates to the probabilities of its constituent parts (e.g., words or phrases). Second, we pursue a *dual perspective* that integrates algebraic and geometric viewpoints: the linear structure underlying probability models and the geometric insights emerging from distributed representations.

To achieve this, we employ the methodology of *algebraic statistics*[42], which utilizes tools from algebraic geometry to study statistical models. We begin by formulating the probability of a sentence as a joint distribution over words. This joint distribution is then parameterized, through a so-called *toric model*, as an algebraic structure that expresses probabilistic constraints as polynomial relations on the probability simplex. These polynomial constraints reveal that statistical dependencies and independencies correspond to geometric structures—specifically, subvarieties such as Segre or Secant varieties[59].

The proposed formulation thus provides a unified framework that bridges linguistic composition, algebraic structure, and geometric representation of probability distributions. By grounding language modeling in this algebraic–geometric correspondence, we aim to make explicit the hidden structural regularities that underlie natural language probability spaces.

3.2.1 Incidence matrix to model language

Modeling sentences A central challenge to formalize probability model of language is to capture the compositional nature of language, where meaning is constructed by combining smaller units into larger structures. To formally investigate this phenomenon, it is crucial to adopt a representation that explicitly models these whole-part relationships. This section outlines a framework for modeling the relationship between words and sentences using a clear and mathematically tractable structure.

To achieve a compositional representation, we explicitly model the relationship between words (the parts) and sentences (the wholes) by employing an incidence matrix. An incidence matrix is defined as a binary matrix representing the inclusion relationship between two disjoint sets. This approach provides a direct and interpretable model of how basic linguistic units constitute larger utterances within a corpus.

Let V be the vocabulary set, containing all unique words in a corpus, and let N be the total number of sentences. The word-sentence structure of the corpus can then be represented by an incidence matrix $M \in \{0,1\}^{|V|\times N}$. Each row of M corresponds to a unique word $v_i \in V$, and each column corresponds to a sentence s^j. An entry M_{ij} of this matrix is defined as follows:

$$M_{ij} = \begin{cases} 1 & \text{if word } v_i \text{ is contained in sentence } s^j \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

This binary matrix serves as a fundamental representation of constituency, explicitly mapping elementary components to the larger structures they form. While n-gram and PCFG (Probabilistic Context Free Grammar) models are limited to local sequences or syntactically adjacent pairs respectively, our incidence model focuses on the word combination as an integral whole. This holistic view enables the simultaneous formulation of the probability of the sentence and the occurrence probability of its words. Note that, by abstracting away word order to model these combinations, this incidence-based representation essentially functions as a bag-of-words model. We adopt the convention of using a lower index for rows (words) and an upper index for columns (sentences) to make the correspondence explicit.

This incidence-based model is not only descriptive but also generative, as it allows for the derivation of fundamental statistical relationships. Specifically, it can be used to compute the word co-occurrence matrix, a cornerstone of distributional semantics. Let $\boldsymbol{p}$ be a probability vector of dimension N, where each element p^j represents the probability of the j-th sentence s^j. From this vector, we can construct a diagonal matrix $P \in \mathbb{R}^{N\times N}$ where the diagonal elements are the probabilities from $\boldsymbol{p}$ (i.e., $P_{jj} = p^j$ and $P_{jk} = 0$ for $j \neq k$).

The word co-occurrence matrix $C \in \mathbb{R}^{|V|\times|V|}$ can then be derived through the following matrix multiplication:

$$C = MPM^\top \tag{3.2}$$

Each element C_{ik} of the resulting matrix quantifies the frequency with which words v_i and v_k appear together in the same sentences. This demonstrates how a simple structural model based on incidence can directly yield rich statistical information about lexical relationships. In graph theory, it is known that for an incidence matrix M, the product $MM^\top$ can be decomposed into the sum of the degree matrix D and the adjacency matrix B, such that $MM^\top = D + B$ ([24] Theorem 2.3.1).

Example: A Toy Corpus and Its Incidence Matrix To illustrate the basic idea of representing sentences as compositional probability structures, we begin with a minimal SVO corpus where a sentence consists of subject(S), verb(V) and object(O). The vocabulary contains only six words:

$$S = \{ann, bob\}, \qquad V = \{eats, likes\}, \qquad O = \{fish, vegetable\}. \tag{3.3}$$

By taking the Cartesian product $S \times V \times O$, we obtain eight possible sentences:

$$\begin{array}{llll} \textit{Ann eats fish}. & \textit{Ann eats vegetable}. & \textit{Ann likes fish}. & \textit{Ann likes vege}. \\ \textit{Bob eats fish}. & \textit{Bob eats vegetable}. & \textit{Bob likes fish}. & \textit{Bob likes vegetable}. \end{array} \tag{3.4}$$

These will serve as the column set of an incidence matrix, while the rows correspond to the six lexical items.

The incidence matrix $M \in \{0,1\}^{6\times 8}$ encodes whether a word appears in a sentence (1) or not (0). Ordering the words as

$$\{ann, bob, eats, likes, fish, vegetable\},$$

and the sentences in the order listed above, the matrix becomes:

$$M = \begin{array}{c} \\ ann \\ bob \\ eats \\ likes \\ fish \\ vegetable \end{array} \begin{array}{c} \begin{array}{cccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array}$$

Each column describes one sentence as a binary vector over the vocabulary. This representation is precisely the *incidence matrix* of the bipartite relation between words and sentences, and serves as the design matrix for the corresponding Toric model discussed later.

This toy example provides a concrete reference point for the concepts introduced in the following sections. In particular, the SVO structure induces a $2 \times 2 \times 2$ probability tensor, whose rank-one constraints, vanishing 2-minors, and various independence properties will be analyzed in detail in the remainder of this chapter.

It is important to note that the incidence matrix above reflects only the surface co-occurrence between words and sentences. However, such a matrix does not fit real linguistic data, because natural language exhibits rich latent structure: semantic relations among words, context-dependent interactions, and abstract properties that are not directly observable at the lexical level. For example, if *ann* is interpreted as a "vegetarian," then her

distribution over objects should correlate negatively with *fish* and positively with *vegetable*; such effects are not recoverable from the word–sentence incidence alone. These semantic attributes can be treated as *latent variables* and introduced as additional rows in the design matrix, on the same footing as observable words. Likewise, polysemy or part-of-speech variation can be modeled by viewing the row vector of a word as being partially projected depending on the context in which the word appears. The goal, therefore, is to identify from corpus statistics the appropriate configuration matrix that best explains the data. This corresponds to performing *structural learning*: selecting the rows of the design matrix so as to uncover hidden semantic relations and contextual structure encoded in the underlying probability model.

Comparison vs NLP practices The word co-occurrence matrix is a cornerstone of distributional semantics and serves as a foundational element in many Natural Language Processing (NLP) tasks, including the generation of word embeddings like those from Word2Vec. Standard practice for constructing this matrix involves counting word co-occurrences within a sliding window of a fixed size (typically 5 to 10 words [68]), including applying weighting schemes that assign less importance to words that are further apart.

Our incidence matrix formulation naturally leads to a co-occurrence matrix, but it differs from these standard practices in several key aspects. The most significant distinction lies in the scope of co-occurrence counting. Whereas conventional methods use a fixed-size sliding window, our approach defines the scope as the entire sentence. This can lead to different counts; for instance, words that are far apart in a long sentence will be counted as co-occurring in our model, whereas they would not be in a narrow-window model.

Furthermore, the binary nature of our incidence matrix, $M \in \{0,1\}^{|V|\times N}$, does not account for multiple occurrences of the same word within a single sentence. This limitation can be readily addressed by extending the matrix entries to the set of natural numbers, $M \in \mathbb{N}^{|V|\times N}$, allowing for multiple counts.

A more fundamental consequence of our sentence-based approach is the loss of all positional information within the sentence, as it effectively treats each sentence as a bag of words. This raises the question of whether this loss of information negatively impacts the model's validity. We argue that this is not a critical drawback for two primary reasons. First, the translational symmetry inherent in sliding-window counting effectively neutralizes or averages out absolute positional information in conventional methods. Second, and more importantly, our framework is designed to re-introduce a more structured notion of order at later stages. We will incorporate relative positioning by treating linguistic features, such as Part-of-Speech (POS) tags and syntactic roles, as random variables. On top of this, we will re-introduce ordering and asymmetry through the algebraic structure of the model itself, specifically by employing a non-commutative Clifford algebra, as will be detailed in a subsequent section.

3.2.2 Toric Model for Language Modeling

Model description Our incidence matrix formalism provides the basis for modeling probability distributions in natural language. To this end, we introduce a *toric model*, which is a fundamental tool in the field of algebraic statistics. The toric model provides a framework for representing high-dimensional structured discrete distributions using tools from algebraic geometry. In particular, it is well suited to describe joint probability distributions over multiple discrete variables, and it naturally aligns with the incidence matrix introduced in the previous section to represent word–sentence relations. A toric model can be regarded as a specific subclass of the exponential family.

Let $\boldsymbol{p} \in \mathbb{R}^N_{\geq 0}$ be a probability vector over N joint states, satisfying $\boldsymbol{p}^\top \mathbf{1} = 1$. $\boldsymbol{p}$ is a point in the $(N-1)$-simplex, which is defined as the set $\Delta^{N-1} = \{\boldsymbol{x} \in \mathbb{R}^N \mid \boldsymbol{x} \geq \mathbf{0},\ \mathbf{1}^\top \boldsymbol{x} = 1\}$. The probability of the j-th state, denoted p^j, is given by

$$p^j = \frac{1}{Z} \exp\bigl(\boldsymbol{\theta} \cdot \boldsymbol{a}^j\bigr) \tag{3.5}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is a parameter vector, $\boldsymbol{a}^j \in \{0,1\}^d$ is a binary feature vector characterizing the association between the j-th state and the parameters, and $\cdot$ denotes the standard inner product.

The normalization constant (partition function) is

$$Z := \sum_{j=1}^{N} \exp\bigl(\boldsymbol{\theta} \cdot \boldsymbol{a}^j\bigr). \tag{3.6}$$

A homogeneity condition is often imposed so that the sum of the elements of each $\boldsymbol{a}^j$ is constant across all j. This constraint ensures that the resulting image of the model, when expressed in monomial form, lies on an algebraic torus in the probability simplex, which gives rise to the name *toric* model.

By defining the *design matrix*

$$A = \begin{bmatrix} \boldsymbol{a}^1 & \boldsymbol{a}^2 & \cdots & \boldsymbol{a}^N \end{bmatrix} \in \{0,1\}^{d \times N}, \tag{3.7}$$

the entire probability vector can be expressed compactly as

$$\boldsymbol{p}^\top = \frac{1}{Z} \exp\bigl(\boldsymbol{\theta}^\top A\bigr), \tag{3.8}$$

where the exponential is applied elementwise. The matrix A defines the structural configuration of the model and is sometimes called a *configuration matrix*. Because the log-probability vector $\log \boldsymbol{p}$ is linearly parameterized by $\boldsymbol{\theta}$ through A, the row space of A determines the parameter-invariant algebraic relations among the entries of $\boldsymbol{p}$.

Classification within the Exponential Family To situate the toric model within the broader landscape of statistical models, it is helpful to understand its relationship to the exponential family, which is defined as follows.

Given m random variables X_i for $i = 1, \dots, m$, with each taking a value in a state space as $x_i \in \mathcal{X}_i$, where $\mathcal{X}_i$ denotes the set of possible states, we consider the joint random vector $\boldsymbol{X} = (X_1, \dots, X_m)$. This vector takes values $\boldsymbol{x} = (x_1, \dots, x_m)$ in the Cartesian product space $\mathcal{X} = \prod_{i=1}^{m} \mathcal{X}_i$. Let $\phi = (\phi_\alpha)_{\alpha \in \mathcal{I}}$ be a collection of functions $\phi_\alpha : \mathcal{X} \to \mathbb{R}$, known as *potential functions* or *sufficient statistics*. Here, the index set $\mathcal{I}$ specifies the features that define the model structure. Corresponding to ϕ, let $\theta = (\theta_\alpha)_{\alpha \in \mathcal{I}}$ be an associated vector of *canonical* or *exponential* parameters, where $\mathcal{I}$ is an index set with $d = |\mathcal{I}|$ elements.

Then, the exponential family associated with ϕ consists of the following parametrized collection of density functions defined over the realizations $\mathbf{x} = (x_1, \dots, x_m) \in \mathcal{X}^m$ of the random vector $\mathbf{X}$:

$$p_\theta(\mathbf{x}) = \exp\big(\theta \cdot \phi(\mathbf{x}) - \psi(\theta)\big), \tag{3.9}$$

where $\langle \cdot, \cdot \rangle$ is a standard inner product. The quantity $\psi(\theta)$, known as the *log-partition function* or *cumulant generating function*, is the normalization constant defined by

$$\phi(\theta) = \log \sum_{\mathbf{x} \in \mathcal{X}^m} \exp\big(\theta \cdot \phi(\mathbf{x})\big) \tag{3.10}$$

for discrete random variables, or by the corresponding integral for continuous random variables. Although the *log-sum-exp* notation is standard in information geometry to exploit convexity and duality, the notation Z (representing the *sum-exp* form) is traditional in statistical mechanics. We adopt the latter convention throughout this dissertation consistent with the notation used in the previous paragraph.

A fundamental principle of the exponential family is that it comprises distributions that maximize entropy subject to constraints on the sufficient statistics [140]. This principle gives rise to a hierarchy of models ranging from the general to the specific. We introduce them in the following order:

1. The most general class is the **exponential family** itself.

2. A key subset is the **finite discrete exponential family**, which is equivalent to the **log-affine model** and is characterized by a design matrix with real-valued entries.

3. The **toric model** is, in turn, a subclass of the log-affine model where the entries of the design matrix are restricted to be non-negative integers. This integer constraint is what gives rise to the model's characteristic representation via a monomial map.

4. A further specialization is the **log-linear model** in contingency table analysis, which can be viewed as a specific toric model where the design matrix is binary, containing only entries of $\{0, 1\}$.

5. Finally, **hierarchical log-linear models** form a subset of log-linear models where

the interaction terms among random variables are determined by the combinatorial structure of a simplicial complex [133].

Justification for Using a Toric Model The distinctive advantage of a toric model in probability modeling lies in its ability to decompose a generative system into two components: the *parameter part*, which specifies numerical weights, and the *parameter-invariant structural part*, which can be represented by a design matrix. Our primary focus is on this invariant structural component, namely the *row space of the configuration (design) matrix*, which encodes the algebraic and combinatorial relations among probabilities. Furthermore, the use of a toric model for language modeling is justified for several reasons. It naturally captures the whole–part and combinatorial nature of linguistic composition through the joint probability of multiple discrete variables. This approach is equivalent to modeling the probability of a sentence as a joint probability of multiple random variables, each of which represents syntactic roles (grammatical functions) such as Subject (S), Verb (V), and Object (O).

$$P(\text{sentence}) = P(S, V, O). \tag{3.11}$$

The formulation as a joint probability is not only theoretically sound but also provides a powerful framework for statistical language modeling.

- *Exclusivity of Roles via Disjoint Vocabularies* : In standard syntactic parsing, the core grammatical functions of Subject, Verb, and Object are fulfilled by distinct constituents. To rigorously model this within a joint probability framework, we introduce a hard structural constraint: the sets of candidate words for each role are treated as disjoint. Even if a surface form w (e.g., a noun that can appear in multiple positions) functions as both a Subject and an Object, we theoretically distinguish them as distinct entities, denoted as w_S and w_O. While raw corpora conflate these forms, syntactic parsing allows us to recover this distinction. Consequently, the random variables corresponding to S, V, and O are defined over these disjoint sample spaces. This formulation ensures that the assignment of a word to a specific slot is unambiguous, thereby allowing us to treat the variables as structurally distinct components whose potential statistical independence or dependence can be explicitly modeled.

- *Modeling compositionality* : By treating S, V, O and other roles as random variables, the generative probability of a sentence, $P(\text{sentence})$, can be expressed as the joint probability of these variables taking on specific values (e.g., particular words or phrases), denoted as $P(S = s, V = v, O = o)$. This represents a significant departure from traditional n-gram models, which primarily capture the linear sequence of words. Instead, our approach directly incorporates the combinatorial structure of the sentence into the probabilistic model, thereby offering a more principled way to capture linguistic compositionality. Crucially, this framework inherently reflects compositionality: the joint distribution of m variables can be recursively decomposed

into subsystems (e.g., of size k and $m-k$). This perspective implies that local algebraic constraints within these subsystems naturally scale to shape the global linguistic structure.

- Potential refinements: Strictly speaking, the constituents filling the S and O roles are often phrases (e.g., noun phrases) rather than single words. For the sake of model simplicity, it is a common and viable strategy to let the random variable be represented by the *head word* of the constituent phrase. We acknowledge that this foundational SVO model is most powerful for simple, active declarative sentences. Extending the framework to handle more complex syntactic structures—such as passive constructions, subordinate clauses, or sentences with different argument structures—would require a corresponding extension of the set of syntactic roles and the introduction of additional random variables. Delineating this scope is crucial for interpreting the model's results and guiding future work.

Other mathematical advantages of using a Toric model include:

- The design matrix—which in our framework corresponds to the incidence matrix from the previous section—provides a compact representation of the compositional and hierarchical structures of sentence probabilities. In particular, hierarchical dependencies among linguistic contexts can be modeled by analyzing the algebraic structure of the row space of the design matrix.
- The model admits linear-algebraic analysis: the kernel of the design matrix reveals parameter-invariant constraints among probabilities, identifying algebraic relations independent of specific parameter values.
- From a geometric perspective, these constraints correspond to polynomial relations defining an ideal. The vanishing set of this ideal forms an algebraic variety—specifically, a toric variety—on which the probability distributions lie.

Thus, the toric model enables us to study probabilistic structures in language through the lens of algebraic geometry.

Equivalence of the Toric Model and the Softmax Function The Toric model, as utilized in this context, is a specific formulation within the exponential family of distributions. The probability p_i for a given state i is defined as:

$$p_i = \frac{1}{Z(\theta)} \exp\left(\theta \cdot a^i\right)$$

Here, θ represents the natural parameter vector, a^i is the column vector from the design matrix A corresponding to state i, and $Z(\theta)$ is the partition function, or normalization

constant, ensuring that $\sum_j p_j = 1$:

$$Z(\theta) = \sum_j \exp\left(\theta \cdot a^j\right)$$

It is crucial to recognize that this mathematical structure is formally identical to the softmax function, which is a standard component in modern deep learning architectures, particularly for multi-class classification.

We can establish a direct equivalence by defining the exponent term as the logit for state i:

$$\text{logit}_i = \theta \cdot a^i$$

This logit_i represents the unnormalized log-probability, which is derived from a linear combination of the parameters θ and the design features a^i.

Substituting this definition back into the probability equation yields the canonical form of the softmax function:

$$p_i = \frac{\exp\left(\text{logit}_i\right)}{\sum_j \exp\left(\text{logit}_j\right)} \equiv \text{softmax}(\boldsymbol{L})_i$$

where $\boldsymbol{L}$ is the vector of logits $(\text{logit}_1, \ldots, \text{logit}_N)$.

This interpretation is not merely a convenient analogy; it is a fundamental identity. The conceptual framework—wherein a linear structure in the logit space is mapped to a valid probability distribution on the simplex via the normalized exponential (softmax) function—is well-established.

3.2.3 Algebraic variety to represent a probability

Monomials, Ideals and Varieties The toric model, introduced in the previous section, can be viewed as *monomial mapping* from a parameter space to a probability space, using the same notation as in 3.2.2,

$$\phi_A : \mathbb{R}^d \to \mathbb{R}^N, \quad \text{defined by} \quad \boldsymbol{p} = \phi_A(\boldsymbol{\theta}), \tag{3.12}$$

since each entry of $\boldsymbol{p}$ is a monomial in the parameters,

$$p^j = \exp\langle \boldsymbol{\theta}, \boldsymbol{a}^j \rangle = (e^{\theta_1})^{a_{1j}} (e^{\theta_2})^{a_{2j}} \cdots (e^{\theta_d})^{a_{dj}} = q_1^{a_{1j}} q_2^{a_{2j}} \cdots q_d^{a_{dj}} \tag{3.13}$$

where we transform parameters as $q_i := \exp \theta_i$. The map ϕ_A is characterized by the associating design matrix A. Later, we show that the image of this mapping $\phi_A(\mathbb{R}^d)$ forms an algebraic variety, representing a lower-dimensional structure within the probability simplex.

Now, we introduce a fundamental theorem of algebraic statistics [41] that provides a tool to analyze a Toric model algebraically after we provide some preparatory definitions.

Definition 3.2.1 (Monomial)**.** A *monomial* in $x_1, \ldots, x_n$ is a product of the form

$$x_1^{a_1} \cdot x_2^{a_2} \cdots x_n^{a_n}, \tag{3.14}$$

where all of the exponents $a_1, \ldots, a_n$ are non-negative integers. We denote

$$\boldsymbol{x}^{\boldsymbol{a}} := x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}, \tag{3.15}$$

for two vectors $\boldsymbol{x} = [x1, \ldots, x_n] \in \mathbb{R}^n$ and $\boldsymbol{a} = [a_1, \ldots, a_n] \in \mathbb{N}^n$, sometimes referring to it as a *monomial power* of $\boldsymbol{x}$ to $\boldsymbol{a}$.

Definition 3.2.2 (Polynomial and Polynomial ring)**.** A *polynomial* f in $x_1, \ldots, x_n$ with coefficients in a field $\mathbb{K}$ is a finite linear combination with coefficients $k \in \mathbb{K}$ of monomials. We write a polynomial f in the form

$$f = \sum_i k_i \boldsymbol{x}^{\boldsymbol{a}_i}, \quad k_i \in \mathbb{K} \tag{3.16}$$

where $\boldsymbol{x}^{\boldsymbol{a}_i}$ are monomials. The set of all polynomials in $x_1, \ldots, x_n$ with coefficients in $\mathbb{K}$ is denoted $\mathbb{K}[x_1, \ldots, x_n]$.

It can be easily shown that $\mathbb{K}[x_1, \ldots, x_n]$ is a *commutative ring*, and is called *polynomial ring*.

Definition 3.2.3 (Ideal)**.** Let $\mathbb{K}[x_1, \ldots, x_m]$ be the commutative ring of all polynomials in indeterminates $x_1, \ldots, x_m$ with coefficients in $\mathbb{K}$. A subset $I \subseteq \mathbb{R}[x_1, \ldots, x_m]$ is an *ideal* if it satisfies:

(i) $0 \in I$, where 0 is a constant zero function.

(ii) If $f, g \in I$, then $f + g \in I$.

(iii) If $f \in I$ and $h \in \mathbb{K}[x_1, \ldots, x_m]$, then $hf \in I$.

Definition 3.2.4 (Ideal generation)**.** If $f_1, \ldots, f_s \in \mathbb{K}[x_1, \ldots, x_m]$, then we define

$$\langle f_1, \ldots, f_s \rangle := \left\{ \sum_{i=1}^{s} h_i f_i \mid h_1, \ldots, h_s \in \mathbb{K}[x_1, \ldots, x_m] \right\} \tag{3.17}$$

that we call *ideal generated by* a set of polynomials $\{f_1, \ldots, f_s\}$. An ideal is a *binomial ideal* if it has a generating set consisting of binomials.

It can be easily proved that $\langle f_1, \ldots, f_s \rangle$ is indeed an ideal. We now show the Sturmfeld's fundamental theorem of algebraic statistics[41].

Theorem 3.2.1 (Fundamental theorem)**.** Let $A \in \{0, 1\}^{d \times N}$ be a binary design matrix associated with a Toric model given by $\boldsymbol{p}^\top = \frac{1}{Z} \exp\langle \boldsymbol{\theta}^\top A \rangle$ as defined in 3.2.2. Then the toric

ideal I_A is a binomial ideal defined by

$$I_A := \langle \boldsymbol{p}^{\boldsymbol{u}} - \boldsymbol{p}^{\boldsymbol{v}} \mid \boldsymbol{u}, \boldsymbol{v} \in \mathbb{N}^N \text{ and } A\boldsymbol{u} = A\boldsymbol{v} \rangle. \tag{3.18}$$

Let $\mathrm{rowspan}(A)$ denote the linear span of the row vectors of A over $\mathbb{R}$. If the all-ones vector $\mathbf{1}_N = [1, 1, \ldots, 1]^\top \in \mathbb{R}^N$ belongs to $\mathrm{rowspan}(A)$, then I_A is *homogeneous*; that is, for every binomial $\boldsymbol{p}^{\boldsymbol{u}} - \boldsymbol{p}^{\boldsymbol{v}}$ in the generating set, the terms have the same total degree (i.e., $\sum u_i = \sum v_i$) Note that a polynomial with two terms is called a *binomial*, and the equation $\boldsymbol{p}^{\boldsymbol{u}} - \boldsymbol{p}^{\boldsymbol{v}} = 0$ is called a *binomial constraint*.

To prove the theorem, we need the following lemma.

Lemma 3.2.2. Let $\ker A = \{\boldsymbol{u} \in \mathbb{R}^N \mid A\boldsymbol{u} = \mathbf{0}\}$ denote the kernel of the design matrix A. If a probability vector $\boldsymbol{p}$ is generated by the monomial mapping associated with A , then for any vector $\boldsymbol{u} \in \ker A$, the following holds:

$$\boldsymbol{p}^{\boldsymbol{u}} = 1. \tag{3.19}$$

Proof. Note that $\boldsymbol{u} \in \ker A$ implies that $A\boldsymbol{u} = \mathbf{0} \in \mathbb{R}^d$ by definition. The monomial power of probability vector $\boldsymbol{p}$ to $\boldsymbol{u}$ is

$$\begin{aligned}
\boldsymbol{p}^{\boldsymbol{u}} &= p_1^{u_1} p_2^{u_2} \cdots p_N^{u_N} && (3.20)\\
&= (e^{\theta_1 a_{11}} e^{\theta_2 a_{21}} \cdots e^{\theta_d a_{d1}})^{u_1} \cdots (e^{\theta_1 a_{1N}} e^{\theta_2 a_{2N}} \cdots e^{\theta_d a_{dN}})^{u_N} && (3.21)\\
&= \exp\left(\theta_1 a_{11} u_1 + \theta_2 a_{21} u_1 + \cdots + \theta_i a_{ij} u_j + \cdots + \theta_d a_{dN} u_N\right) && (3.22)\\
&= \exp\left(\boldsymbol{\theta}^\top A \boldsymbol{u}\right) && (3.23)\\
&= \exp(0) && (3.24)\\
&= 1, && (3.25)
\end{aligned}$$

which proves the lemma. Under the theorem notation, $A\boldsymbol{u} = A\boldsymbol{v}$ implies $\boldsymbol{u} - \boldsymbol{v} \in \ker A$. □

Proof of Theorem 3.2.1. Using Lemma 3.2.2, we can conclude the first part of the theorem. Since $A\boldsymbol{u} = A\boldsymbol{v}$ implies $\boldsymbol{u} - \boldsymbol{v} \in \ker A$, we have

$$\boldsymbol{p}^{\boldsymbol{u}-\boldsymbol{v}} = \frac{\boldsymbol{p}^{\boldsymbol{u}}}{\boldsymbol{p}^{\boldsymbol{v}}} = 1, \quad \text{thus} \quad \boldsymbol{p}^{\boldsymbol{u}} - \boldsymbol{p}^{\boldsymbol{v}} = 0. \tag{3.26}$$

Note that $\boldsymbol{p}^{\boldsymbol{u}}$ is a monomial in the polynomial ring $\mathbb{R}[p_1, \ldots, p_N]$ and $\boldsymbol{p}^{\boldsymbol{u}} - \boldsymbol{p}^{\boldsymbol{v}}$ is a binomial.

To see that I_A is homogeneous, the condition $\mathbf{1} \in \mathrm{rowspan}(A)$ ensures that $A\boldsymbol{u} = A\boldsymbol{v}$ implies $\mathbf{1}^\top \boldsymbol{u} = \mathbf{1}^\top \boldsymbol{v}$ (i.e., $\sum u_i = \sum v_i$). Thus all of the generating binomials $\boldsymbol{p}^{\boldsymbol{u}} - \boldsymbol{p}^{\boldsymbol{v}} \in I_A$ are homogeneous. □

Example 3.2.3 (Independent distribution with two variables)**.** Let us illustrate how a toric model works with the simplest case. Let X_1, X_2 be two binary variables whose values are $0, 1$. Suppose that these variables are independent and their joint probabilities are given

by a product of their marginal probability $Pr(X_1 = i, X_2 = j) = Pr(X_1 = i)Pr(X_2 = j)$ for $i, j = 0, 1$, where parameters are given as $Pr(X_1 = 0) = q_1, Pr(X_1 = 1) = q_2, Pr(X_2 = 0) = q_3, Pr(X_2 = 1) = q_4$. To model this probability distribution by a toric model, the associated design matrix is

$$A = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} = \begin{bmatrix} E_2 \otimes \mathbf{1}_2^\top \\ \mathbf{1}_2^\top \otimes E_2 \end{bmatrix}, \tag{3.27}$$

where E_2 is an identity matrix of order 2 and $\mathbf{1}_2$ is 2-dimensional all-one vector and $\otimes$ is the Kronecker product.

By converting parameters by $\theta_i = \log q_i$ and setting as $\theta = [\theta_1, \ldots, \theta_4]$, the probability vector is given by:

$$\boldsymbol{p}^\top = \frac{1}{Z} \exp\{\theta^\top A\} = \frac{1}{Z} \left[e^{\theta_1+\theta_3},\ e^{\theta_1+\theta_4},\ e^{\theta_2+\theta_3},\ e^{\theta_2+\theta_4} \right] \tag{3.28}$$

with the normalizing constant factors as $Z = (e^{\theta_1} + e^{\theta_2})(e^{\theta_3} + e^{\theta_4})$.

Since a vector $\boldsymbol{u} = \begin{bmatrix} 1 & -1 & -1 & 1 \end{bmatrix} \in \ker(A)$, we can set

$$\boldsymbol{u}^+ = \begin{bmatrix} 1 & 0 & 0 & 1 \end{bmatrix} \tag{3.29}$$

$$\boldsymbol{u}^- = \begin{bmatrix} 0 & 1 & 1 & 0 \end{bmatrix} \tag{3.30}$$

so that $\boldsymbol{u} = \boldsymbol{u}^+ - \boldsymbol{u}^-$. By the fundamental theorem (Theorem 3.2.1) ,

$$\boldsymbol{p}^{\boldsymbol{u}^+} - \boldsymbol{p}^{\boldsymbol{u}^-} = p_{00}p_{11} - p_{01}p_{10} = 0 \tag{3.31}$$

where $p_{ij} = Pr(X_1 = i, X_2 = j)$.

Algebraic variety The binomial constraint derived from the 2×2 independence model, as shown in Equation 3.31 is not a mere algebraic object, but it defines a geometric structure that constrains the model's probability distributions. The set of all probability vectors $\boldsymbol{p} = (p_{00}, p_{01}, p_{10}, p_{11})$ that satisfy this equation, $p_{00}p_{11} - p_{01}p_{10} = 0$, forms a specific hypersurface within the probability simplex. Crucially, this single equation is necessary and sufficient to define the model structure, as the kernel of the associated design matrix is one-dimensional and spanned uniquely by the basis vector $\boldsymbol{u} = [1, -1, -1, 1]^\top$. This geometric object is an instance of an *affine algebraic variety*. More formally, we define it as follows:

Definition 3.2.5 (Affine algebraic variety)**.** Let $f_1, \ldots, f_s \in \mathbb{R}[x_1, \ldots, x_n]$ be polynomials defined in Definition 3.2.2. Here, we identify each polynomial f_i with the induced function $f_i : \mathbb{R}^n \to \mathbb{R}$ via evaluation. Then the set

$$V(f_1, \ldots, f_s) := \{(a_1, \ldots, a_n) \in \mathbb{R}^n \mid f_i(a_1, \ldots, a_n) = 0, \quad \forall i = 1, \ldots, s\} \tag{3.32}$$

is the *affine algebraic variety* defined by $f_1, \ldots, f_s$. In this thesis, we sometimes simply refer to it as an *algebraic variety* or a *variety*.

Definition 3.2.6 (Ideal of affine variety)**.** Let $V \subseteq \mathbb{R}^n$ be an arbitrary subset that is an affine variety. Then we set

$$\mathbf{I}(V) := \{f \in \mathbb{R}[x_1, \ldots, x_n] \mid f(a_1, \ldots, a_n) = 0 \ \forall (a_1, \ldots, a_n) \in V\}. \tag{3.33}$$

We call this set $\mathbf{I}(V)$ the *ideal of* V.

Remark 3.2.1. Note that there is a slight abuse of notation here which is standard in algebraic geometry. While $V(f_1, \ldots, f_s)$ denotes the zero set defined by specific polynomials, the symbol V is used to denote the resulting geometric set (the variety) itself. Unless specified otherwise, V represents a subset of $\mathbb{R}^n$ that satisfies $V = V(f_1, \ldots, f_s)$ for some defining polynomials.

Lemma 3.2.4. If a subset $V \subseteq \mathbb{R}^n$ is an affine variety, then $\mathbf{I}(V) \subseteq \mathbb{R}[x_1, \ldots, x_n]$ is indeed an ideal.

Proof. Suppose that $f, g \in \mathbf{I}(V)$ and $h \in \mathbb{R}[x_1, \ldots, sx_n]$. Pick any point $\boldsymbol{a} = (a_1, \ldots, a_n) \in V$, then

$$f(\boldsymbol{a}) + g(\boldsymbol{a}) = 0 + 0 = 0 \tag{3.34}$$

$$h(\boldsymbol{a}) f(\boldsymbol{a}) = h(\boldsymbol{a}) \cdot 0 = 0 \tag{3.35}$$

$$\tag{3.36}$$

Thus $\mathbf{I}(V)$ is an ideal. $0 \in \mathbf{I}(V)$ is the zero polynomial vanishes in $\mathbb{R}^n$, in particular on V. □

Lemma 3.2.5. Let $f_1, \ldots, f_s \in \mathbb{R}[x_1, \ldots, x_n]$. Let $\langle f_1, \ldots, f_s \rangle$ denote the ideal generated by these polynomials. Consider the affine variety defined by them, denoted as $W = \mathbf{V}(f_1, \ldots, f_s) \subseteq \mathbb{R}^n$. Then, the following inclusion holds:

$$\langle f_1, \ldots, f_s \rangle \subseteq \mathbf{I}(W). \tag{3.37}$$

This lemma reveals a fundamental duality at the heart of algebraic geometry: the correspondence between algebraic objects (∅ideals) and geometric objects (*varieties*). While a set of polynomials $\{f_i\}$ defines a variety $V(f_1, \ldots, f_s)$ as their common zero set, a variety in turn defines an ideal $I(V)$ consisting of all polynomials that vanish on it.

The profound implication for our language model is that the probability vector $\boldsymbol{p}$, which resides in a potentially vast, high-dimensional space of joint states, is not free to exist anywhere within the probability simplex. Instead, under the assumption of independence among certain variables, $\boldsymbol{p}$ is constrained to lie on a much lower-dimensional geometric object defined by the parameter-invariant binomials in the toric ideal I_A. The specific

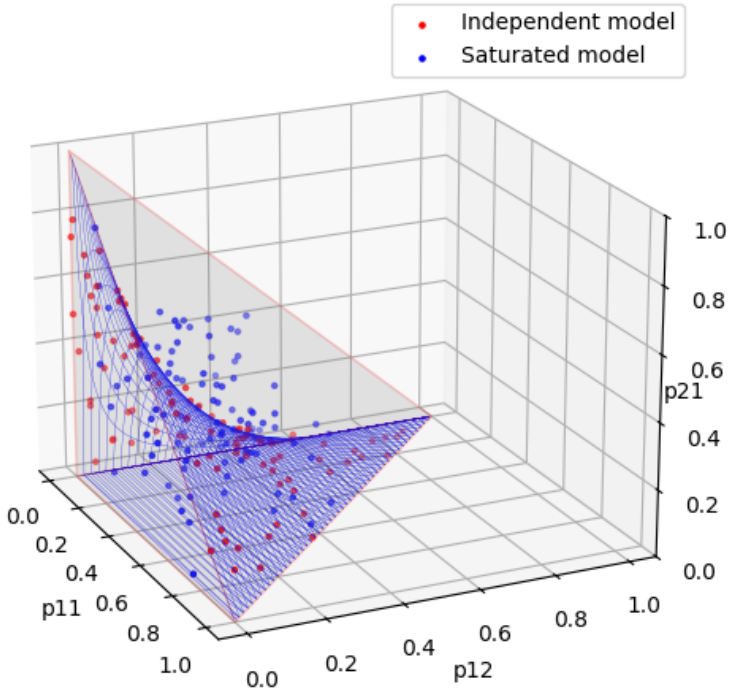


Figure 3.1: Visualization of the Segre variety for the 2×2 independence model. The red dots represent probability vectors sampled from the independence model, which lie strictly on the hypersurface (Segre variety) within the 3-simplex. In contrast, the blue dots represent unconstrained random probability vectors, which generally lie off the hypersurface.

variety corresponding to the statistical model of independence is known as the **Segre variety**. It is the image of the monomial map that takes the Cartesian product of individual probability simplices (one for each random variable) into the simplex of their joint distribution. A visualization of this geometric constraint for the 2×2 case is shown in Figure 3.1. Furthermore, this geometric framework naturally extends to more complex models. For instance, a mixture of independence models corresponds geometrically to the *secant variety* of the underlying Segre varieties, a topic that opens avenues for modeling more intricate statistical dependencies. This provides a precise, algebraic formulation of the well-known manifold hypothesis in machine learning, which posits that high-dimensional real-world data concentrates near a low-dimensional manifold. Our toric model explicitly identifies this manifold as an algebraic variety.

3.2.4 Linear Algebraic Analysis of Model Structure

A primary advantage of formulating our probabilistic framework as a toric model is the ability to analyze its structure using the tools of linear algebra. The parameter-invariant constraints that define the geometry of the model are not arbitrary; as established by the fundamental theorem, the generating binomials of the toric ideal I_A are entirely determined by the *kernel* (or null space) of the design matrix A. In other words, the algebraic structure

of the probability model is a direct reflection of the linear-algebraic structure of the row space of its design matrix.

We can illustrate this principle with the 2×2 independence model. The design matrix is given by

$$A = \begin{bmatrix} E_2 \otimes \mathbf{1}_2^\top \\ \mathbf{1}_2^\top \otimes E_2 \end{bmatrix}, \tag{3.38}$$

and the basis for its kernel is spanned by the vector $\boldsymbol{u} = [1, -1, -1, 1]^\top$. This vector directly yields the familiar binomial constraint $p_{00}p_{11} - p_{01}p_{10} = 0$.

The crucial insight comes from observing the Kronecker product structure of this kernel vector itself. Letting $\tilde{\mathbf{1}}_2 = [1, -1]^\top$, we can express the kernel basis as $\boldsymbol{u} = \tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2$. The reason this vector lies in the kernel of A becomes evident when we consider the orthogonality condition $\mathbf{1}_2^\top \tilde{\mathbf{1}}_2 = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = 0$. Now,we remind the mixed-product property of the Kronecker product as a lemma.

Lemma 3.2.6. Let $A \in \mathbb{R}^{n \times m}, B \in \mathbb{R}^{p \times q}, C \in \mathbb{R}^{m \times l}, D \in \mathbb{R}^{q,r}$ be matrices of different sizes, and $\otimes$ be a Kronecker product, then

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \tag{3.39}$$

Applying Lemma 3.2.6 to the upper and lower blocks of the design matrix A, we see that:

$$(E_2 \otimes \mathbf{1}_2^\top)(\tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2) = (E_2\tilde{\mathbf{1}}_2) \otimes (\mathbf{1}_2^\top \tilde{\mathbf{1}}_2) = \tilde{\mathbf{1}}_2 \otimes 0 = \mathbf{0}_2 \tag{3.40}$$

$$(\mathbf{1}_2^\top \otimes E_2)(\tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2) = (\mathbf{1}_2^\top \tilde{\mathbf{1}}_2) \otimes (E_2\tilde{\mathbf{1}}_2) = 0 \otimes \tilde{\mathbf{1}}_2 = \mathbf{0}_2 \tag{3.41}$$

This calculation provides a constructive explanation for the model's invariant. We can interpret the vector $\mathbf{1}_2$ as a "copy" operator and $\tilde{\mathbf{1}}_2$ as an "anti-copy" operator. The algebraic constraint arises from a purely structural *copy–anti-copy cancellation*: the copy factor $\mathbf{1}_2^\top$ present in the design matrix is precisely nullified by the anti-copy factor $\tilde{\mathbf{1}}_2$ in the kernel vector. This cancellation is independent of any parameter values $\boldsymbol{\theta}$, thus inducing an invariant—the vanishing 2-minor—for any probability distribution generated by the model.

This copy–anti-copy mechanism offers a deeper reason for the equivalence between the statistical condition of independence and the algebraic condition of the probability matrix having rank one. The structural symmetry of the design matrix, captured by this copy/ anti-copy principle, universally enforces the algebraic constraint on the probability space.

This principle is not limited to the simple 2×2 case. As we will demonstrate, this copy–anti-copy mechanism forms a fundamental building block for constructing and analyzing more complex probabilistic models. The same mechanism can be generalized to induce higher-order binomial invariants. For instance, *the No-Three-Way-Interaction model*

for $2 \times 2 \times 2$ case, which we will discuss later, generates fourth-order binomial invariants through a similar, albeit more complex, kernel structure.

Summary and Key Contributions The central message of this section is that the toric model, which constitutes a specific class of discrete exponential families (log-linear models), provides a principled way to formulate and analyze the parameter-invariant structure of a probability distribution generated by its associated design matrix. In the context of language modeling, this framework allows us to define and detect intrinsic regularities in word distributions that are invariant across corpora, even when specific parameters vary.

Geometrically, these regularities manifest as an algebraic variety—the solution set of polynomial equations constraining the probabilities—which forms a low-dimensional structure embedded in a high-dimensional probability space. This allows for the powerful idea of characterizing the intrinsic structure of language as geometry. Crucially, the combinatorial structure of the design matrix not only defines this algebraic variety but also relates directly to the vector geometry of word meanings. Specifically, it provides a theoretical basis for the analogical parallelograms observed in word embedding spaces, as discussed in [136].

Furthermore, our observation that the copy-anticopy mechanism provides a constructive explanation for this regularity leads to the notion of a minimal structural unit. We will later formally define this as *Minimum Invariant Constraints (MIC)*, an atomic notion of Context-Sensitive Independence [19], which serves as a fundamental building block for any regularity that is invariant to model parameters. These MICs will therefore be key structural features to be sought in linguistic data.

Now that we have established this powerful algebraic machinery, we will demonstrate its utility in the subsequent sections. We will show how this framework can not only reproduce classical independence models typically described by Probabilistic Graphical Models (PGMs) but also capture more complex statistical structures that lie beyond the expressive power of standard graphical representations.

3.3 Independence and Invariance

To situate our algebraic approach within the broader context of statistical modeling, we now connect it to the well-established framework of Probabilistic Graphical Models (PGMs)[74]. Our central thesis in this section is that when viewed through the lens of a toric model, the various independence structures typically represented by graphs—such as marginal, joint, and conditional independence—all manifest in a unified algebraic form: as a set of binomial constraints defined by the kernel of a specific design matrix. In the following, we will substantiate this claim by systematically cataloging these classical models and demonstrating how each corresponds to a particular toric ideal.

3.3.1 Prior studies: Probabilistic Graphical Models (PGM)

A probabilistic graphical model is a statistical model in which a graph is used to represent the dependency structure between random variables [87]. The nodes of the graph correspond to the random variables of interest, while the edges (or lack thereof) encode a set of conditional independence (CI) assumptions.

The concept of conditional independence itself has deep roots, formally defined in terms of σ-algebras by M. Loève [85] and later systematically studied in a statistical context by A. P. Dawid [39]. However, it was J. Pearl [110] who recognized the profound significance of CI for probabilistic reasoning and artificial intelligence, establishing a formal connection between graph topology and statistical independence through axioms known as the *semi-graphoid* properties. In particular, Bayesian Networks, which use directed acyclic graphs, provided a powerful framework for reading CI relations directly from the graph via the *d-separation* criterion [74], laying the groundwork for modern causal analysis.

Despite the framework's power, the limitations of a purely topological representation became apparent. The search for a more comprehensive theory, ensuring *soundness* (all independencies implied by the graph hold in the distribution) and *completeness* (all independencies in the distribution are represented by the graph), has been an active area of research. This quest led to the development of more expressive, non-graphical formalisms. Studený [132], for instance, introduced a linear-algebraic representation using *imsets* to characterize CI structures that defy graphical representation. Concurrently, the field of algebraic statistics demonstrated how tools from algebraic geometry could provide an exact description for these complex statistical models. Building on this latter tradition, we will now proceed to catalogue several key CI models within the framework of algebraic statistics, showing how each is defined by a specific toric ideal.

3.3.2 The Algebraic Unification of Independence Models

We will show that a wide spectrum of statistical models—from classical independence to context-specific and higher-order interactions—can be systematically constructed through the single, unified principle of specifying a design matrix A and analyzing its kernel.

First, we establish our notation. Let $X = \{X_1, X_2, \dots, X_m\}$ be a set of m discrete random variables, where each variable X_k has a finite state space $\mathcal{X}_k \in \mathbb{Z}$ of size $n_k := |\mathcal{X}_k|$. The joint state space is the Cartesian product of these sets $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_m$, with a total of $N = \prod_{k=1}^{m} n_k$ states. A probability distribution over this space is represented by a vector $\boldsymbol{p} \in \Delta^{N-1}$ where $\Delta^{N-1} := \left\{ \boldsymbol{p} \in \mathbb{R}^N \;\middle|\; \boldsymbol{p} \geq \mathbf{0} \text{ and } \mathbf{1}^\top \boldsymbol{p} = 1 \right\}$, whose entries correspond to the joint probabilities $P(x_1, \dots, x_m)$ and are ordered lexicographically. For instance, with three binary variables ($m = 3, n_k = 2$ for every k), the vector $\boldsymbol{p}$ has entries $(p_{000}, p_{001}, p_{010}, p_{011}, p_{100}, p_{101}, p_{110}, p_{111})^\top$. We treat these lexicographically ordered tuples as coordinates in an N-dimensional vector space.

To bridge our algebraic results with established statistical theory, we first formally define the conditional independence statement and state its equivalent characterization in

terms of vanishing binomials.

Definition 3.3.1 (Conditional independence [133])**.** Following the notation in [133], let $X_1, \ldots, X_m$ be random variables and $A, B, C \subseteq [m] := \{1, \ldots, m\}$ be pairwise disjoint subsets. We denote by X_A the random vector $(X_k)_{k \in A}$ indexed by the subset A. The random vector X_A is *conditionally independent* of X_B given X_C if and only if

$$p_{A \cup B|C}(x_A, x_B \mid x_C) = p_{A|C}(x_A \mid x_C) \cdot p_{B|C}(x_B \mid x_C) \tag{3.42}$$

for all $x_A \in \prod_{k \in A} \mathcal{X}_k$ and similarly for x_B, x_C. The notation $X_A \perp\!\!\!\perp X_B \mid X_C$ is used to denote this independence.

Theorem 3.3.1. If X is a discrete random vector, then the conditional independent statement $X_A \perp\!\!\!\perp X_B \mid X_C$ holds if and only if

$$p_{i_A, i_B, i_C, +} p_{j_A, j_B, i_C, +} - p_{i_A, j_B, i_C, +} p_{j_A, i_B, i_C, +} = 0 \tag{3.43}$$

for all $i_A, j_A \in \prod_{k \in A} \mathcal{X}_k$, $i_B, j_B \in \prod_{k \in B} \mathcal{X}_k$ and all $i_C \in \prod_{k \in C} \mathcal{X}_k$. The notation $p_{i_A, i_B, i_C, +}$ denotes the joint probability that is marginalized over the remaining variables $D = [m] \setminus (A \cup B \cup C)$, defined as:

$$p_{i_A, i_B, i_C, +} := \sum_{k \in \prod_{l \in D} \mathcal{X}_l} P(X_A = i_A, X_B = i_B, X_C = i_C, X_D = k). \tag{3.44}$$

Proof. By marginalizing, we can assume that $A \cup B \cup C = [m]$. By conditioning, we may assume that C is empty. By aggregating the states indexed by $i \in \prod_{k \in A} \mathcal{X}_k$, $j \in \prod_{k \in B} \mathcal{X}_k$, respectively, we can treat this as a conditional independence statement of two aggregated variables $X_1 \perp\!\!\!\perp X_2$. In this setting, the conditional independence constraints in (3.43) are equivalent to saying that all 2×2 minors of the following matrix,

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n_2} \\ p_{21} & p_{22} & \cdots & p_{2n_2} \\ \vdots & \vdots & \cdots \vdots & \\ p_{n_1 1} & p_{n_1 2} & \cdots & p_{n_1 n_2} \end{bmatrix} \tag{3.45}$$

are zero, where $n_1 = |\prod_{k \in A} \mathcal{X}_k|$, $n_2 = |\prod_{k \in B} \mathcal{X}_k|$. Since P is not the zero matrix, this implies that P has rank one. Thus there exist two vectors $\boldsymbol{p} \in \mathbb{R}^{n_1}$ and $\boldsymbol{q} \in \mathbb{R}^{n_2}$ such that $\sum_i \boldsymbol{p}(i) = 1$ and $\sum_j \boldsymbol{q}(j) = 1$ where $P = \boldsymbol{p}\boldsymbol{q}^T$,implying that $p_{ij} = \boldsymbol{p}(i)\boldsymbol{q}(j)$ for $i = 1, \ldots, n_1; j = 1, \ldots, n_2$ and hence $X_1 \perp\!\!\!\perp X_2$. Conversely, if $X_1 \perp\!\!\!\perp X_2$, P must be a rank one matrix.

□

The above proof follows [133].

We will now examine three canonical design matrices for the case of three binary variables. For each matrix, we will first derive its associated toric ideal and then, by invoking Theorem 3.3.1, identify the classical independence model it generates.

Model 1: The Fully Symmetric Design Matrix Let us first examine the model generated by a design matrix where each row block corresponds to the parameters for a single marginal distribution:

$$A_1 = \begin{bmatrix} E_2 \otimes \mathbf{1}_2^\top \otimes \mathbf{1}_2^\top \\ \mathbf{1}_2^\top \otimes E_2 \otimes \mathbf{1}_2^\top \\ \mathbf{1}_2^\top \otimes \mathbf{1}_2^\top \otimes E_2 \end{bmatrix}. \tag{3.46}$$

The kernel of this matrix contains vectors encoding pairwise constraints. For instance, vectors of the form $\boldsymbol{u}_i = \boldsymbol{e}_i \otimes \tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2$ lie in $\ker(A_1)$ where $\boldsymbol{e}_i$ is a standard basis vectors for $i = 1, 2$. Since the two anti-copy factors $\tilde{\mathbf{1}}_2$ can be placed in any two of the three positions in the Kronecker product, there are the number of combinations $\binom{3}{2} = 3$ such types of kernel vectors.

- $\boldsymbol{u}_i = \boldsymbol{e}_i \otimes \tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2$ for $i = 1, 2$
- $\boldsymbol{v}_i = \tilde{\mathbf{1}}_2 \otimes \boldsymbol{e}_i \otimes \tilde{\mathbf{1}}_2$ for $i = 1, 2$
- $\boldsymbol{w}_i = \tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2 \otimes \boldsymbol{e}_i$ for $i = 1, 2$

Each type generates a set of binomial equations, such as:

$$p_{i_1,j_1,k} p_{i_2,j_2,k} - p_{i_1,j_2,k} p_{i_2,j_1,k} = 0. \tag{3.47}$$

where the indices correspond to one of the following permutations:

- $i_{1,2} \in \mathcal{X}_1, j_{1,2} \in \mathcal{X}_2, k \in \mathcal{X}_3$ (implies $X_1 \perp\!\!\!\perp X_2 \mid X_3$),
- $i_{1,2} \in \mathcal{X}_1, j_{1,2} \in \mathcal{X}_3, k \in \mathcal{X}_2$ (implies $X_1 \perp\!\!\!\perp X_3 \mid X_2$),
- $i_{1,2} \in \mathcal{X}_2, j_{1,2} \in \mathcal{X}_3, k \in \mathcal{X}_1$ (implies $X_2 \perp\!\!\!\perp X_3 \mid X_1$).

This motivates introducing the notations I, J, K as disjoint subsets of variables, where we permute the assignment of X_1, X_2, X_3 to each of I, J, K to cover all pairwise conditional independence statements. By Theorem 3.3.1, these binomials are precisely the constraints for pairwise independence between any two variables, namely, $X_2 \perp\!\!\!\perp X_3 \mid X_1, X_1 \perp\!\!\!\perp X_3 \mid X_2$ and $X_1 \perp\!\!\!\perp X_2 \mid X_3$. As derived above, the binomials generated by $\ker(A_1)$ ensure that the conditional independence statements $X_A \perp\!\!\!\perp X_B \mid X_C$ hold for all distinct permutations of indices $\{A, B, C\} = \{1, 2, 3\}$. The collection of all such constraints generated by $\ker(A_1)$ requires these three conditional independence statements to hold simultaneously. This simultaneous satisfaction defines the **complete independence model ($X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3$)**, whose joint probability can be shown to factorizes into the product of its marginals.

Model 2: The Partitioned Design Matrix The second algebraic structure we investigate is a design matrix that partitions the variables into two groups, $\{X_1, X_2\}$ and $\{X_3\}$. This is encoded by assigning one row block for the joint parameters of (X_1, X_2) and another for

the marginal parameters of X_3:

$$A_2 = \begin{bmatrix} E_2 \otimes E_2 \otimes \mathbf{1}_2^\top \\ \mathbf{1}_2^\top \otimes \mathbf{1}_2^\top \otimes E_2 \end{bmatrix}. \tag{3.48}$$

The kernel of this matrix generates binomial constraints that enforce independence between these two partitioned sets of variables. The resulting toric ideal is generated by binomials of the form:

$$p_{i_1 j_1 k} p_{i_2 j_2 k} - p_{i_1 j_2 k} p_{i_2 j_1 k} = 0. \tag{3.49}$$

for $I = \{X_1, X_2\}, J = \{X_3\}, K = \varnothing$. Applying Theorem 3.3.1 with variable sets $A = \{1, 2\}$, $B = \{3\}$, and $C = \emptyset$, we see that these constraints are the exact conditions for the joint variable (X_1, X_2) to be independent of X_3. We therefore conclude that the design matrix A_2 generates the **joint independence model (**$(X_1, X_2) \perp\!\!\!\perp X_3$**)**, which corresponds to the factorization $P(X_1, X_2, X_3) = P(X_1, X_2)P(X_3)$.

Model 3: The Shared Variable Design Matrix The third model is generated by a design matrix whose structure reflects a shared variable, X_1, that mediates the interaction between X_2 and X_3. This is encoded by row blocks corresponding to the joint variables (X_1, X_2) and (X_1, X_3), where the two standard basis vectors $\boldsymbol{e}_1, \boldsymbol{e}_2 \in \{0, 1\}^2$ appearing in the first position of both blocks provide conditioning:

$$A_3 = \begin{bmatrix} \boldsymbol{e}_1^\top \otimes E_2 \otimes \mathbf{1}_2^\top \\ \boldsymbol{e}_1^\top \otimes \mathbf{1}_2^\top \otimes E_2 \\ \boldsymbol{e}_2^\top \otimes E_2 \otimes \mathbf{1}_2^\top \\ \boldsymbol{e}_2^\top \otimes \mathbf{1}_2^\top \otimes E_2 \end{bmatrix}. \tag{3.50}$$

The kernel vectors for this matrix take the form $\boldsymbol{u}_i = \boldsymbol{e}_i \otimes \tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2$ with $i = 1, 2$. This structure is key: $\boldsymbol{e}_i$ selects a specific state of the shared variable X_1, and for that fixed state, the anti-copy factors $\tilde{\mathbf{1}}_2$ enforce an independence constraint between X_2 and X_3. This process directly generates binomials of the form:

$$p_{i_1 j_1 k} p_{i_2 j_2 k} - p_{i_1 j_2 k} p_{i_2 j_1 k} = 0. \tag{3.51}$$

for $I = \{X_2\}, J = \{X_3\}, K = \{X_1\}$ According to Theorem 3.3.1 with $A = \{2\}$, $B = \{3\}$, and $C = \{1\}$, these are precisely the constraints for the conditional independence of X_2 and X_3 given X_1. Thus, the design matrix A_3 generates the **conditional independence model (**$X_2 \perp\!\!\!\perp X_3 \mid X_1$**)**, commonly represented by the graph $X_2 \leftarrow X_1 \rightarrow X_3$.

Model 4: The Hybrid Design Matrix Finally, the flexibility of our framework is demonstrated by its ability to capture granular, context-specific constraints. Consider the following

'hybrid' design matrix:

$$A_4 = \begin{bmatrix} E_2 \otimes \mathbf{1}_2^\top \otimes \mathbf{1}_2^\top \\ \boldsymbol{e}_1^\top \otimes E_2 \otimes \mathbf{1}_2^\top \\ \boldsymbol{e}_1^\top \otimes \mathbf{1}_2^\top \otimes E_2 \\ \boldsymbol{e}_2^\top \otimes E_2 \otimes E_2 \end{bmatrix}. \tag{3.52}$$

Here, the algebraic structure is selectively applied. The second and third row blocks impose the shared variable structure of A_3, but only for the $X_1 = 0$ context (via the projector $\boldsymbol{e}_1^\top$). Conversely, the fourth row block allows for arbitrary dependence when $X_1 = 1$ by modeling the full joint distribution $P(X_2, X_3 \mid X_1 = 1)$. The kernel of this matrix will therefore only generate constraints within the $X_1 = 0$ slice of the probability space, leading to binomials such as:

$$p_{i_1 j_1 k} p_{i_2 j_2 k} - p_{i_1 j_2 k} p_{i_2 j_1 k} = 0. \tag{3.53}$$

for $I = \{X_2\}, J = \{X_3\}, k = 0 \in \mathcal{X}_1$ These constraints satisfy the conditions of Theorem 3.3.1 only when the conditioning variable X_1 is fixed to the state 0. This shows how local regularities can be specified directly at the algebraic level, and we conclude that A_4 generates the **context-specific independence model ($X_2 \perp\!\!\!\perp X_3 \mid X_1 = 0$)**. We note that the notion of Context-Specific Independence is proposed by Boutillier[19] in the domain of PGM.

Transcending PGM: The No-Three-Way-Interaction Model The true power of the algebraic approach is revealed in its capacity to naturally describe models that are problematic for standard graphical representations. The No-Three-Way-Interaction (No3Way) model is a prime example. In PGM, it corresponds to a complete graph on three nodes, which only states that no variables are marginally or conditionally independent, failing to capture the specific higher-order constraint that defines the model.

In our framework, this model is defined by a design matrix corresponding to the set of all pairwise interactions, $\mathcal{F} = \{\{X_1, X_2\}, \{X_1, X_3\}, \{X_2, X_3\}\}$ that represents a simplicial complex:

$$A_{\text{No3Way}} = \begin{bmatrix} E_2 \otimes E_2 \otimes \mathbf{1}_2^\top \\ E_2 \otimes \mathbf{1}_2^\top \otimes E_2 \\ \mathbf{1}_2^\top \otimes E_2 \otimes E_2 \end{bmatrix}. \tag{3.54}$$

The kernel of this matrix is one-dimensional, spanned uniquely by the vector containing three anti-copy factors:

$$\boldsymbol{u} = \tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2 \otimes \tilde{\mathbf{1}}_2 = \begin{bmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix}^\top. \tag{3.55}$$

This algebraic structure directly induces a *quartic* invariant on the probability vector, a genuine higher-order interaction that cannot be expressed as a simple conditional inde-

pendence statement. The resulting constraint is:

$$p_{000}p_{011}p_{101}p_{110} - p_{001}p_{010}p_{100}p_{111} = 0. \tag{3.56}$$

This example shows that the copy-anti-copy mechanism is not limited to quadratic binomials but naturally generalizes to generate polynomial constraints of arbitrary degree, providing a unified language for a vast class of statistical models.

From Independence to Structural Invariance: The MIC

The preceding examples reveal a profound and unifying pattern. Seemingly disparate statistical concepts—marginal, joint, conditional, and even context-specific independence—along with higher-order interactions like the No3Way model, all manifest algebraically in the exact same way: as a set of polynomial equations that define an algebraic variety on which the probability distribution must lie. These defining equations are, in turn, derived directly from the kernel of a design matrix A.

This unification strongly suggests that concepts like 'independence' or 'conditional independence', while computationally and conceptually useful, are not the most fundamental principles for defining a probabilistic model. They are, rather, prominent special cases—specific and highly structured manifestations of a deeper, more foundational concept. We argue that this foundational concept is the **structural invariance** encoded algebraically within the design matrix. The copy-anti-copy mechanism is the precise algebraic machinery that generates this invariance. It is a purely structural property, independent of any specific parameter values, and thus independent of any particular corpus.

If structural invariance is the fundamental building block of all parameter-free regularities, it necessitates a definition for its most elementary, irreducible unit. The constraints we have observed are binomials arising from vectors in $\ker(A)$. The most basic form of such a constraint is an irreducible binomial, which cannot be factored into simpler polynomial constraints. This leads us to the central definition of our framework.

Definition 3.3.2 (Minimum Invariant Constraints (MIC))**.** We say that a pair of selected states from each variable set $i_1, i_2 \in \mathcal{X}_I$ and $j_1, j_2 \in \mathcal{X}_J$ are **Minimum Invariant Constraints** under context k if their 2-minor $p_{i_1 j_1 k}p_{i_2 j_2 k} - p_{i_1 j_2 k}p_{i_2 j_1 k}$ vanishes.

The *MIC* is the elementary particle of statistical structure. It is the minimal, indivisible unit of regularity that is guaranteed to hold for any probability distribution generated by the model, regardless of the specific parameters. By defining this fundamental unit, we move beyond a mere catalog of different independence models. Instead, we provide a unified "alphabet" from which all parameter-invariant structures can be constructed. Classical conditional independence is just one type of "molecule" built from these atoms; context-specific independence and higher-order interactions are others. This perspective implies that what we should seek in data are not just pre-defined notions of independence,

but the underlying atomic CSIs that compose the true structural signature of the data-generating process.

3.4 Walsh Transformation

3.4.1 Subspace of a Design Matrix

In the preceding section, we established an algebraic-geometric framework to capture the parameter-invariant structures of probability distributions. This framework is crucial for characterizing the stochastic patterns observed across different linguistic corpora. The key instruments in this approach were the design matrix, its kernel, the resulting toric ideal of binomials, and our novel concept of the **MIC** as a fundamental building block. A central finding was that the geometric constraints on the probability vector are entirely determined by the kernel of the design matrix.

This finding brings to light the fundamental concept of duality in linear algebra: the kernel and the row space of a matrix are orthogonal complements, and the structure of one uniquely determines the structure of the other. Thus, if the kernel defines the constraints on the model, the row space defines the structure of the model itself. This motivates a shift in our analytical focus. Having understood the geometry of the constraints, we now pose our next research questions:

1. How can we characterize the algebraic structure of the row space of the design matrix in a way that reveals its combinatorial properties?

2. What is the most natural basis for describing this row space and its duality with the kernel?

As demonstrated by the examples in 3.3.2, different classes of independence models are indeed determined by the specific configuration, and therefore the row space, of their design matrices. We are thus motivated to characterize these row spaces directly. A simple measure like the rank of the matrix is insufficient for this purpose, as it provides only the dimension of the subspace and reveals nothing about its internal combinatorial structure—such as how it might be decomposed.

To address this, we leverage the fact that our design matrices are binary. This property makes them particularly amenable to analysis via the **Walsh-Hadamard Transformation**, a form of discrete Fourier analysis that uses square waves (the Walsh functions) instead of sine and cosine waves. From a broader mathematical viewpoint, Fourier transforms are a tool of harmonic analysis on groups. The standard Fourier transform applies to continuous Lie groups like the circle group (S^1), while the Walsh-Hadamard transform is the analogous tool for finite abelian groups, most notably the dyadic group $\mathbb{Z}_2^m$. Harmonic analysis has deep connections to representation theory and is increasingly used in machine learning to discover hidden symmetries in data. Crucially, the Maschke's theorem guarantees that such transformations provide a decomposition of the space into *irreducible components*[45],

which is precisely what we need to uncover the fundamental building blocks of our models. This formulation provides a powerful and practical means to analyze and operate on the algebraic structure underlying the configuration matrix.

3.4.2 The Walsh-Hadamard Transformation

The Walsh-Hadamard transformation provides a change of basis, transforming a vector in the "sequence domain" (our lexicographically ordered state space) into the "frequency domain", a coordinate representation over a set of orthogonal, symmetric, and involutory basis vectors. We will introduce this basis in two complementary ways: first, through the intuitive, recursive construction of the Walsh-Hadamard matrix, and second, through the more formal and powerful definition of Walsh functions, which will be essential for constructing the Clifford algebra in the next section.

The Walsh–Hadamard Matrix The Walsh–Hadamard matrix of order 2^m, denoted H_m, is constructed recursively:

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad H_m = \begin{bmatrix} H_{m-1} & H_{m-1} \\ H_{m-1} & -H_{m-1} \end{bmatrix} = H_1 \otimes H_{m-1}. \tag{3.57}$$

The rows of the Walsh-Hadamard matrix are mutually orthogonal. Since H_m is a $2^m \times 2^m$ matrix, its 2^m non-zero, mutually orthogonal row vectors are linearly independent and thus form an orthogonal *basis* for the vector space $\mathbb{R}^{2^m}$. For $m = 3$, the matrix is:

$$H_3 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{bmatrix} \tag{3.58}$$

As we will see, each row of this matrix can be generated by a corresponding Walsh function.

Walsh functions The Walsh basis, presented as a set of row vectors in the matrix from Eq. (3.58), is concrete, though its inherent algebraic structure is not immediately apparent. Thus, we now redefine these basis vectors in terms of *Walsh functions* to provide a rigorous mathematical foundation for the subsequent link to Clifford algebra. This reformulation is not merely a notational convenience; it is essential because it allows us to interpret the row space of a design matrix simultaneously as a geometric object (via its kernel) and as an algebraic system equipped with the multilinear operations of a Clifford algebra, a structure

that emerges naturally when the space spanned by the Walsh functions is endowed with a geometric product.

Definition 3.4.1 (Walsh functions)**.** Let $X = \{0,1\}^m = \mathbb{F}_2^m$ be the m-fold Cartesian product of the binary field. Each element of X is represented by a binary vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_m] \in X$. For any subset $S \subseteq [m] := \{1, 2, \ldots, m\}$, we define the *Walsh function*

$$h_S : X \to \{+1, -1\}, \quad h_S(\boldsymbol{x}) = (-1)^{\sum_{i \in S} x_i}. \tag{3.59}$$

For example, when $m = 3$, we have $X = \{000, 001, \ldots, 111\}$. For $S = \{1, 2\}$ and $\boldsymbol{x} = 110$, the function evaluates to $h_{\{1,2\}}(110) = (-1)^{1+1} = +1$. Evaluating this function $h_{\{1,2\}}$ over all points in X (in lexicographical order) yields the vector:

$$\begin{bmatrix} +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \end{bmatrix}, \tag{3.60}$$

which is the fifth row of the Walsh matrix in Eq. (3.58). We can identify each function h_S with its corresponding vector representation, which we denote as e_S. The full set of Walsh basis vectors for $m = 3$ is:

$$e_0 := h_\varnothing(X) = [+1, +1, +1, +1, +1, +1, +1, +1] \tag{3.61}$$
$$e_1 := h_{\{1\}}(X) = [+1, +1, +1, +1, -1, -1, -1, -1] \tag{3.62}$$
$$e_2 := h_{\{2\}}(X) = [+1, +1, -1, -1, +1, +1, -1, -1] \tag{3.63}$$
$$e_3 := h_{\{3\}}(X) = [+1, -1, +1, -1, +1, -1, +1, -1] \tag{3.64}$$
$$e_{12} := h_{\{1,2\}}(X) = [+1, +1, -1, -1, -1, -1, +1, +1] \tag{3.65}$$
$$e_{13} := h_{\{1,3\}}(X) = [+1, -1, +1, -1, -1, +1, -1, +1] \tag{3.66}$$
$$e_{23} := h_{\{2,3\}}(X) = [+1, -1, -1, +1, +1, -1, -1, +1] \tag{3.67}$$
$$e_{123} := h_{\{1,2,3\}}(X) = [+1, -1, -1, +1, -1, +1, +1, -1] \tag{3.68}$$

Therefore, the set of Walsh functions $\{h_S\}_{S \subseteq [m]}$, when evaluated over $\mathbb{F}_2^m$, yields a set of vectors $\{e_S\}_{S \subseteq [m]}$ that forms a complete orthogonal basis for the vector space of all real-valued functions on X. This function space is isomorphic to $\mathbb{R}^{2^m}$.

The Walsh functions satisfy a fundamental multiplicative relation that forms the basis of their algebraic structure.

Lemma 3.4.1. For any two subsets $S, T \subseteq [m]$, the Walsh functions satisfy

$$h_{S \Delta T}(\boldsymbol{x}) = h_S(\boldsymbol{x})\, h_T(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{F}_2^m \tag{3.69}$$

where $S \Delta T = (S \cup T) \setminus (S \cap T)$ is the symmetric difference of the sets.

Proof. The exponent of -1 in $h_S(\boldsymbol{x})\, h_T(\boldsymbol{x})$ is $\sum_{i \in S} x_i + \sum_{i \in T} x_i$. Since $x_i \in \{0, 1\}$, addition modulo 2 corresponds to the XOR operation. The sum of exponents modulo 2 is equivalent

to summing over the symmetric difference:

$$\left(\sum_{i\in S} x_i + \sum_{i\in T} x_i\right) \pmod 2 \cong \left(\sum_{i\in S\Delta T} x_i + 2\sum_{i\in S\cap T} x_i\right) \pmod 2 \tag{3.70}$$

$$\cong \sum_{i\in S\Delta T} x_i \pmod 2. \tag{3.71}$$

Thus, $h_S(\boldsymbol{x})\, h_T(\boldsymbol{x}) = (-1)^{\sum_{i\in S\Delta T} x_i} = h_{S\Delta T}(\boldsymbol{x})$. □

Definition 3.4.2 (Order of a Walsh Basis Vector)**.** The **order** of a Walsh basis vector e_S (or its corresponding Walsh function h_S) is defined as the cardinality of its indexing set $S \subseteq [m]$.

- **Zeroth-Order:** The basis vector $e_\emptyset$ (often denoted e_0) is the unique vector of order 0, as $|\emptyset| = 0$. It represents the constant or mean component in signal processing.
- **First-Order:** A basis vector e_S is of first order if $|S| = 1$. These vectors, such as $e_{\{i\}}$ (often denoted e_i), correspond to the fundamental frequencies or main effects of the individual variables.
- **Higher-Order:** More generally, a basis vector e_S is of order k if its cardinality is $|S| = k$. These vectors, such as $e_{\{i,j\}}$, represent k-way interactions among the variables.

When this vector space is endowed with a product structure inspired by this multiplicative property, as we will formalize in the next section, it becomes isomorphic to a Clifford algebra.

A Note on the Algebraic Structures To clarify the relationship between the algebraic systems at play, we provide this brief supplement. The construction proceeds through a hierarchy of structures:

1. **The Domain: Field, Group, and Vector Space Structures:** Our starting point is the set $X = \{0,1\}^m$. It is crucial to distinguish the two fundamental algebraic structures built upon the set $\{0,1\}$. The *group* $(\mathbb{Z}_2, +)$ considers only a single operation: addition modulo 2. In contrast, *field* $(\mathbb{F}_2, +, \times)$ is a richer structure, equipped with both addition and multiplication. This distinction is vital: the field $\mathbb{F}_2$ provides the necessary scalars to define X as an m-dimensional *vector space*, while the additive group structure is what we analyze with characters. Thus, we view X in two compatible ways: as an m-dimensional vector space over the field $\mathbb{F}_2$, and concurrently, as a finite abelian group under vector addition, which we denote as $(\mathbb{Z}_2)^m$.
2. **Walsh Functions as Irreducible Characters:** A Walsh function h_S is a map from the *additive group* $((\mathbb{Z}_2)^m, +)$ to the *multiplicative group* $(\{+1,-1\}, \times)$. This map preserves the group structure—i.e., $h_S(\boldsymbol{x}+\boldsymbol{y}) = h_S(\boldsymbol{x}) \cdot h_S(\boldsymbol{y})$—and is therefore a *group character*, whose mathematical definition will given soon in the next paragraph. In the language of representation theory, for an abelian group like $(\mathbb{Z}_2)^m$, the characters are precisely

its one-dimensional *irreducible representations*[45]. The term "irreducible" is critical: it signifies that these characters are the most fundamental, non-decomposable units of symmetry for the group. The fact that any function on the group can be uniquely decomposed into these irreducible components is, for finite groups, a foundational result of representation theory, known as the complete reducibility (a consequence of Maschke's theorem) [45]. The set of all 2^m Walsh functions constitutes the complete set of these irreducible characters for $(\mathbb{Z}_2)^m$.

3. **From Characters to Basis Vectors:** Mathematically, a character of a group $(G, +)$ is a group homomorphism $\chi : G \to \mathbb{C}^*$, where $\mathbb{C}^*$ is the multiplicative group of non-zero complex numbers. The defining property is that for any $g_1, g_2 \in G$, the map preserves the structure: $\chi(g_1 + g_2) = \chi(g_1) \times \chi(g_2)$, where $+$ represents the additive operation of the group G and $\times$ means a multiplication of the field $\mathbb{C}$. A character is therefore an abstract function that maps a group's structure into a multiplicative one. We obtain a concrete vector from this abstract function by evaluating a character h_S on all 2^m points of its domain X in a fixed (lexicographical) order. Since the Walsh functions take values in $\{+1, -1\}$, this evaluation map transforms the character h_S into a real-valued vector $e_S \in \mathbb{R}^{2^m}$.

4. **The Real Vector Space and Irreducible Decomposition:** The set of all such vectors $\{e_S\}_{S \subseteq [m]}$ forms a complete orthogonal basis for the real vector space $\mathbb{R}^{2^m}$. The linear combinations of these basis vectors (over the field of real numbers $\mathbb{R}$) constitute this entire space. Because each basis vector originates from an irreducible character, this basis provides an **irreducible decomposition** of the space $\mathbb{R}^{2^m}$, where each axis corresponds to a fundamental symmetry of the underlying group.

Thus, we begin with an algebraic structure defined over the finite field $\mathbb{F}_2$ and, by leveraging its irreducible characters, we construct an orthogonal basis for a real vector space ($\mathbb{R}^{2^m}$) that reflects an irreducible decomposition of the original group's symmetries. It is within this real vector space $\mathbb{R}^{2^m}$, spanned by the irreducible characters of the underlying group, that we will introduce the multilinear operations of a Clifford algebra.

3.4.3 Spectral Decomposition of Design Matrices

Having established the Walsh basis, we now employ it to characterize the row space of our design matrices. The goal is to translate each row vector—a sequence of 0s and 1s representing a specific parameter or state—into a linear combination of Walsh basis vectors. This change of basis from the computational domain to the **spectral domain** will reveal the deep algebraic structure of our models, allowing us to compare their structures in a unified framework.

The transformation is a standard orthogonal projection. Any vector $\boldsymbol{a} \in \mathbb{R}^{2^m}$ can be uniquely decomposed into its Walsh basis components. Using the unnormalized

orthogonal basis vectors $\{e_S\}$, this decomposition is given by the formula:

$$\boldsymbol{a} = \sum_{S \subseteq [m]} \frac{\langle \boldsymbol{a}, e_S \rangle}{\|e_S\|^2} e_S, \tag{3.72}$$

where the squared norm for any basis vector is $\|e_S\|^2 = 2^m$. We now apply this transformation to the spanning vectors of the row spaces for the models introduced in Section 2.6.

Model 1: Complete Independence ($X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3$)

The design matrix A_1 for the complete independence model consists of six row vectors. Each row is the indicator vector for a marginal state of one of the three variables (e.g., the first row for $X_1 = 0$, the second for $X_1 = 1$, and so on). Transforming these vectors into the Walsh basis reveals their spectral components as follows:

$$A_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} e_0 + e_1 \\ e_0 - e_1 \\ e_0 + e_2 \\ e_0 - e_2 \\ e_0 + e_3 \\ e_0 - e_3 \end{bmatrix} \tag{3.73}$$

This decomposition reveals that the row space of A_1 is precisely the 4-dimensional space spanned by the Walsh basis vectors of order one or less: $\{e_0, e_1, e_2, e_3\}$. The absence of higher-order basis vectors like e_{12} reflects the total lack of interaction between variables.

Model 2: Joint Independence ($(X_1, X_2) \perp\!\!\!\perp X_3$)

The row space of A_2 is spanned by two sets of vectors: the indicator vectors for the joint states of (X_1, X_2) and those for the marginal states of X_3. Their spectral decomposition is:

$$A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} e_0 + e_1 + e_2 + e_{12} \\ e_0 + e_1 - e_2 - e_{12} \\ e_0 - e_1 + e_2 - e_{12} \\ e_0 - e_1 - e_2 + e_{12} \\ 2(e_0 + e_3) \\ 2(e_0 - e_3) \end{bmatrix} \tag{3.74}$$

The row space is the 5-dimensional space spanned by $\{e_0, e_1, e_2, e_3, e_{12}\}$. The presence of higher-order basis vectors e_{12} reflects the interaction between two variables X_1, X_2.

Model 3: Conditional Independence ($X_2 \perp\!\!\!\perp X_3 \mid X_1$)

The row space of A_3 is spanned by the indicator vectors for the joint states of the interacting pairs, (X_1, X_2) and (X_1, X_3). Their spectral decomposition is:

$$A_3 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} e_0 + e_1 + e_2 + e_{12} \\ e_0 + e_1 - e_2 - e_{12} \\ e_0 - e_1 + e_2 - e_{12} \\ e_0 - e_1 - e_2 + e_{12} \\ e_0 + e_1 + e_3 + e_{13} \\ e_0 + e_1 - e_3 - e_{13} \\ e_0 - e_1 + e_3 - e_{13} \\ e_0 - e_1 - e_3 + e_{13} \end{bmatrix} \tag{3.75}$$

The row space is thus the 6-dimensional space spanned by $\{e_0, e_1, e_2, e_3, e_{12}, e_{13}\}$. The presence of the second-order terms e_{12} and e_{13} algebraically encodes the interactions between (X_1, X_2) and (X_1, X_3), respectively.

Model 4: Context-Specific Independence ($X_2 \perp X_3 \mid X_1 = 0$)

The row space of A_4 is spanned by a heterogeneous set of vectors that define the context-specific structure. These include vectors for the marginal states of the context variable X_1, vectors defining the independence for the $X_1 = 0$ context, and vectors defining arbitrary dependence for the $X_1 = 1$ context. Their spectral decomposition is:

$$A_4 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} = \frac{1}{8} \begin{bmatrix} 4(e_0 + e_1) \\ 4(e_0 - e_1) \\ 2(e_0 + e_1 + e_2 + e_{12}) \\ 2(e_0 + e_1 - e_2 - e_{12}) \\ 2(e_0 + e_1 + e_3 + e_{13}) \\ 2(e_0 + e_1 - e_3 - e_{13}) \\ e_0 - e_1 + e_2 + e_3 - e_{12} + e_{13} - e_{23} - e_{123} \\ e_0 - e_1 + e_2 - e_3 - e_{12} - e_{13} + e_{23} + e_{123} \\ e_0 - e_1 - e_2 + e_3 + e_{12} - e_{13} - e_{23} + e_{123} \\ e_0 - e_1 - e_2 - e_3 + e_{12} + e_{13} + e_{23} - e_{123} \end{bmatrix} \tag{3.76}$$

Similarly, the row space of the design matrix A_4, which corresponds to a Minimum Invariant Constraint (MIC), is the 8-dimensional space spanned by all Walsh bases.

The Emergent Algebraic Pattern

The spectral decomposition of these diverse models reveals a striking commonality. The basis vectors of the row spaces are not arbitrary linear combinations of Walsh vectors; instead, they are consistently formed by highly structured combinations, such as $\frac{1}{2}(e_0 \pm e_i)$

and $\frac{1}{4}(e_0 \pm e_i \pm e_j \pm e_{ij})$. This recurring pattern suggests that these combinations are not accidental but represent fundamental algebraic building blocks for constructing statistical models.

This observation motivates the next step of our inquiry. To understand the properties of these building blocks and to manipulate them algebraically, we must first formalize a product structure between the Walsh basis vectors themselves. This will be the task of the next section, where we introduce the Clifford algebra.

3.5 Clifford Algebra

3.5.1 Constructing the Clifford Algebra

The Geometric Product and the Clifford Algebra Structure

As established in Lemma 3.4.1, the Walsh functions possess a natural multiplicative structure, $h_S(\boldsymbol{x})h_T(\boldsymbol{x}) = h_{S\Delta T}(\boldsymbol{x})$, that is based on a set-theoretic operation. This property is intrinsically linked to the concept of order (Definition 3.4.2), where the absence or presence of higher-order basis vectors like e_{12} in a model's spectral decomposition indicates the degree of interaction among its random variables. This motivates us to elevate this multiplicative property from a relationship between abstract functions to a concrete algebraic operation on the entire vector space $\mathbb{R}^{2^m}$.

Our goal, however, is not to introduce a simple commutative product. To eventually model the order-sensitive and asymmetric structures inherent in language, we require an algebra that can handle non-commutativity. Following the approach of prior work that constructs Clifford algebras from Walsh functions [57, 1], we define a product that intentionally incorporates a sign function to introduce the necessary alternating property.

The choice of this sign function determines the fundamental properties of the algebra, specifically whether the square of a basis vector, e_i^2, is $+1$ or -1. This choice corresponds to the signature of the space, as seen in physics where Minkowski spacetime is modeled with a signature of $(+,+,+,-)$. As the geometric properties of language are yet unknown, we adopt the most straightforward and neutral assumption: that all basis vectors represent equivalent dimensions. This leads us to define a product where all first-order basis vectors square to $+1$, which generates the Clifford algebra of Euclidean space, $\mathrm{Cl}(m, 0)$.

Definition 3.5.1 (The Geometric Product)**.** Let $\{e_S\}_{S\subseteq[m]}$ be the unnormalized Walsh basis for the vector space $\mathbb{R}^{2^m}$. The **geometric product** (or Clifford product) is a bilinear operation on this space, defined on the basis vectors as:

$$e_S e_T := \omega(S,T)\, e_{S\Delta T}, \tag{3.77}$$

where $S\Delta T = (S \cup T) \setminus (S \cap T)$ is the symmetric difference of the sets, and $\omega(S,T)$ is a sign function defined by:

$$\omega(S,T) := (-1)^{\sum_{i\in S}\sum_{j\in T}\mathbb{I}_{i>j}}, \tag{3.78}$$

with $\mathbb{I}_{i>j}$ being the indicator function that is 1 if $i > j$ and 0 otherwise.

The function $\omega(S, T) \in \{+1, -1\}$ is not arbitrary; it is specifically designed to satisfy the **2-cocycle condition**, which is the mathematical requirement that ensures the geometric product is associative, i.e., $(e_S e_T)e_U = e_S(e_T e_U)$ for all basis vectors. While guaranteeing associativity, this cocycle simultaneously introduces the necessary non-commutativity into the algebra. This definition ensures two fundamental properties for the first-order basis vectors (e_i for $i \in [m]$):

$$e_i e_j = -e_j e_i \quad (i \neq j), \qquad e_i^2 = e_0, \tag{3.79}$$

where e_0 is the multiplicative identity. These are precisely the defining relations of the real Clifford algebra $\mathrm{Cl}(m, 0)$. Hence, the vector space $\mathbb{R}^{2^m}$, equipped with this multiplication rule, becomes an algebra isomorphic to $\mathrm{Cl}(m, 0)$. This construction shows that the Clifford algebra can be realized as a *twisted group ring* of $(\mathbb{Z}_2)^m$, with the Walsh basis providing a concrete representation [86].

Remark on an alternative formulation This product is sometimes expressed using binary vectors instead of sets, a formulation attributed to Brauer and Weyl [21]. If we identify each set S with a binary vector $\boldsymbol{s} \in \{0, 1\}^m$ and the symmetric difference $S \Delta T$ with the XOR operation $\boldsymbol{s} \oplus \boldsymbol{t}$, the product can be written as:

$$e_{\boldsymbol{s}} e_{\boldsymbol{t}} := (-1)^{\sum_{i>j} s_i t_j} e_{\boldsymbol{s} \oplus \boldsymbol{t}}. \tag{3.80}$$

This highlights the deep connection between the set-theoretic, group-theoretic, and vector-space perspectives.

Fundamental Properties of the Clifford Algebra

Having defined the geometric product in the preceding section, we now systematically enumerate the fundamental algebraic properties it bestows upon our vector space. These properties confirm that we have constructed a well-behaved, consistent algebraic system and provide the foundational rules for all subsequent manipulations.

Proposition 3.5.1 (Properties of the Geometric Product)**.** The geometric product, defined on the Walsh basis as $e_S e_T = \omega(S, T)\, e_{S \Delta T}$, satisfies the following properties for all subsets $S, T, U \subseteq [m]$:

(i) **Closure:** The set of Walsh basis vectors $\{e_S\}_{S \subseteq [m]}$ is closed under the geometric product.

(ii) **Associativity:** $(e_S e_T)e_U = e_S(e_T e_U)$

(iii) **Identity Element:** The zeroth-order basis vector $e_\emptyset$ (denoted e_0) serves as the multiplicative identity: $e_0 e_S = e_S e_0 = e_S$.

(iv) **Inverse Elements:** Every basis element with order 1 (i.e., $|S| = 1$) is its own inverse: $e_S e_S = e_0$, which implies $e_S^{-1} = e_S$.

(v) **Anticommutation of First-Order Bases:** For any distinct $i, j \in [m]$, the first-order basis vectors anticommute: $e_i e_j = -e_j e_i$.

Proof. (i) **Closure:** Since $S\Delta T$ is always a subset of $[m]$, the product of any two basis elements $e_S e_T$ results in another basis element $e_{S\Delta T}$ (up to a sign), which is a member of the original set.

(ii) **Associativity** is guaranteed by the fact that the sign function $\omega(S, T)$ satisfies the 2-cocycle condition, as was necessary for its definition.

(iii) **Identity Element:** This follows from the properties of the symmetric difference, as $S\Delta\emptyset = S$, and the fact that $\omega(S, \emptyset) = \omega(\emptyset, S) = 1$. Thus, $e_\emptyset e_S = e_S e_\emptyset = e_S$.

(iv) **Inverse Elements:** The product of a basis element with itself is $e_S e_S = \omega(S, S) e_{S\Delta S}$. Since $S\Delta S = \emptyset$, this simplifies to $e_S e_S = \omega(S, S) e_\emptyset$. The definition of $\omega(S, S)$ yields $(-1)^{\sum_{i\in S}\sum_{j\in S}\mathbb{I}_{i>j}}$, and as the condition $i > i$ is never satisfied. Thus, $\omega(S, S) = (-1)^0 = 1$, which gives $e_S^2 = \omega(S, S) e_\emptyset = e_0$.

(v) **Anticommutation:** For one-element subsets $S = \{i\}$ and $T = \{j\}$ with $i \neq j$:

$$e_i e_j = \omega(\{i\}, \{j\})\, e_{\{i,j\}} = (-1)^{\mathbb{I}_{i>j}}\, e_{\{i,j\}},$$
$$e_j e_i = \omega(\{j\}, \{i\})\, e_{\{i,j\}} = (-1)^{\mathbb{I}_{j>i}}\, e_{\{i,j\}}.$$

Since $\mathbb{I}_{i>j} + \mathbb{I}_{j>i} = 1$ for $i \neq j$, it follows that $e_i e_j = -e_j e_i$.

□

Decomposition of the Geometric Product: Inner and Outer Products

The geometric product is powerful precisely because it unifies multiple operations. To understand and utilize it, we must first decompose it into its fundamental components. This decomposition relies on the natural hierarchical structure of the algebra, known as a grading , which classifies elements based on their order or complexity.

A basis vector e_S has a **grade** equal to the cardinality of its index set, $|S|$. A linear combination of grade-k basis vectors is called a **k-vector**. The set of all k-vectors forms a subspace of the algebra, and the entire algebra is the direct sum of these graded subspaces: $Cl(m, 0) = \bigoplus_{k=0}^{m} Cl^k(m, 0)$. Any element of the algebra, called a **multivector**, can be uniquely written as a sum of its k-vector parts. We use the notation $\langle M \rangle_k$ to denote the grade-k part of a multivector M. The most fundamental k-vectors are **k-blades**, which are elements that can be factored into the geometric product of k orthogonal vectors (our basis vectors e_S are simple k-blades).

With this grading machinery, we can define two specialized products that are contained within the geometric product[88].

Definition 3.5.2 (Inner and Outer Products)**.** Let A be a j-vector and B be a k-vector.

1. The **inner product** of A and B is the lowest-grade part of their geometric product:

$$A \cdot B := \langle AB \rangle_{|j-k|} \tag{3.81}$$

2. The **outer product** (or wedge product) of A and B is the highest-grade part of their geometric product:

$$A \wedge B := \langle AB \rangle_{j+k} \tag{3.82}$$

A foundational property of the geometric product is that for a vector a and any multivector B, it decomposes perfectly into these two parts: $aB = a \cdot B + a \wedge B$. For the special case of two vectors, e_i and e_j, this leads to a particularly elegant and useful decomposition.

Proposition 3.5.2. The geometric product of two vectors decomposes into a symmetric part (the inner product) and an antisymmetric part (the outer product):

$$e_i \cdot e_j = \frac{1}{2}(e_i e_j + e_j e_i) \tag{3.83}$$

$$e_i \wedge e_j = \frac{1}{2}(e_i e_j - e_j e_i) \tag{3.84}$$

These two products have distinct and complementary geometric interpretations. The **inner product** is a *contraction*; it reduces grade and captures metric information like projections, angles, and magnitudes. It is the algebraic analogue of projection. The **outer product** is *generative*; it increases grade and creates new elements representing higher-dimensional subspaces. For instance, the outer product of two vectors, $e_i \wedge e_j$, is a bivector that represents the oriented plane spanned by them. With these specialized tools defined, we are now equipped to analyze the structure of our design matrices and algebraically model probabilistic operations like conditioning and marginalization.

3.5.2 Algebraic Representation via Idempotent Projectors

Formalizing the Algebraic Pattern: Idempotent Projectors

In 3.4.3, our spectral decomposition of the design matrices revealed a recurring algebraic motif: row vectors consistently transformed into structured combinations of the form $\frac{1}{2}(e_0 \pm e_i)$. This observation suggested that these were not accidental patterns but fundamental building blocks. With the geometric product now at our disposal, we can rigorously analyze the properties of these combinations and formalize their role as projection operators.

Definition 3.5.3 (Idempotent Projectors)**.** For each first-order basis vector e_i (where $i \in \{1, \ldots, m\}$), we define a pair of **idempotent projectors** as:

$$P_i^{(+)} := \frac{e_0 + e_i}{2} \tag{3.85}$$

$$P_i^{(-)} := \frac{e_0 - e_i}{2} \tag{3.86}$$

These elements are called "projectors" because, when applied via the geometric product, they project multivectors onto specific subspaces. They possess a set of elegant and powerful properties that make them the ideal building blocks for our algebraic models.

Proposition 3.5.3 (Properties of Projectors)**.** The idempotent projectors possess the following properties, where we denote $P_i^{(+)}$ as P_i and $P_i^{(-)}$ as $\bar{P}_i$ for brevity:

(i) **Idempotency:** $(P_i)^2 = P_i$ and $(\bar{P}_i)^2 = \bar{P}_i$.

(ii) **Orthogonality:** $P_i\bar{P}_i = 0$.

(iii) **Completeness:** $P_i + \bar{P}_i = e_0$, where e_0 is the multiplicative identity.

(iv) **Eigenspace Projection:** $e_iP_i = P_i$ and $e_i\bar{P}_i = -\bar{P}_i$.

Proof. The proofs follow directly from the properties of the geometric product established in 3.5.1, namely that e_0 is the identity ($e_0e_i = e_i$) and each first-order basis vector squares to the identity ($e_i^2 = e_0$ for $i > 0$).

(i) **Idempotency:** We show this for P_i:

$$(P_i)^2 = \left(\frac{e_0 + e_i}{2}\right)^2 = \frac{e_0^2 + e_0e_i + e_ie_0 + e_i^2}{4} = \frac{e_0 + e_i + e_i + e_0}{4} = \frac{2e_0 + 2e_i}{4} = P_i. \quad (3.87)$$

(ii) **Orthogonality:**

$$P_i\bar{P}_i = \left(\frac{e_0 + e_i}{2}\right)\left(\frac{e_0 - e_i}{2}\right) = \frac{e_0^2 - e_0e_i + e_ie_0 - e_i^2}{4} = \frac{e_0 - e_i + e_i - e_0}{4} = 0. \quad (3.88)$$

(iii) **Completeness:**

$$P_i + \bar{P}_i = \frac{e_0 + e_i}{2} + \frac{e_0 - e_i}{2} = \frac{2e_0}{2} = e_0. \quad (3.89)$$

(iv) **Eigenspace Projection:**

$$e_iP_i = e_i\left(\frac{e_0 + e_i}{2}\right) = \frac{e_ie_0 + e_i^2}{2} = \frac{e_i + e_0}{2} = P_i. \quad (3.90)$$

The proof for $e_i\bar{P}_i = -\bar{P}_i$ is analogous.

□

These properties are highly significant. Idempotency confirms that these operators are indeed projectors. Orthogonality and completeness ensure that they partition the entire algebra into a set of mutually exclusive, complementary subspaces. Finally, the fourth property reveals the precise nature of these subspaces: P_i projects onto the subspace where the "frequency component" e_i is invariant (its eigenvalue is +1), while $\bar{P}_i$ projects onto the

subspace where e_i is anti-invariant (its eigenvalue is -1). Thus, the pair $(P_i, \bar{P}_i)$ forms a complete set of orthogonal projectors that diagonalizes the basis vector e_i.

This decomposition effectively categorizes the entire algebraic space based on how its elements "respond" to the operator e_i. Conceptually, these projectors partition the space into two fundamental sectors: P_i isolates the components that are in perfect agreement with the symmetry of e_i (the +1 eigenspace), while $\bar{P}_i$ isolates those that stand in direct opposition to it (the -1 eigenspace). By resolving the space into these concordant and discordant factions, we reduce the complex action of e_i to a binary orientation, allowing us to analyze the algebra through the lens of this specific symmetry.

An Algebraic Catalogue of Independence Models

With the geometric product defined and the idempotent projectors identified as fundamental building blocks, we are now positioned to fulfill the primary goal of this chapter: to represent the structure of statistical models in a purely algebraic language. In this section, we revisit the design matrices from our examples and translate their row spaces into the language of projectors. This translation will not only simplify their representation but also make their underlying structural similarities and differences immediately apparent.

The key to this translation is the factorization of the spectral components. For instance, the indicator vector for the state $X_1 = 0$, which decomposes to $\frac{1}{2}(e_0 + e_1)$, is now simply written as $P_1^{(+)}$. Similarly, the indicator vector for the joint state $(X_1, X_2) = (0, 0)$, which decomposes to $\frac{1}{4}(e_0 + e_1 + e_2 + e_{12})$, can be factorized as follows:

$$\frac{1}{4}(e_0 + e_1 + e_2 + e_{12}) = \frac{1}{4}(e_0 + e_1)(e_0 + e_2) = \left(\frac{e_0 + e_1}{2}\right)\left(\frac{e_0 + e_2}{2}\right) = P_1^{(+)}P_2^{(+)}. \tag{3.91}$$

This reveals that the algebraic object for a joint state is the *product* of the objects for the marginal states. This principle allows us to build a catalogue of our models.

Model 1: Complete Independence ($X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3$) The row space of A_1 is spanned by the indicator vectors for the marginal states of each variable. In the projector formalism, their spectral decomposition is remarkably simple:

$$A_1 = \frac{1}{2}\begin{bmatrix} e_0 + e_1 \\ e_0 - e_1 \\ e_0 + e_2 \\ e_0 - e_2 \\ e_0 + e_3 \\ e_0 - e_3 \end{bmatrix} = \begin{bmatrix} P_1^+ \\ P_1^- \\ P_2^+ \\ P_2^- \\ P_3^+ \\ P_3^- \end{bmatrix} \tag{3.92}$$

The algebraic form consists of a simple collection of individual projectors, $\{P_1^{(\pm)}, P_2^{(\pm)}, P_3^{(\pm)}\}$. This transparently reveals that the row space is the 4-dimensional space spanned by the Walsh basis vectors of order one or less: $\{e_0, e_1, e_2, e_3\}$. The absence of any

products of projectors algebraically reflects the complete absence of interactions between variables; the model's structure is, in essence, a direct sum along three independent axes.

Model 2: Joint Independence ($(X_1, X_2) \perp\!\!\!\perp X_3$) The row space of A_2 is spanned by indicator vectors for the joint states of (X_1, X_2) and the marginal states of X_3. This hybrid structure is elegantly captured in the projector formalism:

$$A_2 = \frac{1}{4}\begin{bmatrix} e_0 + e_1 + e_2 + e_{12} \\ e_0 + e_1 - e_2 - e_{12} \\ e_0 - e_1 + e_2 - e_{12} \\ e_0 - e_1 - e_2 + e_{12} \\ 2e_0 + 2e_3 \\ 2e_0 - 2e_3 \end{bmatrix} = \frac{1}{4}\begin{bmatrix} \frac{1}{2}(e_0 + e_1)\frac{1}{2}(e_0 + e_2) \\ \frac{1}{2}(e_0 + e_1)\frac{1}{2}(e_0 - e_2) \\ \frac{1}{2}(e_0 - e_1)\frac{1}{2}(e_0 + e_2) \\ \frac{1}{2}(e_0 - e_1)\frac{1}{2}(e_0 - e_2) \\ \frac{1}{2}(e_0 + e_3) \\ \frac{1}{2}(e_0 - e_3) \end{bmatrix} = \begin{bmatrix} P_1^+ P_2^+ \\ P_1^+ P_2^- \\ P_1^- P_2^+ \\ P_1^- P_2^- \\ P_3^+ \\ P_3^- \end{bmatrix} \tag{3.93}$$

This algebraic form, $\{P_1^{(\pm)} P_2^{(\pm)}, P_3^{(\pm)}\}$, provides a beautiful picture of the model's structure as a direct sum of two distinct blocks. One block, $\{P_1^{(\pm)} P_2^{(\pm)}\}$, represents the interacting pair of variables, while the other, $\{P_3^{(\pm)}\}$, represents the independent variable. The resulting row space is the 5-dimensional space spanned by $\{e_0, e_1, e_2, e_3, e_{12}\}$, which can be seen as the direct sum of the subspaces $W_1 = \mathrm{span}\{e_0, e_1, e_2, e_{12}\}$ and $W_2 = \mathrm{span}\{e_0, e_3\}$ (sharing the identity e_0).

Model 3: Conditional Independence ($X_2 \perp\!\!\!\perp X_3 \mid X_1$) The row space of A_3 is spanned by the indicator vectors for the joint states of the pairs (X_1, X_2) and (X_1, X_3). The factorization of their spectral components into projectors is as follows:

$$A_3 = \frac{1}{4}\begin{bmatrix} e_0 + e_1 + e_2 + e_{12} \\ e_0 + e_1 - e_2 - e_{12} \\ e_0 - e_1 + e_2 - e_{12} \\ e_0 - e_1 - e_2 + e_{12} \\ e_0 + e_1 + e_3 + e_{13} \\ e_0 + e_1 - e_3 - e_{13} \\ e_0 - e_1 + e_3 - e_{13} \\ e_0 - e_1 - e_3 + e_{13} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(e_0 + e_1)\frac{1}{2}(e_0 + e_2) \\ \frac{1}{2}(e_0 + e_1)\frac{1}{2}(e_0 - e_2) \\ \frac{1}{2}(e_0 - e_1)\frac{1}{2}(e_0 + e_2) \\ \frac{1}{2}(e_0 - e_1)\frac{1}{2}(e_0 - e_2) \\ \frac{1}{2}(e_0 + e_1)\frac{1}{2}(e_0 + e_3) \\ \frac{1}{2}(e_0 + e_1)\frac{1}{2}(e_0 - e_3) \\ \frac{1}{2}(e_0 - e_1)\frac{1}{2}(e_0 + e_3) \\ \frac{1}{2}(e_0 - e_1)\frac{1}{2}(e_0 - e_3) \end{bmatrix} = \begin{bmatrix} P_1^+ P_2^+ \\ P_1^+ P_2^- \\ P_1^- P_2^+ \\ P_1^- P_2^- \\ P_1^+ P_3^+ \\ P_1^+ P_3^- \\ P_1^- P_3^+ \\ P_1^- P_3^- \end{bmatrix} \tag{3.94}$$

Here, the algebraic structure reveals the statistical model with perfect clarity. The products of projectors, such as $P_1^{(+)} P_2^{(+)}$, represent the interaction between variables X_1 and X_2. The fact that projectors for X_1 (i.e., $P_1^{(\pm)}$) appear in *both* sets of products, $\{P_1^{(\pm)} P_2^{(\pm)}, P_1^{(\pm)} P_3^{(\pm)}\}$, algebraically encodes its role as the central, conditioning variable. Consequently, the row space is the 6-dimensional space spanned by $\{e_0, e_1, e_2, e_3, e_{12}, e_{13}\}$. This space can be understood as the union of two interacting blocks, $W_1 = \mathrm{span}\{e_0, e_1, e_2, e_{12}\}$ and $W_2 = \mathrm{span}\{e_0, e_1, e_3, e_{13}\}$, which share the common conditioning factor represented by $\mathrm{span}\{e_0, e_1\}$.

Model 4: Context-Specific Independence ($X_2 \perp\!\!\!\perp X_3 \mid X_1 = 0$) The row space of A_4 is spanned by a heterogeneous set of vectors that define the context-specific structure. These include vectors for the marginal states of the context variable X_1, vectors defining the independence for the $X_1 = 0$ context, and vectors defining arbitrary dependence for the $X_1 = 1$ context. Their spectral decomposition and subsequent factorization into projectors are as follows:

$$A_4 = \frac{1}{8} \begin{bmatrix} 4e_0 + 4e_1 \\ 4e_0 - 4e_1 \\ 2e_0 + 2e_1 + 2e_2 + 2e_{12} \\ 2e_0 + 2e_1 - 2e_2 - 2e_{12} \\ 2e_0 + 2e_1 + 2e_3 + 2e_{13} \\ 2e_0 + 2e_1 - 2e_3 - 2e_{13} \\ e_0 - e_1 + e_2 + e_3 - e_{12} + e_{13} - e_{23} - e_{123} \\ e_0 - e_1 + e_2 - e_3 - e_{12} - e_{13} + e_{23} + e_{123} \\ e_0 - e_1 - e_2 + e_3 + e_{12} - e_{13} - e_{23} + e_{123} \\ e_0 - e_1 - e_2 - e_3 + e_{12} + e_{13} + e_{23} - e_{123} \end{bmatrix} = \begin{bmatrix} P_1^+ \\ P_1^- \\ P_1^+ P_2^+ \\ P_1^+ P_2^- \\ P_1^+ P_3^+ \\ P_1^+ P_3^- \\ P_1^- P_2^+ P_3^+ \\ P_1^- P_2^+ P_3^- \\ P_1^- P_2^- P_3^+ \\ P_1^- P_2^- P_3^- \end{bmatrix} \tag{3.95}$$

The resulting algebraic expression reveals the model's context-specific nature with remarkable clarity. The elements prefixed by $P_1^{(+)}$ describe the model for the $X_1 = 0$ context: a structure of conditional independence between X_2 and X_3. In contrast, the elements prefixed by $P_1^{(-)}$ describe the model for the $X_1 = 1$ context: a fully saturated model where all three variables interact, represented by the triple product of projectors. This powerfully demonstrates that conditioning on a state is algebraically represented by left-multiplication of the corresponding projector. As the spanning set contains all possible products of projectors, the row space is the full 8-dimensional space spanned by all Walsh bases.

Summary: A Unified View This algebraic catalogue reveals a profound structural correspondence. Different independence models, which are described by disparate graphical topologies or complex probability factorizations, are now represented as simple expressions in a single, unified algebra. The structural differences between models are reduced to clear algebraic distinctions:

- **Addition** of projectors corresponds to a **direct sum** of model components (e.g., separating independent variables).
- **Multiplication** of projectors corresponds to **interaction** or **conditioning** between variables.

Consequently, a structural parallel emerges among three distinct mathematical domains: the independence constraints in probability theory, the composition of the row space in linear algebra, and the grade of the elements in our Clifford algebra.

Application: Analyzing a Variety Through Algebra

As previously established, a probability vector $\boldsymbol{p}$ generated by a toric model is constrained to lie on an algebraic variety, which is the image of a monomial map $\phi_A(\theta) = \frac{1}{Z(\theta)} \exp(\theta^\top \cdot A)$ defined by the design matrix A. The parameter-invariant constraints that define this variety are captured by the toric ideal I_A, which is entirely determined by the kernel of the design matrix, $\ker(A)$. Having established a bridge from the design matrix A to its expression in Clifford algebra, we are now positioned to leverage the analytical power of this new representation. Specifically, we can now analyze the internal combinatorial structure of the row space of the design matrix through direct algebraic operations, a task for which conventional tools like matrix rank are insufficient.

This algebraic framework provides a new computational perspective for model analysis, allowing us to derive statistical properties through direct manipulation of equations. To illustrate this, let us revisit the conditional independence model $X_2 \perp\!\!\!\perp X_3 \mid X_1$, which in graphical models corresponds to a structure with interaction pairs (X_1, X_2) and (X_1, X_3). The multivector representing the logit part of this model, $\mathcal{L}_3 = \boldsymbol{\theta}^\top A_3$, can be algebraically factorized as follows:

$$\begin{aligned}\mathcal{L}_3 &= \theta_1 P_1^{(+)} P_2^{(+)} + \theta_2 P_1^{(+)} P_2^{(-)} + \theta_3 P_1^{(-)} P_2^{(+)} + \theta_4 P_1^{(-)} P_2^{(-)} \\ &\quad + \theta_5 P_1^{(+)} P_3^{(+)} + \theta_6 P_1^{(+)} P_3^{(-)} + \theta_7 P_1^{(-)} P_3^{(+)} + \theta_8 P_1^{(-)} P_3^{(-)} \qquad (3.96) \\ &= P_1^{(+)} (\theta_1 P_2^{(+)} + \theta_2 P_2^{(-)} + \theta_5 P_3^{(+)} + \theta_6 P_3^{(-)}) \\ &\quad + P_1^{(-)} (\theta_3 P_2^{(+)} + \theta_4 P_2^{(-)} + \theta_7 P_3^{(+)} + \theta_8 P_3^{(-)}) \qquad (3.97)\end{aligned}$$

This algebraic refactoring has a profound structural interpretation. The initial expression represents the model as a sum of interactions between pairs (X_1, X_2) and (X_1, X_3). The final expression, factored by the projectors $P_1^{(+)}$ and $P_1^{(-)}$, reveals an equivalent structure: within the subspace defined by $X_1 = 0$ (projected by $P_1^{(+)}$), the variables X_2 and X_3 are independent, and the same holds for the subspace defined by $X_1 = 1$ (projected by $P_1^{(-)}$). This algebraic manipulation is a direct, computable proof of the conditional independence relationship $X_2 \perp\!\!\!\perp X_3 \mid X_1$.

This is precisely the concept of **d-separation**, a cornerstone of reasoning in Probabilistic Graphical Models (PGMs)[74]. While PGMs derive independence from the topology of a graph, our framework offers a direct algebraic representation and computational method for the same underlying structural relationship. It provides a computationally transparent algebraic foundation for graphical rules like d-separation and presents an alternative perspective to other algebraic formalisms, such as Studený's imsets[132].

The power of this algebraic lens extends to all the models discussed. The structure of independence is read directly from the Clifford algebra expression:

- **Model A_1 (Complete Independence)**: The algebraic form consists only of individual projectors $\{P_1^{(\pm)}, P_2^{(\pm)}, P_3^{(\pm)}\}$. The absence of any products of projectors algebraically reflects the complete absence of interactions.

- **Model A_2 (Joint Independence)**: The expression $\{P_1^{(\pm)}P_2^{(\pm)},\ P_3^{(\pm)}\}$ clearly partitions the model into an interacting block for (X_1, X_2) and a separate, independent block for X_3, which corresponds to a direct-sum decomposition of the underlying vector space.

- **Model A_4 (Context-Specific Independence)**: The algebraic structure makes the context-dependency explicit. The subspace projected by $P_1^{(+)}$ (context $X_1 = 0$) is a direct sum reflecting independence, while the subspace projected by $P_1^{(-)}$ (context $X_1 = 1$) contains interaction terms, representing a dependent structure.

In summary, the proposed Clifford algebra framework represents the structure behind graphical rules like d-separation through elegant algebraic computation. This provides a unified and computationally transparent method for analyzing, comparing, and understanding the structure of multivariate probability distributions.

3.6 Log Semiring

3.6.1 Extension to Probability Space

In the preceding sections, we developed an algebraic formalism defined entirely on a linear space, where both parameters and design matrices reside. The goal of this section is to extend that algebra to the probability space itself—an exponential space—so that both domains can be treated within a unified algebraic framework. Specifically, probabilities are modeled by a toric model expressed in the softmax form:

$$\boldsymbol{p}^\top = \mathrm{softmax}(\text{logit}), \qquad \text{logit} = \theta^\top A, \tag{3.98}$$

where $\boldsymbol{p}$ is the probability vector, θ denotes the parameter vector, and A is the design matrix. This section establishes a bridge between the linear structure of the logit—reflecting the row-space geometry of A—and the probability vector $\boldsymbol{p}$ obtained after the softmax transformation. We introduce the *log-semiring* as the algebraic device that enables probabilistic operations to be handled in a consistent, linear-like manner within our framework.

The need for this extension arises from three distinct but interconnected notions of linearity in language modeling:

1. *Empirical linearity* — observed in vector spaces derived from probabilities. Distributional semantics has shown that word vectors exhibit remarkable linear regularities, such as

$$\boldsymbol{v}_{\text{king}} - \boldsymbol{v}_{\text{man}} + \boldsymbol{v}_{\text{woman}} \approx \boldsymbol{v}_{\text{queen}},$$

suggesting that semantic relations form approximately parallelogram-shaped structures in embedding space.

2. *Operational linearity* — required for probabilistic computation. A co-occurrence matrix results from marginalizing a higher-order probability tensor, which is an additive

operation over probabilities.

3. *Theoretical linearity* — inherent in the parameter space of the toric model. The probability vector $\boldsymbol{p}$ belongs to an exponential family governed by $\boldsymbol{p} \propto \exp(\theta^\top A)$. Although $\boldsymbol{p}$ itself is non-linear in θ, its canonical parameters and design matrix lie in a linear log domain.

These three aspects highlight a central challenge: how to connect the multilinear structure of Clifford algebra to both the theoretical linearity of the model's parameters and the empirical regularities observed in probability space. A direct linear treatment of probabilities is infeasible, yet a systematic link between the linear log domain and the non-linear exponential domain is essential for compositional reasoning and analytical tractability.

The connection between these domains is mediated by the monotone functions $\exp$ and $\log$. While they do not preserve linearity, they preserve the *order structure* between the two spaces. This correspondence can be formalized as a Galois connection (adjunction) between partially ordered sets.

Lemma 3.6.1 (Galois Connection of Exp/Log)**.** Let $\exp : \overline{\mathbb{R}} \to \mathbb{R}_{\geq 0}$ and $\log : \mathbb{R}_{\geq 0} \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty\}$. Both mappings are monotone, and for any $a \in \overline{\mathbb{R}}$ and $b \in \mathbb{R}_{\geq 0}$,

$$\exp(a) \leq b \quad \Longleftrightarrow \quad a \leq \log(b). \tag{3.99}$$

Hence $(\exp \dashv \log)$ forms a Galois connection between the two ordered sets.

Proof. The equivalence $e^a \leq b \Leftrightarrow a \leq \log b$ follows directly from the strict monotonicity of $\exp$ and $\log$. □

This adjunction provides the formal foundation for linking the log and probability domains. It guarantees an order-preserving correspondence even without linearity. To make this connection computationally useful—particularly for representing probabilistic summation within the log domain—we introduce the *log-semiring*. By equipping our Clifford algebra with this structure, we integrate all three notions of linearity into a single algebraic system.

The log-semiring offers a coherent way to perform probabilistic calculations on a logarithmic scale. Its key idea is to define addition and multiplication by transporting operations through the exp–log mapping: exponentiate from the log domain, perform ordinary arithmetic in the exponential domain, and map back via the logarithm.

Definition 3.6.1 (Log-Semiring)**.** Let $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty\}$ be the set of extended real numbers of negative infinity equipped with two binary operations:

$$x \boxplus y := \mathrm{LSE}(x, y) = \log(e^x + e^y), \qquad \text{(Log-Sum-Exp addition)} \tag{3.100}$$

$$x \boxtimes y := x + y, \qquad \text{(standard addition as multiplication)} \tag{3.101}$$

with additive identity $\mathbf{0} = -\infty$ and multiplicative identity $\mathbf{1} = 0$. The algebraic structure $(\overline{\mathbb{R}}, \boxplus, \boxtimes)$ is called the **log-semiring**.

The two operations correspond directly to probabilistic requirements:

- $\boxplus$ (**LSE addition**) mirrors the summation of probabilities in the exponential domain.
- $\boxtimes$ (**log-space addition**) represents multiplication of probabilities, which becomes addition in the log domain.

It is a semiring because it lacks additive inverses; for any $x \in \mathbb{R}$, there is no x' such that $\log\left(e^x + e^{x'}\right) = -\infty$. This mirrors the fact that $\exp$ maps to $\mathbb{R}_{\geq 0}$, where standard addition has no inverse for positive values.

This semiring is closely related to the tropical (max-plus) algebra. As the logarithm base tends to infinity, the LogSumExp operation converges to the maximum function:

$$\lim_{b \to \infty} \log_b(b^x + b^y) = \max(x, y),$$

a process known as *tropicalization* or *dequantization*[89] . Thus, the log-semiring can be viewed as a smooth deformation of the tropical semiring. With this formal structure in place, we are now prepared to define probabilistic marginalization within the Clifford-algebraic framework.

3.6.2 Marginalization as Ordered Projection Contraction

In this section, we reformulate probabilistic marginalization as a geometric contraction within the Clifford algebra. Unlike classical marginalization, which is defined as a commutative sum over variables, our formulation preserves the *order* of projection operations, thereby allowing us to track the sequential structure of variable elimination. This order-tracking is essential: it encodes syntactic or causal dependencies that would otherwise be lost in a purely commutative probability space.

We consider m binary random variables $\{X_i\}_{i \in [m]}$, where each X_i takes values in $\mathcal{X}_i = \{+1, -1\}$. The joint configuration space is $\mathcal{X} := \{-1, +1\}^m$.

For each variable X_i, we introduce a pair of idempotent projectors $P_i^{(+)}$ and $P_i^{(-)}$, defined as in Section 3.5.2:

$$P_i^{(\pm)} = \frac{1}{2}(e_0 \pm e_i),$$

which satisfy the properties of idempotency, orthogonality, completeness, and eigenspace projection proven there. We do not restate those properties here, but we make essential use of them below.

In the Clifford algebra, the basis elements e_i satisfy $e_i e_j = -e_j e_i$ for $i \neq j$. Therefore, the projectors $P_i^{(\pm)}$ do not commute in general:

$$P_i^{(\sigma_i)} P_j^{(\sigma_j)} \neq P_j^{(\sigma_j)} P_i^{(\sigma_i)}, \quad \sigma_i, \sigma_j \in \{+, -\}.$$

Hence, each joint configuration must be represented not merely as a set of states $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_m)$, but as an *ordered projection sequence*

$$P^{\boldsymbol{\sigma}} := P_1^{(\sigma_1)} P_2^{(\sigma_2)} \cdots P_m^{(\sigma_m)}. \tag{3.102}$$

The order of multiplication determines the resulting subspace: $P_i^{(+)} P_j^{(+)}$ and $P_j^{(+)} P_i^{(+)}$ correspond to distinct geometric projections within the algebra. This non-commutativity encodes the syntactic or directional structure among variables.

Logit representation and marginalization in the Clifford–log semiring framework. In the toric model, the logit is a linear form

$$\text{logit} = \boldsymbol{\theta}^\top A,$$

where $A \in \mathbb{R}^{d \times n}$ is the design matrix and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ is the parameter vector. Each row of A can be represented by a Clifford polynomial constructed from the elementary idempotent projectors $P_k^{(\pm)} = \frac{1}{2}(e_0 \pm e_k)$ introduced in Section 3.5.2. Let Π_i denote the Clifford element corresponding to the i-th row:

$$\Pi_i = \prod_{k \in S_i} P_k^{(\sigma_{ik})}, \qquad \sigma_{ik} \in \{+, -\}, \tag{3.103}$$

where $S_i \subseteq [m]$ specifies which variables participate in the feature. Each Π_i encodes the structural composition of that feature within the Clifford algebra.

The Clifford embedding of the toric logit is then written as

$$\mathcal{L} = \sum_{i=1}^{d} \theta_i \, \Pi_i, \tag{3.104}$$

where the coefficients θ_i belong to the log-semiring $(\overline{\mathbb{R}}, \boxplus, \boxtimes)$ with $\boxplus = \log\left(e^a + e^b\right)$ and $\boxtimes = +$. This representation preserves the linear structure of the toric model in the log-domain while making explicit the polynomial composition of features in the projector basis $\{P_k^{(\pm)}\}$.

Let $K \subseteq [m]$ be a subset of variables to be marginalized out and $K^c = [m] \setminus K$ the set to be retained. In the classical probability space, marginalization corresponds to summing over the states of X_K. In the Clifford formulation, the same operation is realized as a combination of (1) a *geometric contraction* of the projectors associated with K and (2) a *log-semiring aggregation* of the corresponding coefficients.

Definition 3.6.2 (Marginalization operator on the Clifford–log semiring)**.** Let $\pi_K : C \to C$ be the Clifford projection that replaces each projector $P_j^{(\pm)}$ with the identity e_0 for $j \in K$, removing the corresponding directions from the product. The marginalization operator

M_K acting on $\mathcal{L} = \sum_i \theta_i \Pi_i$ is defined as

$$M_K(\mathcal{L}) = \sum_{\Pi' \in \pi_K(\{\Pi_i\})} \Big(\boxplus_{\{\, i | \pi_K(\Pi_i) = \Pi' \,\}} \theta_i \Big) \Pi'. \tag{3.105}$$

The inner $\boxplus$ performs the log-sum-exp aggregation over coefficients belonging to identical projected structures Π', and the outer summation reconstructs the resulting element in the Clifford algebra of the retained variables.

Remark (domain of applicability). Equation (3.105) is to be applied only after $\mathcal{L}$ has been expanded in the full-state idempotent basis $\{\Pi_x = \prod_j P_j^{(x_j)}\}_{x \in \{\pm\}^{[m]}}$ (i.e., each coefficient θ_i corresponds to a complete assignment). In particular, if $\mathcal{L}$ is given in a factorized (partial-monomial) form, it must first be rewritten into this full expression by inserting identities $e_0 = P_j^{(+)} + P_j^{(-)}$ for the missing variables and collecting identical Π_x terms before applying M_K.

Operationally, M_K executes three steps:

1. **Projection:** apply π_K to contract all $P_j^{(\pm)}$ with $j \in K$ to the identity e_0, geometrically eliminating those variables;
2. **Aggregation:** for each resulting basis element Π', combine its log-coefficients through $\boxplus = \log\big(e^a + e^b\big)$ in the log-semiring;
3. **Reconstruction:** form the new Clifford element as the linear sum of Π' weighted by the aggregated coefficients.

Thus, marginalization in probability space—a summation over hidden variables— is expressed in the Clifford algebra as a contraction of projectors together with a log-semiring aggregation of the corresponding coefficients. This formulation unifies the algebraic structure of the toric model with the probabilistic operation of marginalization within a single Clifford–log-semiring framework. Importantly, M_K preserves the algebraic order of the retained variables: the resulting element remains within the ordered Clifford algebra, whereas only the scalar coefficients are aggregated in the commutative log-semiring. The structural basis, built from non-commuting projectors, retains the geometric signature and orientation of the original projection sequence. Hence, M_K does not merely "sum out" variables, but performs an *ordered contraction*—a process that records the sequence of geometric reductions leading to the marginal subspace.

Example: Binary 2×2 case under the marginalization operator Let $X_1, X_2 \in \{\pm 1\}$ and consider the Clifford–log-semiring representation $\mathcal{L} = \sum_i \theta_i \Pi_i$. We compare two structural configurations of the design matrix and observe how the marginalization operator $M_{\{1\}}$ behaves differently.

(a) Independent structure. The independent configuration corresponds to the Clifford basis

$$\{\, P_1^{(+)},\, P_1^{(-)},\, P_2^{(+)},\, P_2^{(-)} \,\}, \tag{3.106}$$

so that

$$\mathcal{L}_{\text{indep}} = \theta_1 P_1^{(+)} + \theta_2 P_1^{(-)} + \theta_3 P_2^{(+)} + \theta_4 P_2^{(-)}. \tag{3.107}$$

To apply $M_{\{1\}}$ in the sense of Definition 3.6.2, we first rewrite $\mathcal{L}_{\text{indep}}$ in the full-state basis $\{P_1^{(\pm)} P_2^{(\pm)}\}$ by inserting $e_0 = P_2^{(+)} + P_2^{(-)}$ (and $e_0 = P_1^{(+)} + P_1^{(-)}$) and collecting identical terms:

$$\begin{aligned}
\mathcal{L}_{\text{indep}} &= \theta_1 P_1^{(+)}(P_2^{(+)} + P_2^{(-)}) + \theta_2 P_1^{(-)}(P_2^{(+)} + P_2^{(-)}) \\
&\quad + \theta_3 (P_1^{(+)} + P_1^{(-)}) P_2^{(+)} + \theta_4 (P_1^{(+)} + P_1^{(-)}) P_2^{(-)} \\
&= (\theta_1 + \theta_3)\, P_1^{(+)} P_2^{(+)} + (\theta_1 + \theta_4)\, P_1^{(+)} P_2^{(-)} + (\theta_2 + \theta_3)\, P_1^{(-)} P_2^{(+)} + (\theta_2 + \theta_4)\, P_1^{(-)} P_2^{(-)}.
\end{aligned} \tag{3.108}$$

Marginalizing X_1 (i.e., applying $\pi_{\{1\}} : P_1^{(\pm)} \mapsto e_0$ and aggregating coefficients by $\boxplus$ for each fixed $P_2^{(\pm)}$) yields

$$\begin{aligned}
M_{\{1\}}(\mathcal{L}_{\text{indep}}) &= \big((\theta_1 + \theta_3) \boxplus (\theta_2 + \theta_3)\big)\, P_2^{(+)} + \big((\theta_1 + \theta_4) \boxplus (\theta_2 + \theta_4)\big)\, P_2^{(-)} \\
&= \big((\theta_1 \boxplus \theta_2) + \theta_3\big)\, P_2^{(+)} + \big((\theta_1 \boxplus \theta_2) + \theta_4\big)\, P_2^{(-)}.
\end{aligned} \tag{3.109}$$

Here, the term $(\theta_1 \boxplus \theta_2) = \log\big(e^{\theta_1} + e^{\theta_2}\big)$, if already normalized, represents the log-sum-exp aggregation of X_1's two states and equals to $\log 1 = 0$,

$$\mathcal{L}_{\text{indep}} = \theta_3\, P_2^{(+)} + \theta_4\, P_2^{(-)}. \tag{3.110}$$

Since P_1 and P_2 were separable, the marginalization simply removes the P_1 component without introducing any interaction between $P_2^{(+)}$ and $P_2^{(-)}$. The resulting model remains factorized—X_1 and X_2 are independent.

(b) Dependent (joint) structure. For a joint configuration, the design matrix rows correspond to the products

$$\{\, P_1^{(+)} P_2^{(+)},\; P_1^{(+)} P_2^{(-)},\; P_1^{(-)} P_2^{(+)},\; P_1^{(-)} P_2^{(-)} \,\},$$

giving

$$\mathcal{L}_{\text{joint}} = \theta_1 P_1^{(+)} P_2^{(+)} + \theta_2 P_1^{(+)} P_2^{(-)} + \theta_3 P_1^{(-)} P_2^{(+)} + \theta_4 P_1^{(-)} P_2^{(-)}.$$

Marginalizing X_1 ($K = \{1\}$) now yields

$$M_{\{1\}}(\mathcal{L}_{\text{joint}}) = (\theta_1 \boxplus \theta_3)\, P_2^{(+)} + (\theta_2 \boxplus \theta_4)\, P_2^{(-)}.$$

Unlike the independent case, the coefficients of $P_2^{(+)}$ and $P_2^{(-)}$ are aggregated across different states of X_1. The resulting element is no longer a simple sum of $P_2^{(+)}$ and $P_2^{(-)}$ terms with

independent weights: the log-sum-exp couples the contributions of θ_1, θ_3 and θ_2, θ_4, reflecting statistical dependence between X_1 and X_2. In the probability domain, this corresponds to

$$p_{X_2=+} \propto e^{\theta_1} + e^{\theta_3}, \qquad p_{X_2=-} \propto e^{\theta_2} + e^{\theta_4},$$

which coincides with the classical marginal but retains the ordered Clifford structure encoding the original dependence pattern. *Conceptually, this means that after marginalizing X_1, the parameters originally attached to $P_1^{(\pm)}$ still influence the resulting coefficients of $P_2^{(\pm)}$ through the LSE aggregation.* Consequently, the marginal probabilities of X_2 inherit mixed contributions from both variables, making the interaction between X_1 and X_2 explicitly visible in the Clifford–log-semiring representation. This highlights that the Clifford–log-semiring formalism distinguishes independence and dependence directly at the algebraic level.

3.6.3 Structure preservation through marginalization

Equipping the Clifford algebra with the log-semiring structure and the marginalization operator completes the formulation of our probabilistic language model. At this final stage, we demonstrate how this unified framework allows us to describe—both algebraically and geometrically—how invariant structures are preserved or transformed through probabilistic operations. As a concrete and illuminating example, we examine a *conditional independence model* in the binary $2 \times 2 \times 2$ case. This model can be viewed as the concatenation of two rank-one 2×2 slices, corresponding to a variety containing two Segre subvarieties. Under marginalization, these slices collapse into a mixture of two rank-one components, which in general form a point on the *secant variety* of the original Segre manifold rather than a rank-one point. However, under a specific algebraic constraint—the *co-linear condition for rank-one preservation*—the marginal distribution remains rank-one. Our Clifford–log-semiring formulation enables this condition to be derived and solved directly within the same algebraic system, offering an elegant and transparent view of structural invariance. Although the same result could be obtained by conventional linear-algebraic methods, the present formulation highlights the expressive and unifying power of the log-semiring–augmented Clifford algebra. Finally, we note that marginalization serves as a fundamental mechanism for efficient structural learning, and its relationship to pointwise mutual information (PMI) in co-occurrence matrices will be further explored in the next chapter.

Rank-one preservation under marginalization (co-linear condition) We revisit the conditional-independence form of the Clifford logit in Eq. 3.97:

$$\begin{aligned}\mathcal{L}_{CI} = P_1^{(+)}\big(\theta_1 P_2^{(+)} + \theta_2 P_2^{(-)} + \theta_5 P_3^{(+)} + \theta_6 P_3^{(-)}\big) \\ + P_1^{(-)}\big(\theta_3 P_2^{(+)} + \theta_4 P_2^{(-)} + \theta_7 P_3^{(+)} + \theta_8 P_3^{(-)}\big).\end{aligned} \tag{3.111}$$

Each bracketed term corresponds to one slice $X_1 = \pm$ of the $2 \times 2 \times 2$ conditional-independence model. Within each slice, the logit decomposes additively into independent contributions of X_2 and X_3, hence each slice is rank-one (a Segre component) in the probability space.

Applying the marginalization operator. We now marginalize X_1 by applying $M_{\{1\}}$ to Eq. (3.111). Since Definition 3.6.2 applies to a full-state expansion, we first rewrite $\mathcal{L}_{CI}$ on the idempotent basis $\{P_1^{(\pm)}P_2^{(\pm)}P_3^{(\pm)}\}$ by inserting identities $e_0 = P_3^{(+)} + P_3^{(-)}$ (for the P_2-terms) and $e_0 = P_2^{(+)} + P_2^{(-)}$ (for the P_3-terms), and then collecting identical monomials:

$$\begin{aligned}
\mathcal{L}_{CI} &= \theta_1 P_1^{(+)}P_2^{(+)}(P_3^{(+)} + P_3^{(-)}) + \theta_2 P_1^{(+)}P_2^{(-)}(P_3^{(+)} + P_3^{(-)}) \\
&\quad + \theta_5 P_1^{(+)}(P_2^{(+)} + P_2^{(-)})P_3^{(+)} + \theta_6 P_1^{(+)}(P_2^{(+)} + P_2^{(-)})P_3^{(-)} \\
&\quad + \theta_3 P_1^{(-)}P_2^{(+)}(P_3^{(+)} + P_3^{(-)}) + \theta_4 P_1^{(-)}P_2^{(-)}(P_3^{(+)} + P_3^{(-)}) \\
&\quad + \theta_7 P_1^{(-)}(P_2^{(+)} + P_2^{(-)})P_3^{(+)} + \theta_8 P_1^{(-)}(P_2^{(+)} + P_2^{(-)})P_3^{(-)} \\
&= (\theta_1 + \theta_5)\, P_1^{(+)}P_2^{(+)}P_3^{(+)} + (\theta_1 + \theta_6)\, P_1^{(+)}P_2^{(+)}P_3^{(-)} \\
&\quad + (\theta_2 + \theta_5)\, P_1^{(+)}P_2^{(-)}P_3^{(+)} + (\theta_2 + \theta_6)\, P_1^{(+)}P_2^{(-)}P_3^{(-)} \\
&\quad + (\theta_3 + \theta_7)\, P_1^{(-)}P_2^{(+)}P_3^{(+)} + (\theta_3 + \theta_8)\, P_1^{(-)}P_2^{(+)}P_3^{(-)} \\
&\quad + (\theta_4 + \theta_7)\, P_1^{(-)}P_2^{(-)}P_3^{(+)} + (\theta_4 + \theta_8)\, P_1^{(-)}P_2^{(-)}P_3^{(-)}.
\end{aligned} \tag{3.112}$$

Applying $M_{\{1\}}$ now contracts $P_1^{(\pm)} \mapsto e_0$ and performs log-sum-exp aggregation over the two coefficients that share the same projected structure $P_2^{(\pm)}P_3^{(\pm)}$:

$$\begin{aligned}
M_{\{1\}}(\mathcal{L}_{CI}) &= \big((\theta_1 + \theta_5) \boxplus (\theta_3 + \theta_7)\big)\, P_2^{(+)}P_3^{(+)} \\
&\quad + \big((\theta_1 + \theta_6) \boxplus (\theta_3 + \theta_8)\big)\, P_2^{(+)}P_3^{(-)} \\
&\quad + \big((\theta_2 + \theta_5) \boxplus (\theta_4 + \theta_7)\big)\, P_2^{(-)}P_3^{(+)} \\
&\quad + \big((\theta_2 + \theta_6) \boxplus (\theta_4 + \theta_8)\big)\, P_2^{(-)}P_3^{(-)}.
\end{aligned} \tag{3.113}$$

The resulting element lives on the ordered full-state basis of the retained variables $\{P_2^{(\pm)}P_3^{(\pm)}\}$, making explicit that marginalization mixes the two X_1-slices through the $\boxplus$ aggregation of their cellwise log-weights.

Structurally, the marginalized element is expressed on the *full-state* basis $\{P_2^{(\pm)}P_3^{(\pm)}\}$, so cross products $P_2^{(\cdot)}P_3^{(\cdot)}$ appear explicitly. Each coefficient is a nonlinear log-sum-exp mixture of the two X_1-slice cellwise log-weights, e.g., $\big((\theta_1 + \theta_5) \boxplus (\theta_3 + \theta_7)\big)$ for $P_2^{(+)}P_3^{(+)}$. Consequently, marginalization generally breaks the rank-one (Segre) structure: in the probability domain it forms a sum of two rank-one slices, which typically lies on the secant variety (rank-two) unless the parameters satisfy the co-linearity (equal-gap) condition that collapses the secant back onto the Segre locus as we see next.

Why the mixture (secant) is not rank-one. After marginalizing X_1, the (X_2, X_3) joint distribution becomes a mixture of two rank-one slices, corresponding respectively to $X_1 = +$ and $X_1 = -$:

$$p(x_2, x_3) \propto e^{\theta_1+\theta_5} + e^{\theta_3+\theta_7}, \quad e^{\theta_1+\theta_6} + e^{\theta_3+\theta_8}, \quad e^{\theta_2+\theta_5} + e^{\theta_4+\theta_7}, \quad e^{\theta_2+\theta_6} + e^{\theta_4+\theta_8}.$$

These four terms form the 2×2 marginal table on (X_2, X_3) after summing out X_1. Although each slice ($X_1 = +$ or $X_1 = -$) was individually rank-one, their mixture is generally not rank-one. The reason is that the sum of two exponential products cannot, in general, be written as a single multiplicative factorization:

$$e^{a+b} + e^{a'+b'} \neq (e^a + e^{a'})(e^b + e^{b'}).$$

Equivalently,

$$e^{a+b}\left(1 + e^{(a'-a)+(b'-b)}\right)$$

cannot be factorized into a product of one function of X_2 and another of X_3 in genral, because

$$1 + e^{(a'-a)+(b'-b)} \neq (1 + e^{(a'-a)})(1 + e^{(b'-b)}).$$

This *non-separability of the additive exponential mixture* is what breaks the rank-one structure. The geometric (Clifford) layer still represents a direct sum over $\{P_2^{(\pm)}, P_3^{(\pm)}\}$, but the non-linear coupling introduced by the log-sum-exp operation in the coefficient layer destroys multiplicative separability.

When rank-one is restored: co-linearity along one axis (OR condition). Rank-one can be recovered if the two slices ($X_1 = +$ and $X_1 = -$) are *co-linear along either one axis*—that is, if the parameter differences between the two slices are constant (equal-gap) along X_2 or along X_3.

Write the two conditional slices ($X_1 = \pm$) in the *probability* domain as

$$Pr(X_2, X_3 \mid X_1 = 0) = \begin{bmatrix} e^{\theta_1+\theta_5} & e^{\theta_1+\theta_6} \\ e^{\theta_2+\theta_5} & e^{\theta_2+\theta_6} \end{bmatrix} = \underbrace{\begin{bmatrix} e^{\theta_1} \\ e^{\theta_2} \end{bmatrix}}_{\mathbf{u}^{(+)}} \underbrace{\begin{bmatrix} e^{\theta_5} & e^{\theta_6} \end{bmatrix}}_{\mathbf{v}^{(+)\top}} \tag{3.114}$$

$$Pr(X_2, X_3 \mid X_1 = 1) = \begin{bmatrix} e^{\theta_3+\theta_7} & e^{\theta_3+\theta_8} \\ e^{\theta_4+\theta_7} & e^{\theta_4+\theta_8} \end{bmatrix} = \underbrace{\begin{bmatrix} e^{\theta_3} \\ e^{\theta_4} \end{bmatrix}}_{\mathbf{u}^{(-)}} \underbrace{\begin{bmatrix} e^{\theta_7} & e^{\theta_8} \end{bmatrix}}_{\mathbf{v}^{(-)\top}}. \tag{3.115}$$

Thus each slice is rank-one: $P^{(+)} = \mathbf{u}^{(+)}\mathbf{v}^{(+)\top}$, $P^{(-)} = \mathbf{u}^{(-)}\mathbf{v}^{(-)\top}$. Co-linearity of the *row* factors

(the X_2-axis) means $\mathbf{u}^{(-)} = \alpha\, \mathbf{u}^{(+)}$ for some $\alpha > 0$, i.e.

$$\frac{e^{\theta_3}}{e^{\theta_1}} = \frac{e^{\theta_4}}{e^{\theta_2}} \iff \theta_3 - \theta_1 = \theta_4 - \theta_2 \quad (X_2\text{-axis equal-gap}). \tag{3.116}$$

Similarly, co-linearity of the *column* factors (the X_3-axis) means $\mathbf{v}^{(-)} = \beta\, \mathbf{v}^{(+)}$ for some $\beta > 0$, i.e.

$$\frac{e^{\theta_7}}{e^{\theta_5}} = \frac{e^{\theta_8}}{e^{\theta_6}} \iff \theta_7 - \theta_5 = \theta_8 - \theta_6 \quad (X_3\text{-axis equal-gap}). \tag{3.117}$$

The *axis meaning* is therefore precise: $u \parallel u'$ aligns the two slices *along* X_2 (row direction), while $v \parallel v'$ aligns them *along* X_3 (column direction). Under either alignment (the OR condition), the mixture after marginalization factorizes (the LSE contributions share a common additive gap and cancel in the rank-one test), so the (X_2, X_3) marginal recovers rank-one.

Geometric interpretation. Geometrically, each conditional slice ($X_1 = +$ or $X_1 = -$) is a point on the Segre variety, representing a rank-one structure. Marginalization takes the union of these two points and forms their *secant*—a mixture that typically lies off the Segre variety. When the parameters satisfy the co-linearity (equal-gap) condition along one axis, the two Segre points align on a single coordinate line, and the secant point collapses back onto the Segre locus. Thus, the *rank-one restoration* arises not from a change in the Clifford structure but from the cancellation of nonlinearities in the log-semiring layer.

In sum, marginalization transforms two independent (rank-one) slices into their mixture, which generally lies on the secant variety and breaks independence. However, if the slices are co-linear along either the X_2 or the X_3 axis—that is, if one of the equal-gap conditions above holds—then the nonlinear LSE aggregation becomes separable, and the marginal distribution recovers rank-one. The Clifford–log-semiring framework makes this interplay explicit: the Clifford layer preserves structural order, while the log-semiring layer governs how parameter alignment restores multiplicative factorization.

Conjecture. The results obtained so far suggest a deeper interpretation of symmetry in language. We conjecture that linguistic symmetry manifests itself in the *copy structure* of the design matrix—rows such as $[1, 1, \ldots]$ that represent repeated or mirrored configurations across different contexts. In the probabilistic tensor representation, these copy patterns correspond to rank-one structures, reflecting perfect alignment among components. However, as marginalization proceeds, this rank-one alignment becomes less visible due to the nonlinearity introduced by the log-sum-exp operation. Through this process, previously implicit combinations of factors become explicit at each cell, and marginalization mixes the corresponding cellwise log-weights by the log-sum-exp operation. As a consequence, even if each conditional slice is rank-one, the marginal typically entangles the remaining variables and lies off the Segre locus (i.e., exhibits interaction), unless a special parameter alignment holds. In some cases, the rank-one structure can persist—specifically when the

parameters of the marginalized slices are *co-linear*, preserving alignment along at least one axis. More generally, when co-linearity is broken, the marginal becomes a secant-type mixture of rank-one components: the entanglement is not arbitrary but arises from the specific LSE-coupled coefficient pattern induced by summing out.

From a tropical perspective, where log-sum-exp degenerates into the max operator, marginalization selects the dominant slice in each configuration. If a single slice consistently dominates across all configurations, the distribution effectively preserves rank-one despite the marginalization. This conjecture implies that linguistic regularities—symmetries embedded in the combinatorial structure of language—may remain algebraically intact but perceptually obscured by nonlinear aggregation in the probability space. What appears as stochastic variability or weak dependence in observed word distributions might, in fact, be the *tropical shadow* of an underlying algebraic symmetry— a hidden rank-one geometry that persists beneath the surface of linguistic variation.

3.7 Irreducibility of Minimum Invariant Component (MIC)

Section objective In the preceding sections, we have defined the Minimum Invariant Component (MIC)as the fundamental building block underlying invariant structures in probabilistic models. The objective of this final section is to provide a mathematical rationale for treating MIC—characterized as the most granular form of local Context-Specific Independence (CSI)—as a significant and irreducible object that constitutes the minimal unit of invariance across algebraic, geometric, and probabilistic domains.

To achieve this, we draw upon results from *Representation Theory*, a discipline situated at the intersection of group theory and linear algebra, which offers a systematic framework to characterize irreducibility. Within this framework, we reinterpret the Clifford algebra $Cl_{n,0}$ as a *twisted group ring* $\mathbb{R}_\omega[(\mathbb{Z}_2)^n]$, where the twisting factor ω enforces the anticommutation relations that define the Clifford structure.

Moreover, we employ the fact that the Walsh transformation constitutes a discrete Fourier transform on the abelian group $(\mathbb{Z}_2)^n$, thus providing a concrete instance of *harmonic analysis on finite groups*. This harmonic viewpoint allows us to decompose probability models into orthogonal invariant components, clarifying how minimal invariance manifests as rank-one (Segre) structures. Such harmonic representations also form the mathematical foundation for analyzing inductive biases in modern geometric deep neural networks, where invariance and equivariance are realized as constraints on learned representations.

Walsh basis as an irreducible representation (irrep) To establish the minimality of the MIC, we first identify the "atoms" of symmetry within the group algebra. Let us consider the n-dimensional elementary abelian 2-group $G = (\mathbb{Z}_2)^n$. For each $y \in G$, the Walsh function $\chi_y(x) = (-1)^{x \cdot y}$ defines a *character* of G, which corresponds to a one-dimensional irreducible representation (irrep).

In representation theory, by Maschke's theorem [45], any finite group representation

over $\mathbb{C}$ decomposes into a direct sum of irreducible ones. Since G is abelian, a fundamental result dictates that all its irreps are necessarily one-dimensional. This implies that each Walsh function χ_y spans an indivisible invariant subspace that cannot be further decomposed. Consequently, the Walsh basis $\{\chi_y\}_{y \in G}$ provides the most granular spectral decomposition of the function space $\mathbb{C}[G]$:

$$\mathbb{C}[G] \cong \bigoplus_{y \in G} \mathbb{C}\chi_y.$$

This isomorphism signifies that any complex-valued function on the binary space G can be uniquely decomposed into a linear combination of these one-dimensional invariant subspaces. From the perspective of MIC, this decomposition proves that the Walsh functions $\{\chi_y\}$ are the most primitive constituents—or atomic components—of the model's structure. It ensures that any complex invariant property is ultimately constructed from these irreducible units. By realizing both a complete system of irreps and an orthonormal basis for harmonic analysis, the Walsh basis serves as the definitive collection of irreducible building blocks for all higher-order structures, including the MIC.

MIC as a minimal combination of irreducible projectors The minimality of the MIC is a direct inheritance from the atomicity of the Walsh basis. In the group algebra $\mathbb{C}[G]$, each character χ_y defines a minimal idempotent projector P_y. Since these characters are one-dimensional irreps, the corresponding projectors P_y cannot be further decomposed into smaller invariant subspaces.

An MIC is thus defined as the minimal direct sum of these atomic projectors that recovers a specific invariant structure, namely a local context-specific independence (CSI). While more complex invariant structures can exist—such as those defined by higher-degree binomials—the MIC represents the most granular level at which these symmetries manifest.

From a geometric perspective, this corresponds to the fact that the simplest MIC defines an irreducible Segre variety. The vanishing of the 2×2 minor, $p_{00}p_{11} - p_{01}p_{10} = 0$, represents a prime ideal, ensuring that the resulting variety cannot be partitioned into simpler independent components. Even in higher-dimensional systems where the decomposition is less trivial, the MIC serves as the irreducible building block, providing the smallest possible support in the group algebra that maintains the model's structural invariance.

Important conjecture Based on the correspondence established among representation theory, algebraic geometry, and probability, we state a conjecture concerning the universality of the MIC as the fundamental unit of invariance. We conjecture that *any invariant structure in a probabilistic model can be expressed as a structured composition of MICs*. That is, every higher-order invariant—whether representing multi-variable independence or partial exchangeability—is ultimately decomposable into a set of local rank-one components,

characterized by vanishing 2×2 minors.

The significance of this conjecture lies in its potential to radically simplify the discovery of invariance in empirical data. If we can identify the precise mathematical requirements—such as the *unimodularity* of the design matrix—under which this decomposition is guaranteed, the search for complex invariants can be reduced to the systematic detection of MICs (vanishing 2×2 minors). This would transform a high-dimensional combinatorial problem into a granular, localized algebraic task.

Characterizing these exact requirements, whether they involve unimodularity or specific lattice properties of the model's toric ideal, remains a subtle question for future research. However, clarifying the conditions under which invariance is restricted to MICs offers a powerful strategic advantage: it provides a definitive "algebraic footprin" for researchers to target, ensuring that the fundamental building blocks of model structure are both identifiable and irreducible.

Implication The results developed in this chapter suggest that conventional statistical independence is a specific, elementary manifestation of a broader principle: **invariance**. While independence corresponds to the simplest rank-one invariant structure, more intricate dependencies—such as the No-Three-Way Interaction model defined by quartic binomials—can be reinterpreted as structured compositions or intersections of MICs. This perspective generalizes the notion of independence into a hierarchy of invariant regularities, where the MIC serves as the irreducible building block.

Consequently, patterns of MICs provide a new algebraic taxonomy for invariant structures. Each distinct combination of MICs defines a unique equivalence class of probability models, transcending traditional categorical distinctions between *independent, conditionally independent*, and *correlated.* Instead, we obtain a unified description of invariance as a fundamental algebraic and geometric property of probability distributions.

This unification crystallizes the irreducible correspondence across three levels:

1. **Geometric:** Invariance manifests as the structure of Segre varieties describing rank-one and mixture components in projective space.

2. **Algebraic:** Invariance corresponds to the vanishing of minors and the decomposition of polynomial ideals into irreducible prime components.

3. **Probabilistic:** Invariance generalizes multiplicative factorization and locally independent relations into a unified framework of structural stability.

These three perspectives are not merely analogous but are mutually equivalent views of the same underlying mechanism: the Minimum Invariant Component.

Chapter 4

Learning: Tensor Analysis for Invariant Structure

4.1 Overview of the Chapter

This chapter translates the theoretical framework developed in Chapter 3 into concrete learning procedures that identify invariant probabilistic structures from data. The goal is to operationalize the compositional probability model by discovering *Minimum Invariant Constraints* (MICs)—local rank-one components that remain parameter invariant. Learning here means recovering these invariant substructures directly from empirical probability tensors, thus turning the abstract algebraic theory of invariance into a computational procedure.

In the previous chapter, we introduced MICs as the fundamental building block for modeling high-dimensional discrete probability distributions. An MIC manifests as a local rank-one structure—specifically, a 2×2 cell of the joint probability tensor whose determinant (or 2-minor) vanishes. Complex probabilistic structures containing symmetries can thus be represented not by a single global rank-one tensor, but as intricate combinations of these local rank-one submatrices.

Two complementary paradigms are proposed for this task. The first, a *geometric* or *spatial* approach, inspects the tensor locally by detecting vanishing 2×2 minors that signify MICs as a building block of invariant structures. This method represents the ground-truth geometry of invariance as formulated in toric model of algebraic statistic: each vanishing minor corresponds to a point on the Segre variety, or equivalently, to a rank-one substructure inside the probability tensor. While conceptually transparent, exhaustive minor inspection suffers from exponential growth in computational cost. To address this limitation, the chapter develops heuristic extensions based on marginalization based on *a lattice of variable moments* , which exploit the vertical and horizontal propagation of vanishing minors to reduce the search space.

The second, *harmonic* or *spectral* approach, views the same tensor through the lens of frequency analysis. By transforming probability tensors into the *Walsh basis*, complex

local dependencies become linear relations among orthogonal spectral components. In this representation, a vanishing interaction corresponds simply to a zero coefficient in the spectral spectrum. The Walsh transform offers substantial computational advantages: it captures global invariant patterns that may be invisible in the spatial domain and achieves a near-linear complexity $O(N \log N)$, in contrast to the combinatorial explosion of minor enumeration. The spectral view also reveals hidden algebraic relations—*syzygies*—linking different invariants, thereby providing a unified description of independence and symmetry within the same algebraic space.

Comparative experiments illustrate the complementarity of these two paradigms. Spatial inspection provides fine-grained interpretability and direct access to local structures, whereas spectral analysis offers scalability and the ability to detect global invariants efficiently. Both methods uncover recurring MIC patterns that correspond to linguistic compositionality, demonstrating that invariance can be empirically detected within real probability tensors. In combination, they establish a bidirectional correspondence between geometry and algebra: the geometric approach defines what independence is (a vanishing condition on a toric variety), while the harmonic approach shows how such structure can be found efficiently (as linear combinations of orthogonal basis).

The learning algorithms introduced here thus give operational meaning to the theoretical model of Chapter 3. They transform the algebraic characterization of compositionality into procedures capable of extracting invariant structures from data, forming the empirical and algorithmic basis for the analyses of language distribution data. In this sense, the present chapter serves as the methodological core of the dissertation—connecting the abstract invariance principles of algebraic geometry with the concrete mechanisms of statistical learning.

4.2 Spatial Approach with 2×2 Minor

4.2.1 Vanishing 2×2 Minor

Setting Following the theoretical formulation in Chapter 3, we define the *Minimum Invariant Component* (MIC) via an algebraic condition given by the vanishing of a 2×2 minor in a probability tensor.

Let $X = \{X_1, X_2, \ldots, X_m\}$ be a set of m discrete random variables with finite state space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m$ where $|\mathcal{X}_k| = n_k$. Let $\boldsymbol{p} \in \Delta^{N-1}$, where $\Delta^{N-1} := \left\{\boldsymbol{p} \in \mathbb{R}^N_{\geq 0} \mid \sum_{i=1}^N p_i = 1, \right\}$, be a probability vector of the joint state $x = (x_1, \ldots, x_m) \in \mathcal{X}$ where $N = \prod_{k=1}^m n_k$. Entries of a vector $\boldsymbol{p}$ are lexicographically ordered. For example, in the case of 2 binary variables ($m = 2$), $\boldsymbol{p} = (p_{00}, p_{01}, p_{10}, p_{11})^\top$. We note that the lexicographically ordered tuple of assignments such as $00, 01, 10, 11$ in this example shall be seen as coordinates in a vector space. We denote a variable X_k as its index k interchangeably from time to time.

An MIC represents the minimal direct sum of two irreducible representations whose

equivalence is captured by the polynomial constraint

$$p_{ijk}p_{i'j'k} - p_{ij'k}p_{i'jk} = 0, \tag{4.1}$$

for indices $i, i' \in \mathcal{X}_I$, $j, j' \in \mathcal{X}_J$, and $k \in \mathcal{X}_K$, where $I, J, K \subseteq [m]$ are disjoint subsets of m random variables X_r $(r = 1, \ldots, m)$ with $I \sqcup J \sqcup K = [m]$.

This equation specifies a *rank-1 substructure*, which geometrically corresponds to a point on the Segre variety. It provides a direct, local notion of invariance as a generalized notion of independence within the tensor representation of a probability model.

By identifying where such minors vanish in an empirical or theoretical probability tensor, one can detect subsets of variables exhibiting invariant relations. Each vanishing minor encodes a local constraint, and the collection of these constraints determines an invariant structure of the model. Some of these structures coincide with conventional independence relations (e.g., marginal or conditional independence), while others generalize them within the broader algebraic framework developed in this thesis.

Numerical Example ($2 \times 2 \times 2$ Model) To illustrate the geometric definition of MIC via vanishing minors, we consider three binary random variables that induce a joint distribution on a $2 \times 2 \times 2$ cube. This setting shows how different combinations of MICs characterize distinct invariant structures in probability distributions.

Let X_i $(i = 1, 2, 3)$ be binary with values in $\{0, 1\}$, yielding $2^3 = 8$ joint states (as introduced in Chapter 3). In that context, we considered four design matrices, denoted A_1, A_2, A_3, and A_4 as defined in Section 3, corresponding (up to variable permutation) to *complete independence*, *joint independence*, *conditional independence*, and *context-specific independence*, respectively. We further introduce a fifth matrix A_5 representing the *no–three-way interaction* model:

$$A_5 := \begin{bmatrix} E_2 \otimes E_2 \otimes \mathbf{1}_2^\top \\ E_2 \otimes \mathbf{1}_2^\top \otimes E_2 \\ \mathbf{1}_2^\top \otimes E_2 \otimes E_2 \end{bmatrix}, \tag{4.2}$$

and the saturated model without any kind of independence:

$$A_6 := E_2 \otimes E_2 \otimes E_2 \tag{4.3}$$

Examining the kernel of each design matrix yields a system of 13 binomial relations—12 quadratics and 1 quartic—that together describe the vanishing ideals of the corresponding

models:

$$
\begin{aligned}
i_1 &= p_{000}p_{011} - p_{001}p_{010}, \quad & i_2 &= p_{100}p_{111} - p_{101}p_{110}, \\
i_3 &= p_{000}p_{101} - p_{001}p_{100}, \quad & i_4 &= p_{010}p_{111} - p_{011}p_{110}, \\
i_5 &= p_{000}p_{110} - p_{010}p_{100}, \quad & i_6 &= p_{001}p_{111} - p_{011}p_{101}, \\
i_7 &= p_{000}p_{111} - p_{001}p_{110}, \quad & i_8 &= p_{000}p_{111} - p_{010}p_{101}, \\
i_9 &= p_{000}p_{111} - p_{100}p_{011}, \quad & i_{10} &= p_{001}p_{110} - p_{010}p_{101}, \\
i_{11} &= p_{001}p_{110} - p_{100}p_{011}, \quad & i_{12} &= p_{010}p_{101} - p_{100}p_{011}, \\
i_{13} &= p_{000}p_{011}p_{101}p_{110} - p_{001}p_{010}p_{100}p_{111}.
\end{aligned}
\tag{4.4}
$$

These 12 quadric binomials are the only permissible combinations of indices that restrict the possible constraints satisfying the index criteria given in MIC conditions (4.1) and the quartic binomial reflects the higher-ordered combination of them.

Different subsets of these binomials vanish under different design matrices; for example, all $i_1, \ldots, i_{13}$ vanish for A_1, whereas only i_1 vanishes for A_4. Each model imposes a specific subset of these relations as its defining ideal: the *no–three-way interaction* model A_5 is characterized by the single quartic relation $\{i_{13}\}$, while the *saturated model* corresponds to the empty ideal, meaning none of them vanishes. Details will be given in the next paragraph. These 13 binomials are not algebraically independent; for instance, $i_1 = i_6 = 0$ implies $i_{13} = 0$. Thus, by inspecting which of the 13 relations vanish in empirical data, one can determine the invariant structure present in a given probability tensor.

Learning Method Conventional model selection typically relies on maximum likelihood estimation combined with information criteria such as AIC or BIC[133]. While these approaches effectively penalize model complexity, they depend on explicit parameter optimization and do not directly exploit the algebraic constraints inherent in probabilistic models.

We propose an alternative vanishing binomial approach that replaces likelihood-based selection with an algebraic criterion derived from vanishing binomials, including 2×2 minors. Each probabilistic model is associated with an ideal generated by polynomial equations that vanish for all admissible parameter values of that model. Because these equations are invariant under parameter changes, the ideal serves as an algebraic constraint that uniquely characterizes the model's structural class.

Given a family of models $\{\mathcal{M}_i\}_{i=0}^{14}$, we encode the vanishing pattern of each model by a binary vector $\boldsymbol{m}_i \in \{0,1\}^{13}$ $(i = 0, \ldots, 14)$, where the k-th entry equals 1 if the binomial i_k

in Eq. (4.4) vanishes under $\mathcal{M}_i$ and 0 otherwise. The complete set of these vectors is

			i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}
$\mathcal{M}_0$(*Saturated*)	$\boldsymbol{m}_0^T$		0	0	0	0	0	0	0	0	0	0	0	0	0
$\mathcal{M}_1$(*No3Way*)	$\boldsymbol{m}_1^T$		0	0	0	0	0	0	0	0	0	0	0	0	1
$\mathcal{M}_2$(*Conditional I on* X_1)	$\boldsymbol{m}_2^T$		1	1	0	0	0	0	0	0	0	0	0	0	1
$\mathcal{M}_3$(*Conditional I on* X_2)	$\boldsymbol{m}_3^T$		0	0	1	1	0	0	0	0	0	0	0	0	1
$\mathcal{M}_4$(*Conditional I on* X_3)	$\boldsymbol{m}_4^T$		0	0	0	0	1	1	0	0	0	0	0	0	1
$\mathcal{M}_5$(*Joint I* $(X_1, X_2) - X_3$)	$\boldsymbol{m}_5^T$		1	1	1	1	0	0	1	0	0	0	0	1	1
$\mathcal{M}_6$(*Joint I* $(X_1, X_2) - X_2$)	$\boldsymbol{m}_6^T$		1	1	0	0	1	1	0	1	0	0	1	0	1
$\mathcal{M}_7$(*Joint I* $(X_2, X_3) - X_1$)	$\boldsymbol{m}_7^T$	=	0	0	1	1	1	1	0	0	1	1	0	0	1
$\mathcal{M}_8$(*Complete Independence*)	$\boldsymbol{m}_8^T$		1	1	1	1	1	1	1	1	1	1	1	1	1
$\mathcal{M}_9$(*Context-Specific* $X_1 = 0$)	$\boldsymbol{m}_9^T$		1	0	0	0	0	0	0	0	0	0	0	0	0
$\mathcal{M}_{10}$(*Context-Specific* $X_1 = 1$)	$\boldsymbol{m}_{10}^T$		0	1	0	0	0	0	0	0	0	0	0	0	0
$\mathcal{M}_{11}$(*Context-Specific* $X_2 = 0$)	$\boldsymbol{m}_{11}^T$		0	0	1	0	0	0	0	0	0	0	0	0	0
$\mathcal{M}_{12}$(*Context-Specific* $X_2 = 1$)	$\boldsymbol{m}_{12}^T$		0	0	0	1	0	0	0	0	0	0	0	0	0
$\mathcal{M}_{13}$(*Context-Specific* $X_3 = 0$)	$\boldsymbol{m}_{13}^T$		0	0	0	0	1	0	0	0	0	0	0	0	0
$\mathcal{M}_{14}$(*Context-Specific* $X_3 = 1$)	$\boldsymbol{m}_{14}^T$		0	0	0	0	0	1	0	0	0	0	0	0	0

The model $\mathcal{M}_8$—where all 13 relations vanish—corresponds to complete independence denoted as $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp X_3$. The models $\mathcal{M}_5$, $\mathcal{M}_6$, and $\mathcal{M}_7$ correspond to joint independence ($(X_1, X_2) \perp\!\!\!\perp X_3$, $(X_1, X_3) \perp\!\!\!\perp X_2$, and $(X_2, X_3) \perp\!\!\!\perp X_1$, respectively), while $\mathcal{M}_2$, $\mathcal{M}_3$, and $\mathcal{M}_4$ correspond to conditional independence ($X_2 \perp\!\!\!\perp X_3 \mid X_1$, $X_1 \perp\!\!\!\perp X_3 \mid X_2$, and $X_1 \perp\!\!\!\perp X_2 \mid X_3$, respectively). $\mathcal{M}_2$ corresponds to the no-three-way interaction model, where all pairwise interactions exist but the three-way interaction vanishes. $\mathcal{M}_0$—associated with the empty ideal ($\boldsymbol{m}_0 = \mathbf{0}$)—corresponds to the saturated model. The remaining models $\mathcal{M}_i$ for $i = 9, \ldots, 14$ correspond to context-specific independence (CSI) such as $X_2 \perp\!\!\!\perp X_3 \mid X_1 = 0$. Hence, $\boldsymbol{m}_i$ functions as a model-specific invariant signature.

Given empirical frequency data, we compute $i_1, \ldots, i_{13}$ from Eq. (4.4) to form $\boldsymbol{i} = (i_1, \ldots, i_{13}) \in \mathbb{R}^{13}$. Model selection is then posed as an algebraic matching problem: define a distance $d(\boldsymbol{i}, \boldsymbol{m}_k)$ between the observed feature vector and each model's invariant pattern, and select

$$k^* = \arg \min_{k \in \{0,\ldots,14\}} d(\boldsymbol{i}, \boldsymbol{m}_k).$$

This enables model identification purely from the structure of vanishing minors, yielding a parameter-free invariant criterion that complements conventional likelihood-based methods.

Challenge: Computational Explosion The principal limitation of the direct 2×2 minor approach lies in its computational complexity. The number of minors that must be evaluated grows combinatorially with the dimensionality of the probability tensor, scaling exponentially with the number of variables m. For any realistic model involving multiple interacting variables, an exhaustive inspection of all possible minors quickly becomes infeasible.

In practice, directly computing all possible minors for an $m \times n$ matrix requires evalu-

ating $\binom{m}{2}\binom{n}{2}$ determinants, resulting in a computational cost of $O(m^2n^2)$. For higher-order tensors, even those of moderate size such as $10 \times 10 \times 10 \times 10$, the number of quadratic binomials approaches 200 million. Thus, even for moderate m and n, this quadratic scaling in both dimensions renders brute-force verification computationally prohibitive.

To alleviate this issue, an incremental strategy must be employed. Rather than evaluating every possible minor independently, one can exploit structural relationships among overlapping submatrices. By propagating information through the network of variables—an approach referred to here as *network propagation*— it becomes possible to infer vanishing minors indirectly, thereby reducing the number of explicit minor evaluations required. This idea lays the foundation for the more efficient spatial-domain methods discussed in the following section.

4.2.2 Advanced Spatial Methods with Moments

To mitigate the computational complexity inherent in the direct evaluation of all 2×2 minors, we introduce an advanced *divide-and-conquer* strategy based on marginalization. Instead of inspecting the raw entries of the full probability tensor, this approach prioritizes the exploration of invariant substructures by first examining lower-order variable moments. This is justified by the *marginalization–tensorization adjunction* as we will discuss soon: rank-one structures (independence) in higher dimensions are typically preserved or approximated as rank-one structures in their marginals. The marginalization-based approach also employs heuristic measures, such as generalized forms of Pointwise Mutual Information (PMI), to guide the search process and avoid exhaustive verification of all minors. These heuristics serve as practical indicators of where rank-one constraints are likely to emerge, and they form the basis for the efficient algorithmic framework developed in the subsequent section.

Variable Moments and Lattice For a subset of indices $I = \{i_1, \dots, i_r\} \subseteq [m]$ with $|I| = r$, a moment $\boldsymbol{\mu}_I$ is defined as the marginal distribution obtained by summing out variables in $[m] \setminus I$:

$$\boldsymbol{\mu}_I(x_I) = E_p[\mathbb{I}(X_I = x_I)] = \sum_{x_{[m]\setminus I}} p(x), \tag{4.5}$$

where $x_I = (x_k)_{k\in I}$ means a tuple of states for the variables contained in the set I and function as a coordinate of the moment vector. Note that the moment $\boldsymbol{\mu}_I$ can be also identified with a function whose value depends solely on the assignment of x_{I_r} and is called a r-th variable moment. We emphasize that r-th variable moments should not be confused with r-th order moments, which conventionally refer to expectations of powers of random variables (e.g., $\mathbb{E}[X^2]$ is the second-order moment). All the r-th variable moments defined here are a first ordered moment.

These moments form a lattice under inclusion (see the example in Figure 4.1). The key diagnostic for our strategy is whether a combined moment $\boldsymbol{\mu}_{IJ} := \boldsymbol{\mu}_{I\cup J}$ factorizes as

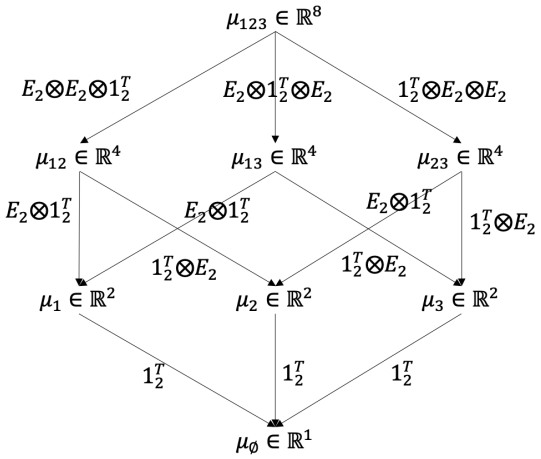


Figure 4.1: Moment lattice defined by all r-variable moments and the partial order given by inclusion.

$\boldsymbol{\mu}_{IJ} \approx \boldsymbol{\mu}_I \otimes \boldsymbol{\mu}_J$ as defined below. If this condition holds (indicating a rank-one structure), we proceed to test higher-order interactions.

A k-th variable moment is obtained from a $(k{+}1)$-th variable moment through marginalization, which acts as a linear map. For example, in case of $2 \times 2 \times 2$ model, 1st variable moments $\boldsymbol{\mu}_X, \boldsymbol{\mu}_Y \in \mathbb{R}^2$ can be derived from 2nd-variable moments $\boldsymbol{\mu}_{XY} \in \mathbb{R}^4$ in the probability of two binary variables as

$$\boldsymbol{\mu}_X = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \boldsymbol{\mu}_{XY}, \qquad \boldsymbol{\mu}_Y = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix} \boldsymbol{\mu}_{XY}, \tag{4.6}$$

where the coordinates of $\boldsymbol{\mu}_{XY}$ are assumed to be ordered lexicographically with $X \prec Y$. In general, marginalization defines a surjective linear map $\mathcal{M}\colon \boldsymbol{\mu}^{(k+1)} \longmapsto \boldsymbol{\mu}^{(k)}$. Thus, variable moments of different r form a lattice under marginalization . Each level corresponds to the set of all r-variable moments, and the partial order is given by inclusion of variable sets. Conversely, under (joint) independence between k variables and another variable, $k+1$ higher-order variable moments can be reconstructed from lower-order ones.

Let $(M_{XY})_{ij} = \Pr\big(X = i, Y = j\big)$ be the joint distribution matrix, and $\boldsymbol{\mu}_{XY} = \mathrm{vec}(M_{XY})$ as its vectorization under the $X \prec Y$ convention. Then X and Y are independent if and only if

$$\boldsymbol{\mu}_{XY} = \boldsymbol{\mu}_X \otimes \boldsymbol{\mu}_Y. \tag{4.7}$$

where $\otimes$ denotes the Kronecker product. Equivalently, $M_{XY} = \boldsymbol{\mu}_X \boldsymbol{\mu}_Y^\top$ has rank one, implying that all 2×2 minors vanish. In this case we say that the moment vector $\boldsymbol{\mu}_{XY}$ is *rank one*. More generally, if a $(k{+}1)$-variable distribution factors as a product of k-variable and 1-variable moments, then $\boldsymbol{\mu}^{(k+1)} = \boldsymbol{\mu}^{(k)} \otimes \boldsymbol{\mu}^{(1)}$.

Remark 4.2.1 (Adjunction on Moment Lattice)**.** We construct an adjunction on the moment lattice by setting categories. Let $\mathcal{U}$ be the category of moments, and let $\mathcal{U} \times \mathcal{U}$ be its

product category where morphisms are defined by equality (i.e., $(\mu_X, \mu_Y) \to (\nu_X, \nu_Y)$ exists iff $\mu_X = \nu_X$ and $\mu_Y = \nu_Y$). We define the mutual information $\delta(\mu)$ for a moment μ as the KL divergence from the joint distribution to the (outer) product of its marginals (marginal vectors):

$$\delta(\mu) := D_{\mathrm{KL}}(\mu \,\|\, \mu_X \otimes \mu_Y). \tag{4.8}$$

A morphism $\mu \to \nu$ exists in $\boldsymbol{\mathcal{U}}$ if and only if they share the same marginals and ν has less dependency than μ:

$$M(\mu) = M(\nu) \quad \text{and} \quad \delta(\mu) \geq \delta(\nu). \tag{4.9}$$

We define the marginalization functor $M : \boldsymbol{\mathcal{U}} \to \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{U}}$ and the tensorization functor $T : \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{U}} \to \boldsymbol{\mathcal{U}}$ as:

$$M(\mu_{XY}) := (\sum_Y \mu_{XY}, \sum_X \mu_{XY}), \qquad T(\nu_X, \nu_Y) := \nu_X \otimes \nu_Y, \tag{4.10}$$

for $\mu_{XY}, \nu_X, \nu_Y \in \boldsymbol{\mathcal{U}}$ and $(\sum_Y \mu_{XY}, \sum_X \mu_{XY}), (\nu_X, \nu_Y) \in \boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{U}}$

Under these definitions, M is the left adjoint and T is the right adjoint ($M \dashv T$). The adjunction is established by the following bi-directional correspondence:

$$\frac{M(\mu_{XY}) \to (\nu_X, \nu_Y)}{\mu_{XY} \to T(\nu_X, \nu_Y)} \tag{4.11}$$

The condition on the bottom holds because $T(\mu_X, \mu_Y)$ is a rank-one moment with zero mutual information ($\delta = 0$). Thus, provided the marginals match (which is the condition on the top), the morphism to the rank-one tensor always exists as the direction of maximum entropy.

Note that the category $\boldsymbol{\mathcal{U}} \times \boldsymbol{\mathcal{U}}$ is treated as a discrete category (morphisms are identities). This is a direct consequence of our definition of morphisms in $\boldsymbol{\mathcal{U}}$, which are strictly marginal-preserving. The adjunction thus operates fiber-wise, identifying the rank-one tensor as the terminal object within each fiber of fixed marginals.

Generalized PMI as a Diagnostic To practically detect these rank-one structures, we employ the *generalized Pointwise Mutual Information (PMI)*, or cumulant, defined elementwise as:

$$\kappa_{I,J} := \log \boldsymbol{\mu}_{IJ} - \log \boldsymbol{\mu}_I \, \mathbf{1}^\top - \mathbf{1} \, (\log \boldsymbol{\mu}_J)^\top, \tag{4.12}$$

where the logarithm is taken elementwise. Let d_I and d_J be the numbers of the joint states of variables in I and J, respectively, so that $\kappa_{I,J} \in \mathbb{R}^{d_I \times d_J}$. The magnitude is defined by the entry-wise maximum norm

$$\|\kappa_{I,J}\|_{\max} := \max_{1 \leq a \leq d_I,\ 1 \leq b \leq d_J} \left|(\kappa_{I,J})_{ab}\right|, \tag{4.13}$$

which serves as a tractable proxy for deviation from the Segre variety. Small values indicate that the variables in I and J are nearly independent, warranting a "deep dive" to check for conditional independence or specific vanishing minors in higher dimensions.

The validity of this marginalization-based approach rests on the observation that rank-one structures are often preserved under the summation inherent in marginalization. The following lemma provides the linear-algebraic condition for this structural persistence.

Lemma 4.2.1 (Sum of two rank-one slices)**.** Let $P^{(1)} = u_1 v_1^\top$ and $P^{(2)} = u_2 v_2^\top$ with $u_\ell \in \mathbb{R}^m_{\geq 0}$, $v_\ell \in \mathbb{R}^n_{\geq 0}$ $(\ell = 1, 2)$. The sum $M := P^{(1)} + P^{(2)}$ is rank one if and only if either $u_2 = \alpha u_1$ or $v_2 = \beta v_1$ for some scalars $\alpha, \beta \geq 0$. We call this necessary and sufficient condition *co-linear condition*

Proof sketch. If $u_2 = \alpha u_1$, then $M = u_1(v_1 + \alpha v_2)^\top$ is rank one; similarly when v_2 is colinear with v_1. If neither pair is colinear, then M has two independent columns (or rows), hence rank two. □

Conceptually, it corresponds to the same mechanism we discussed in Chapter 3, Section 3.6 (Log–Semiring): there, multiplicative independence $p(x, y) = p(x)p(y)$ was translated into additive separability $\log p(x, y) = \log p(x) + \log p(y)$ through the log–semiring. When two rank-one slices are aligned as the colinear condition in Lemma 4.2.1, the *log-sum-exp* operation,

$$\log\left(e^a + e^b\right) = \max(a, b) + \log\left(1 + e^{-|a-b|}\right), \tag{4.14}$$

reduces to exact addition when a and b are equal, and stays close to it when $|a - b|$ is small. Hence, even when slices are not perfectly colinear, the log–semiring smooths their mixture toward an additive (near rank-one) form. Geometrically, this additive closure corresponds to forming convex (linear) mixtures of several Segre components, which in algebraic geometry defines the secant variety of the Segre embedding. This insight justifies our approximation strategy: we first identify near rank-one patterns in lower-order moments and then explore higher-order structures when such evidence is found. The next paragraph provides a practical procedure for structural learning.

Hierarchical Model Selection Algorithm We implement this strategy for the $2 \times 2 \times 2$ problem as a structured decision flow (Algorithm 1). The procedure utilizes three marginal dependency moments $\hat{\boldsymbol{m}} = (\hat{m}_1, \hat{m}_2, \hat{m}_3)$. While the generalized PMI discussed above provides a statistical measure of deviation from independence, for computational efficiency, we instantiate $\hat{\boldsymbol{m}}$ using the magnitude of the 2×2 determinants (minors) of the marginal matrices: $\hat{m}_1 = |\det \mathrm{Mat}(\boldsymbol{\mu}_{X_2X_3})|$ where Mat is an operator to re-vectorize a vector to a matrix. This algebraic measure vanishes if and only if the PMI vanishes (rank-one), serving as a direct and computationally inexpensive proxy for structural learning.

The algorithm relies on three key parameters: the absolute tolerance $\varepsilon_{\text{zero}}$ (corresponding to `tol`), the relative comparison factor α_{sec} (corresponding to `alpha`), and the absolute trigger τ_{rank1} (corresponding to `tau`). Using these parameters, the classification proceeds

Algorithm 1 Concise Two-Stage Model Selection for $2 \times 2 \times 2$ Models

Input: Empirical distribution $\hat{p}$; Thresholds: `tol`, `alpha`, `tau`;
Output: Predicted model ID $\widehat{k} \in \{0, \ldots, 14\}$.

$\hat{\boldsymbol{m}} \leftarrow$ `compute_moments`$(\hat{p})$; `num_zero` $\leftarrow \sum(\hat{\boldsymbol{m}} <$ `tol`$)$ ▷ Stage 1: Compute marginal dependencies
if `num_zero` $= 3$ **then return** 8 ▷ Independence
end if
if `num_zero` $= 2$ **then return** `JIx` model ▷ Joint Independence
end if
$\hat{m}_{\min} \leftarrow \min(\hat{\boldsymbol{m}})$; `minpos` $\leftarrow \operatorname{argmin}(\hat{\boldsymbol{m}})$
if $\hat{m}_{\min} <$ `alpha` $\cdot$ `others` **then return** $2 +$ `minpos` ▷ Conditional Independence
end if
if $\hat{m}_{\min} <$ `tau` **then**
    Check corresponding 2 minors $\hat{I}$
    if $\hat{I}$ minor is $<$ `ths` **then return** `CSIx` model ▷ CSI
    end if
end if
`quartic_val` $\leftarrow$ `QuarticMinor`$(\hat{p})$
if `quartic_val` $<$ `tol` **then return** 1 ▷ No-3-Way
elsereturn 0 ▷ Saturated
end if

in two stages (Algorithm 1). **Stage 1 (Coarse Classification)** first counts the number of vanishing marginal moments to identify Independence or Joint Independence. If ambiguity remains, **Stage 2 (Fine Classification)** analyzes the minimum moment magnitude: a significant relative drop indicates Conditional Independence, while specific 2×2 minors are checked to identify Context Specific Independence (CSI); failing these, the quartic minor determines the final model.

4.3 Harmonic Analysis for Structure Discovery

4.3.1 Walsh Transformation

As introduced in Chapter 3, the invariant structures of a Toric model can be analyzed through the joint use of the Walsh–Hadamard Transform (WHT) and Clifford algebra. Once the underlying symmetries and equivalence relations in a target model are identified, they can be leveraged as inductive biases for efficient parameter learning. However, in the absence of any prior knowledge about such structures, the first step must be to *discover* them directly from data.

The Walsh–Hadamard Transform provides an irreducible representation of the function space over $\{0, 1\}^n$, enabling us to decompose a probability model into orthogonal components that correspond to interactions of different orders. This decomposition makes it possible to detect the locations and patterns where the Minimum Invariant Constraints (MICs) exist, e.g., local structures corresponding to context-specific independence (CSI).

In this section, we describe how the Walsh transform can be applied to empirical data represented as probability vectors. Through this analysis, hidden algebraic structures can be revealed without enumerating all vanishing minors explicitly. The Walsh–Hadamard Transform (WHT) provides a complete and computationally efficient ($O(N \log N)$) *frequency-domain* representation of probabilistic structure. Also, hidden local symmetries such as in context-specific independence (CSI) emerge as linear algebraic relations (syzygies) among spectral coefficients, offering a tractable alternative to the combinatorial explosion of direct minor analysis.

Method: Projection onto the Walsh Basis To analyze this structure, we project the log-probability function onto an orthogonal basis provided by the Walsh–Hadamard Transform (WHT). Formally, let $\boldsymbol{p}(x)$ be a probability distribution over joint states $x \in \{0,1\}^m$ for m binary variables, and define its log-probability vector as

$$\theta(x) = \log \boldsymbol{p}(x). \tag{4.15}$$

Applying the Walsh transform to $\theta(x)$ gives the *structural spectrum*:

$$\hat{\theta}(y) = \sum_{x \in \{0,1\}^m} \theta(x)\,(-1)^{x \cdot y}, \tag{4.16}$$

where $y \in \{0,1\}^m$ indexes the Walsh basis functions $\chi_y(x) = (-1)^{x \cdot y}$ corresponding to subsets of variable indices. Each coefficient $\hat{\theta}(y)$ represents the contribution of an interaction term, or equivalently a monomial in the Clifford basis $e_y = \prod_{i:y_i=1} e_i$, whose *interaction order* follows the definition given in Definition 3.4.1 of Chapter 3.

The structure of the model is then *read from the spectrum*. If $\hat{\theta}(y) = 0$, the corresponding y-interaction term is absent from the log-linear expansion. Conversely, nonzero coefficients indicate the existence of dependencies of that order. For instance, when all higher-grade coefficients (i.e., $|y| \geq 2$) vanish, the model reduces to a complete independent (rank-one) distribution. Furthermore, vanishing coefficients or their linear combinations in the Walsh domain correspond to *vanishing 2-minors* of the probability tensor in the original space. For example, in a $2 \times 2 \times 2$ model, a single vanishing 2-minor on the slice $X_1 = 0$ represents a context-specific independence (CSI) relation $X_2 \perp\!\!\!\perp X_3 \mid X_1 = 0$. In this model, the sum of the coefficients for e_{123} and e_{23} would be zero, implying that the global pairwise interaction is exactly canceled by the three-way interaction in the context of $X_1 = 0$. In essence, the WHT provides a direct way to detect invariant structures by examining the sparsity and algebraic relations within the spectral coefficients $\hat{\theta}$, replacing the need for exhaustive minor-based searches with a tractable, linear-algebraic analysis.

4.3.2 Model Identification using Walsh Transform

Now we describe the steps to identify an unknown model through harmonic analysis using the Walsh transform (Algorithm 2). The procedure is based on the *spectral signature*

Algorithm 2 Model Selection via Walsh Spectral Signatures (`select_by_walsh`)

Input: Empirical distribution $\hat{p}$ over $2\times2\times2$ states; Zero tolerance threshold `tol`; Signature database `PATTERNS`.
Output: Predicted model ID $\widehat{k} \in \{0, \ldots, 14\}$.

Log-Transformation
$\theta \leftarrow \log(\hat{p})$ ▷ Compute logit tensor
Walsh Projection
`w` $\leftarrow$ `compute_walsh_transform`(θ) ▷ Project onto Walsh basis
Identify Zero Basis (Sparsity)
`signature` $\leftarrow$ (|`w`| < `tol`) ▷ Identify vanishing coefficients
Pattern Matching (Standard Models)
`match_id` $\leftarrow$ `LookupModel`(`signature`, `PATTERNS`)
if `match_id` is unambiguous **then**
    return `match_id`
end if
Constraint Verification (CSI Models) ▷ Check algebraic constraints like $e_{12} + e_{123} \approx 0$
`constraints` $\leftarrow$ `ComputeLinearCombinations`(`w`)
`csi_signature` $\leftarrow$ (|`constraints`| < `tol`)
`csi_id` $\leftarrow$ `LookupCSI`(`csi_signature`, `PATTERNS`)
if `csi_id` is valid **then**
    return `csi_id`
else
    return 0 ▷ Default: Saturated
end if

of the log-probability tensor. For the 15 candidate models, the structure identification is implemented as follows:

1. **Log-Transformation and Projection:** First, the empirical probability distribution $\hat{\boldsymbol{p}}$ is transformed into the log-domain logit vector $\theta = \log \hat{\boldsymbol{p}}$. This vector is then projected onto the Walsh basis via the inner product (WHT) to obtain the spectral coefficients w.

2. **Sparsity Detection:** A thresholding operation is applied to these coefficients to identify which basis terms effectively vanish ($|w_y| <$ `tol`).

3. **Signature Matching:** The resulting pattern of zero coefficients constitutes a *signature*. The algorithm attempts to identify the corresponding model by matching this signature against predefined patterns (e.g., Independence, No-3-Way).

4. **Constraint Verification for CSI:** For models involving context-specific independence (CSI), the vanishing structure appears not as single coefficients but as linear algebraic constraints (e.g., $e_{12} + e_{123} = 0$). For these cases, the algorithm additionally computes these linear combination features to distinguish specific CSI structures.

4.3.3 Computational Complexity

Before finishing this section, we briefly compare the computational complexity of the frequency-domain approach (based on the Walsh–Hadamard Transform, WHT) with that of the geometric approach based on 2×2 minors.

In its direct form, the WHT requires $O(N^2)$ operations for a probability vector of length $N = 2^n$, since each Walsh coefficient involves a full summation over all N states. However, due to the recursive structure of the Hadamard matrix, a *fast* variant (FWHT) can be implemented using a divide-and-conquer scheme, reducing the computational cost to $O(N \log N) = O(2^n n)$. Although the present study does not depend on algorithmic implementation, this result indicates that the frequency-domain analysis scales more favorably than exhaustive minor-based enumeration.

For comparison, testing all 2×2 minors over n binary variables involves

$$\binom{n}{2} 2^{n-2}$$

distinct submatrices, each requiring constant-time determinant evaluation, leading to overall cost $O(2^n n^2)$. Therefore, even without optimization, the spectral approach provides a computationally tractable alternative to combinatorial minor enumeration, and its recursive structure allows further acceleration through FWHT.

4.4 Comparative Experiment for Proof of Concept

4.4.1 Experiment design

Objective This Proof of Concept (PoC) experiment aims to empirically verify the central thesis developed in Chapter 3: that the *Minimum Invariant Constraint* (MIC) constitutes the smallest parameter-invariant structure underlying probabilistic models. An MIC remains structurally invariant under random perturbations of parameters, generalizes statistical independence, and can be detected directly from data as a local rank-one relation. It captures the essential symmetry and antisymmetry relationships that organize stochastic structures, revealing that such invariants are not accidents but intrinsic regularities of the model.

The objective of the experiment is to confirm that this invariant structure can be both computed and used for model selection. Detecting MICs by examining the vanishing of 2×2 minors demonstrates structural learnability, while distinguishing models by their pattern of vanishing minors demonstrates usefulness. At the same time, the combinatorial growth of minors poses a computational challenge, motivating the development of more efficient methods derived from the theory: a geometric approach that tests membership in the Segre variety through a lattice of moments, and a spectral approach using the Walsh transformation. The PoC implements these methods on fully observable toy data to assess whether the MIC-based framework achieves feasible computation, reliable identification,

and interpretable diagnostic behavior.

Experiment design The Proof of Concept compares three computational realizations derived from the preceding sections: (1) a direct *brute-force* evaluation of 2×2 minors (quadratic and quartic binomials), (2) a *divide-and-conquer* method based on marginalized moments and cumulants, and (3) a *spectral analysis* using the Walsh transformation to discover invariant structures. Together they embody two complementary theoretical perspectives established earlier: the geometric view, grounded in the Segre structure of probability tensors, and the harmonic view, which captures the same invariance as linear relations in the frequency domain. The goal here is to examine their empirical behavior under identical conditions using the same synthetic data, assessing how consistently each detects the underlying MIC structure across models and noise levels.

The three methods are summarized below.

1. **Brute-force by all binomial indicators (geometric)**
   All 2×2 submatrices of the probability tensor are enumerated. For m binary variables with disjoint subsets I, J, C such that $I \sqcup J \sqcup C = [m]$, each 2×2 minor is defined as

$$\Delta_{i_1,i_2,j_1,j_2,c} := p_{i_1 j_1 c}\, p_{i_2 j_2 c} - p_{i_1 j_2 c}\, p_{i_2 j_1 c}, \tag{4.17}$$

   where $i_1, i_2 \in \prod_{k\in I} \mathcal{X}_k,\ j_1, j_2 \in \prod_{k\in J} \mathcal{X}_k,\ c \in \prod_{k\in C} \mathcal{X}_k$. 2×2 minors with $\Delta_{i_1,i_2,j_1,j_2,c} < \varepsilon$ are regarded as *vanishing*. For $m = 3$, there are 12 such quadratic binomials. The resulting binary pattern of vanishing minors is compared with canonical model signatures via Hamming distance, and the vanishing coordinates indicate the localized MICs. This method provides the most explicit geometric certification of local rank-one structure, although its computational cost grows combinatorially with dimensionality.

2. **Divide and conquer by moments and cumulants (geometric)**
   Compute the first- and second-order moments μ_I, μ_J, and μ_{IJ} for all $I, J \subseteq [m]$ with $|I| = |J| = 1$. Two diagnostics are employed: (i) the spectral ratio $\rho(M) = \sigma_2/\sigma_1$, where $\mathrm{vec}(M) = \mu_{IJ}$, and (ii) the generalized pointwise mutual information (second cumulant) $\kappa_{I,J} = \log \mu_{IJ} - \log \mu_I 1^\top - 1(\log \mu_J)^\top$. When $\rho < \tau_\rho$ and $\|\kappa_{I,J}\|_{\max} < \tau_\kappa$, the corresponding block is regarded as near rank-one and escalated to conditional testing, where conditional moments $\mu_{IJ|C=c}$ define slices along an additional axis. MICs are thus localized in contexts $(I, J \mid C = c)$ whose minors vanish or nearly vanish. This method exploits the lattice structure of moments to propagate invariance efficiently, alleviating the combinatorial explosion of brute-force enumeration.

3. **Spectral analysis by Walsh coefficients (harmonic)**
   Apply the Walsh–Hadamard Transform (WHT) to the log-probability $\theta = \log p$ to obtain the spectral coefficients $\hat{\theta}(y)$. The *order pattern*—which coefficients with $|y| \geq 2$ vanish—and the *linear relations* among Walsh coefficients reveal invariant structures.

For example, a completely independent model yields vanishing coefficients for all $|y| \geq 2$, whereas in a context-specific independence (CSI) model, slice-wise inspection shows an order collapse in particular contexts such as $\hat{\theta}(e_2e_3) + \hat{\theta}(e_1e_2e_3) = 0$ for $X_2 \perp\!\!\!\perp X_3 \mid X_1{=}0$. The WHT linearizes such conditional dependencies: a context-dependent disappearance of interaction in the probability domain appears as a simple equality among spectral coefficients, showing how local contextual structure is algebraically encoded in the Walsh representation.

Synthetic data Synthetic data are generated to cover all possible MIC patterns in the $2 \times 2 \times 2$ model, comprising fifteen distinct structural types: one case of complete independence, three of joint independence, three of conditional independence, six of context-specific independence (CSI), one of No-3-Way interaction, and one of full saturation. For each pattern, a corresponding configuration matrix A_1 through A_5—defined in Chapter 3—was constructed to specify the model's structural constraints together with all relevant permutations of variables and states. Parameter vectors of dimension d for each configuration matrix were drawn from a uniform distribution $\mathcal{U}[0, 1]$ and mapped to eight-dimensional probability vectors via the toric model representation. Iterating this instatiation process yielded a family of the ground-truth probability distributions that consistently satisfy the same invariant relations, thereby allowing the extraction of parameter-invariant structures inherent to each MIC configuration. This synthetic dataset serves as a comprehensive reference for testing the learning algorithms introduced below.

Evaluation For each model class, 1000 distinct models are instantiated (i.e., ground-truth probability vectors), from which empirical probabilities are computed via iterated sampling for sample sizes $n = 1k, 10k, 100k$. These serve as the shared inputs to the three pipelines described above. The overall evaluation follows four criteria: (i) model identification accuracy, (ii) invariance to parameter variation, (iii) robustness to perturbation, and (iv) structural properties of non-vanishing indicators. Benchmark comparisons are conducted against conventional likelihood-based model selection using maximum likelihood estimation with AIC/BIC penalties. This allows direct assessment of the performance gains achieved by invariant, structure-based criteria.

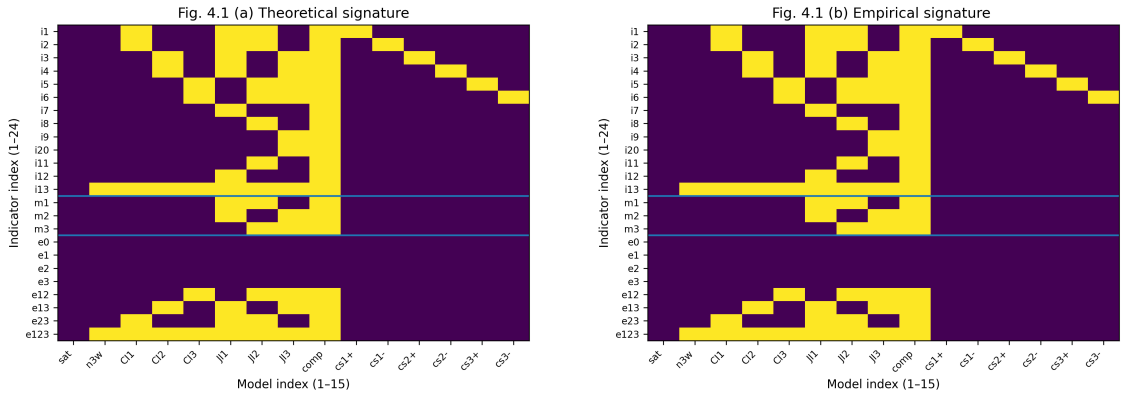
Finally, we compare the ability to select correct models among the geometrical approaches (a,b)—the binomial indicators (BI) and moment and cumulants (Moments)—and the harmonic approach (c) based on the Walsh coefficients (WC). By examining how the same underlying data patterns are reflected across these representations, we can identify which domain yields greater stability and interpretability under noise. Together, these evaluations demonstrate that the MIC-based learning framework provides a consistent and computationally feasible means of detecting structural invariants across probabilistic models.

4.4.2 Results and Analysis

This section presents experimental evidence supporting the theoretical properties of MIC developed in earlier chapters. Five aspects are examined: (i) numerical verification of MIC, (ii) invariance to parameter variation, (iii) robustness to noise, (iv) structural consistency among non-vanishing indicators, and (v) model selection performance.

Numerical Verification of MIC. We first verify that the MIC signatures computed from numerical probabilities exactly match the theoretically derived signatures. Specifically, we compare the Theoretical Signature, defined by the symbolic vanishing ideal (e.g., 2×2 minors), against the Synthetic Signature, which is computed directly from the true probability distributions generated by a Toric model with random parameter instantiation. Unlike empirical validation using sampled data, this synthetic approach allows us to verify the exactness of the constraints without the interference of finite-sample noise.

For each of the fifteen canonical models, we compute all twenty-four indicators—thirteen binomial ideals (rows $i1, \ldots, i13$), three marginal-moment quantities (rows $m1, m2, m3$), and eight Walsh coefficients ($e_0, \ldots, e_{123}$)—and compare them with the theoretical reference. Figure 4.2 presents a heatmap demonstrating their perfect agreement. Across all models and indicators, the Synthetic signatures exhibit perfect agreement with the Theoretical ones. This confirms that the proposed computation pipeline implements the algebraic constraints with complete fidelity. The numerical threshold to binarize computed indicator values is set to 10^{-12} in absolute value.



(a) Theoretical Signature (b) Synthetic Signature

Figure 4.2: Binary heatmap: exact correspondence between (a) theoretical and (b) synthetic signatures across all 15 models is observed. The X-axis represents 15 canonical models for $2 \times 2 \times 2$ probability distribution and the Y-axis represents 24 indicators—13 vanishing binomials, 3 moments, 8 Walsh coefficients. Numerical values are binarized by the threshold 10^{-12} for Synthetic signatures.

Invariance to Parameter Variation. To examine whether MIC signatures remain invariant over specific parameter choices, we perturb each model's parameters randomly across

1000 trials, which generates 1000 different probability vectors $\boldsymbol{p}$ of 8 dimensions. For each perturbed distribution, we recompute the MIC signature and measure the agreement at all theoretically vanishing entries in the same way as in the previous analysis. Figure 4.3 summarizes the resulting distribution of agreement rates by signature category in a box-plot format.

Overall, the synthetic signatures remain exceptionally stable under parameter variation, with agreement rates ranging from 0.917 to 1.000 and a mean of approximately 0.997 with very small variance. Among the three categories of the indicators, the marginal-moment quantities exhibit slightly larger variability than the minors and Walsh coefficients. These results confirm that MIC captures genuine structural invariants rather than parameter-dependent artifacts.

Figure 4.3: Distribution of match rates between Theoretical and Synthetic signatures under 1000 random parameter perturbations. The result shows structural invariance over parameter perturbation especially of 2×2 Minors $(i_1, \ldots, i_{13})$ and Walsh coefficients $(e_0, \ldots, e_{123})$.

Robustness to Noise. We next assess the robustness of MIC signatures to stochastic perturbations. Given a true distribution $\boldsymbol{p}$, we add Gaussian noise in the logit domain, $\log(p_i) + \varepsilon_i; \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 \in \mathbb{R}_{>0}$ denotes the variance, and renormalize the result to obtain a noisy distribution. For each noise level σ, we add a sampled noise to each of 1000 true probabilities for each model and compute the empirical signatures. Figure 4.4 shows the signature agreement as a function of σ.

The result reveals the characteristic regime in which MIC signatures remain structurally robust under perturbations. When Gaussian noise is added in the logit domain, the agreement with the ground–truth signature decreases smoothly as the noise level σ increases from 0 to approximately 5×10^{-3}. Importantly, for sufficiently large σ, all methods approach a plateau near 0.78—the proportion of zero entries in the ground–truth

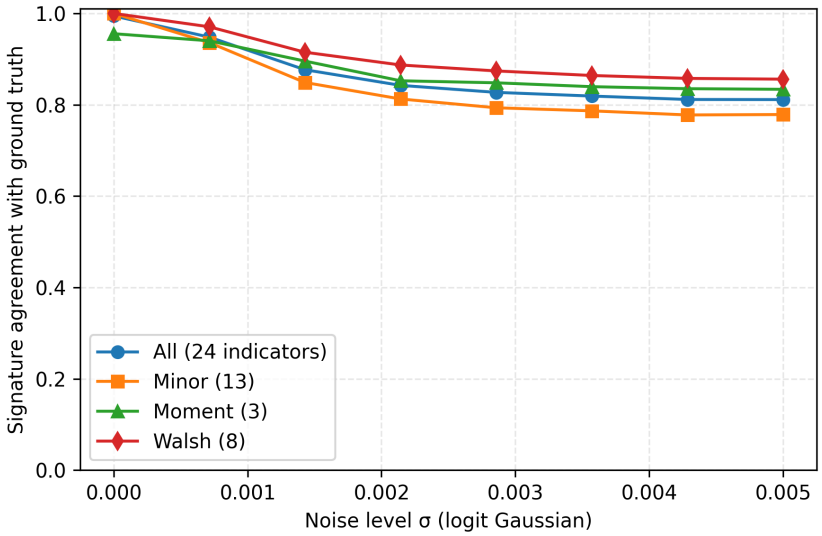


Figure 4.4: Signature agreement under logit–Gaussian noise. Curves correspond to minor, moment, Walsh, and full 24-dimensional signatures and cleanly indicate a robust regime in which MIC signatures tolerate noise.

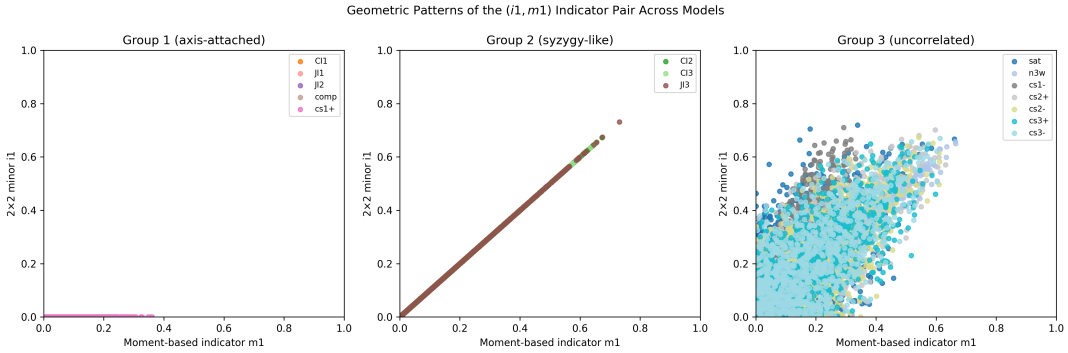
signature (78.1%). The observed decay curve cleanly indicate a robust regime in which MIC signatures tolerate perturbations.

Structural Consistency among Non-Vanishing Indicators. Figure 4.5 summarizes three complementary geometric analyses that demonstrate how structural information persists even when individual indicators do not vanish.

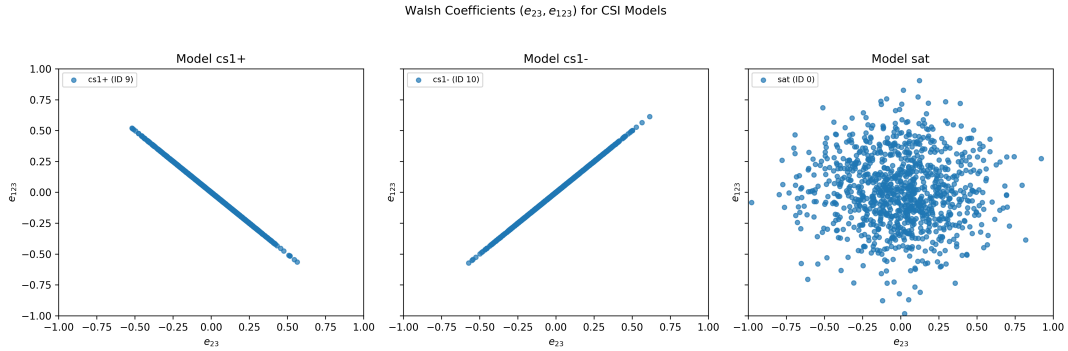
Panel (a) examines the joint behavior of the moment-based indicator m_1 and its minor-based counterpart i_1 across different families of models. The models form three qualitatively distinct geometric regimes—axis-attached, syzygy-like, and uncorrelated—reflecting the algebraic constraints inherent to each model class. Even without vanishing, the relative positioning of points in the (m_1, i_1) plane reveals characteristic structural signatures.

Panel (b) investigates the dependency between Walsh coefficients e_{23} and e_{123}, which encode parity-based interactions among variables. Models exhibiting context-specific independence (CSI) display a clear linear relationship between these coefficients, whereas other models produce diffuse or incoherent patterns. This demonstrates that the Walsh domain captures algebraic structure that is not visible through minors or moments alone, i.e., the symmetrical relationship two CSI models for $X_1 = 0$ and $X_1 = 1$.

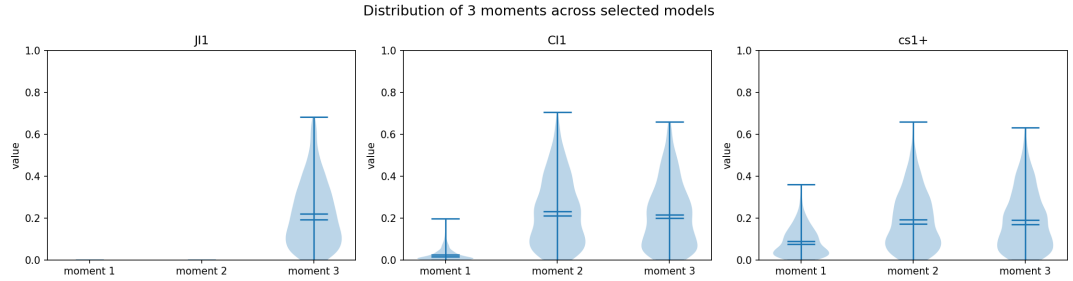
Panel (c) compares the distribution of a moment indicator across representative models, highlighting how different classes—such as conditional independence, CSI, and saturated models—exhibit distinct distributional shapes. These differences persist even when the moment itself is non-zero, illustrating that internal structure of rank-one slice(s) affects the range and variability of the indicator rather than only its vanishing behavior that corresponds to a secant variety. Together, these three analyses show that structural constraints manifest as robust geometric patterns, offering a richer and more nuanced view of model



(a) Different correlation patterns between the moment-based and minor-based indicators (m_1, i_1) across groups of models, effective in identifing a calss of models.



(b) Dependency patterns between Walsh coefficients (e_{23}, e_{123}), revealing characteristic CSI structures.



(c) Distributions of the first-order moment indicator showing internal structure—such as conditional independence or CSI on X_1—even when the moment does not vanish.

Figure 4.5: **Structural properties revealed by geometric relations among non-vanishing indicators.** Each panel highlights a different type of algebraic dependency that remains detectable even when indicators do not vanish, demonstrating that structural regularities extend far beyond simple rank-one signatures and can be captured through their geometric relationships.

Sample size	ML	AIC	BIC	BI-0	BI-N	Moments	WC	RF-P	RF-BI
∞ (true prob.)	0.565	0.982	0.957	**1.000**	**1.000**	0.693	**1.000**	0.479	1.000
$n = 10^5$	0.070	0.702	0.851	0.798	**0.857**	0.624	0.823	0.483	0.871
$n = 10^4$	0.067	0.582	0.637	0.595	**0.661**	0.504	0.596	0.470	0.663
$n = 10^3$	0.067	0.352	0.311	0.315	**0.367**	0.310	0.325	0.398	0.344

Notes. ML = maximum likelihood; AIC = Akaike Information Criteria; BIC = Bayesian Information Criteria; BI-0 = unnormalized binomial indicators (raw minors); BI-N = normalized binomial indicators indicators; WC = Walsh coefficients; RF-P = Random Forest (probability vectors); RF-BI = Random Forest (binomial indicators).

Table 4.1: Comparison of model selection accuracy. The normalized minor-based method (BI-N) demonstrates superior performance across all sample sizes, surpassing the standard BIC. Walsh coefficients remain competitive with BIC, whereas the Moment method shows lower discriminative power despite its computational efficiency. ML performs at the chance level ($1/15 \approx 0.067$) because it consistently selects the saturated model, which has the highest degrees of freedom.

behavior than binary signatures alone.

Model Selection via MIC Signatures. Finally, we evaluate the utility of MIC signatures for model selection. Synthetic datasets are generated by sampling $n = 10^3, 10^4, 10^5$ observations from each underlying true model to construct empirical histograms. To assess the impact of sampling noise on model selection performance, we also utilize theoretical probability distributions (representing the limit of infinite sampling). We compare six selection procedures: (i) full minor-based matching with raw and normalized values (BI-0,BI-N), (ii) a two-step marginal–minor procedure (Moments by Algorithm 1), (iii) Walsh-based matching (WC by Algorithm 2), (iv) maximum likelihood, serving as a baseline (ML), (v) maximum likelihood with AIC (AIC), and (vi) maximum likelihood with BIC (BIC). The normalized minors are computed as:

$$\frac{p_{i_1 j_1 k} p_{i_2 j_2 k} - p_{i_1 j_2 k} p_{i_2 j_1 k}}{p_{i_1 j_1 k} p_{i_2 j_2 k} + p_{i_1 j_2 k} p_{i_2 j_1 k} + \epsilon}, \tag{4.18}$$

where $\epsilon = 10^{-12}$ is added to avoid division by zero. Note that this formulation is mathematically equivalent to applying a $\tanh$ transformation to the log-odds ratio. This effectively acts as a variance-stabilizing transformation for scale-dependent noise (typical of Gamma distributions), ensuring that the metric remains robust across different probability scales. BIC is considered the standard criterion for model selection in this context [133]. To ensure a fair comparison, all methods are evaluated on identical fixed test sets for each sample size n. For reference, to assess the effectiveness of the indicators as features, we also apply a Random Forest classifier to compare raw probability vectors and the 13 normalized binomial features to investigate the extent to which the proposed indicators capture the structural information necessary for model identification.

Table 4.1 presents the classification accuracy for each model selection method. The normalized minor-based method (BI-N) emerged as the best-performing approach, consis-

tently achieving the highest accuracy rates across all sample sizes and outperforming the maximum likelihood based BIC. The Walsh coefficient-based method also demonstrated consistent performance, following BI-N and performing on par with BIC on average. While the Moment method—noted for its computational efficiency—lagged behind the minor and Walsh-based approaches, it maintained a certain level of accuracy, suggesting potential for improvement through further parameter tuning.

Regarding the theoretical limit using true probability distributions, the Minor (BI-0, BI-N) and Walsh methods achieved 100% accuracy, a result consistent with our prior theoretical analysis. In contrast, the Moment method reached an accuracy of only 0.694, indicating that the current algorithm and threshold settings require further refinement. The Maximum Likelihood (ML) baseline yielded an accuracy of 0.067, corresponding to the chance rate ($1/15$). This low performance is expected, as the unpenalized ML inevitably selects the saturated model due to its having the highest degrees of freedom.

Finally, we analyze the Random Forest (RF) results. It should be noted that the RF models were trained on 70% of the data and evaluated on the remaining 30%, making direct numerical comparison with the other procedures inappropriate. However, the significant performance gap observed between RF trained on raw probabilities (RF-P) and RF trained on binomial features (RF-BI) suggests that standard machine learning algorithms struggle to automatically discover the specific structural features, such as binomial invariants, necessary for identifying these algebraic models from raw probability vectors.

4.5 Discussion

Structural Robustness of MIC Signatures The experimental results consistently demonstrate that the MIC signatures contain a stable invariant core. A substantial subset of the 2×2 minors, marginal moments, and Walsh coefficients remains unchanged under parameter perturbation, sampling noise, and even adversarial distortions, allowing us to identify a range within which the structural signature is preserved. This robustness is observed not only at the level of individual indicators but also across the two analytical domains employed in this study. All representations in the geometric approach (vanishing minors and marginal moments) and in the harmonic analysis framework (Walsh coefficients) converge on the same underlying structural constraints, revealing a coherent set of structural regularities shared across these distinct mathematical formalisms. These findings suggest that structural invariants capture the essential qualitative dependencies encoded in a distribution, even when quantitative probability estimates vary substantially. In other words, MIC signatures isolate those aspects of the generative structure that must remain true, regardless of noise, parameter variation, or estimation error. This robustness provides empirical justification for viewing MIC-based structural units as stable and reliable descriptors of linguistic probability distributions.

Structural Learning Based on Invariants Another important aspect is that the proposed model-selection methods rely solely on checking algebraic constraints, which require only simple arithmetic computations (e.g., evaluating minors or indicator functions) directly from the observed data. This stands in sharp contrast to conventional learning methods, which depend on optimizing an objective function—such as likelihood combined with AIC or BIC—involving iterative numerical computation and no guaranteed convergence to the global optimum. In this sense, the proposed approach can be regarded as a form of *structural learning*, in which model discrimination is achieved by identifying the structural properties of each model class rather than by estimating parameters within a fixed model. Such structural invariants, exemplified by MIC, may therefore provide a powerful and computationally efficient alternative perspective on learning, complementing and potentially expanding traditional parameter-centric approaches in machine learning.

Computational Complexity and the Curse of Dimensionality As noted earlier in this chapter, exhaustively computing all 2×2 minors of a high-order probability tensor is computationally infeasible: the number of minors grows combinatorially with the tensor's order and mode sizes, leading to a worst-case complexity that is exponential in the number of variables. To address this, we consider two complementary strategies that reduce the effective search space while preserving structural information.

First, a *divide–conquer* approach based on lower-order moments allows us to systematically prune the exploration tree. By examining marginals and conditional marginals of increasing order, one can eliminate branches that violate necessary structural constraints, yielding an exponential reduction in computation in practice, even though the worst-case guarantees remain open.

Second, a *harmonic analysis* approach leverages the Walsh–Hadamard transform, an instance of the fast Fourier transform on $(\mathbb{Z}_2)^m$. This provides a frequency-domain representation in which many structural invariants (such as CSI constraints) appear as sparse or low-order interactions. The fast Walsh–Hadamard transform computes this representation in $O(N \log N)$ time, offering a substantial improvement over the naive $O(N^2)$ computation of all pairwise interactions where $N = k^m$ with m is the number of variables and k is the number of states.

To compare the scalability of these approaches, consider a binary system with three variables ($m = 3, n = 2$). In this setting, the brute-force method typically requires checking an average of 13 indicators and the Walsh coefficient method utilizes 14, whereas the moment-based approach relies on only 4. Generally, for a system with m variables and n states, the computational complexity for feature identification scales as $O(m^2 n^4)$ for brute force and $O(n^2 m^2)$ for Walsh coefficients, but drops to $O(m + n)$ for the moment-based strategy. Although the moment-based method currently lags in classification performance, this superior linear scaling makes it the most promising candidate for high-dimensional language modeling, provided that parameter tuning is further refined. For application to language models, we propose a stepwise exploration using moments. The effective

computational cost of this strategy will depend on the intrinsic geometry of language: if linguistic compositionality is shallow, examining low-order moments will suffice; conversely, if high-order interactions exist but are sparse, the branching factor for deep exploration remains manageable. Verifying these structural hypotheses and optimizing the pruning process are key objectives for future research.

MIC as an Inductive Bias The structural invariance exhibited by MIC signatures suggests that they can serve as a powerful form of *inductive bias* for learning in probabilistic and neural language models. Because MICs capture algebraically well-defined and interpretable structural regularities, and because these regularities remain stable even under substantial sampling noise or parameter variation, they provide reliable constraints that persist in low-data regimes. Identifying such invariants *prior* to model training effectively reduces the hypothesis space that a learning algorithm must explore, thereby improving both optimization efficiency and generalization performance.

Moreover, MICs can be expressed in multiple coordinate systems—including algebraic, probabilistic, and harmonic views—each highlighting different aspects of the same underlying structural constraints. This representational flexibility suggests, in a more modest sense, that MICs may help bridge structured probabilistic modeling and current neural architectures. While any direct correspondence with mechanisms such as attention or feed-forward layers of Transformer-based large language models remains an open question, the decomposition of distributions into invariant components provides a principled way to formulate structural priors. Such priors could, in future work, be incorporated into neural models to guide learning toward linguistically meaningful dependencies rather than relying solely on data-driven estimation.

Chapter 5

Application: Algebraic Structure in PMI Matrices

5.1 Overview of the Chapter

This chapter presents an empirical study demonstrating that locally rank-one structures extracted from real corpus data correspond to interpretable semantic relations. The work reported here was the starting point of the overall research program developed in this dissertation, originally conducted to understand the internal structure of word representations derived from word co-occurrence statistics.

It has long been observed that word co-occurrence–based vectors encode semantic and syntactic information, and since the advent of Word2Vec [96], it has become widely recognized that even higher-order semantic relations—such as analogies—manifest as geometric configurations in the embedding space. The present study was motivated by the question of *why* such semantic and syntactic regularities should be reflected in the mathematical structure inherent in word vectors and in the underlying co-occurrence distributions from which they are derived. Far from being trivial, this question demands a rigorous theoretical elucidation.

Applying Formal Concept Analysis (FCA) and Non-negative Matrix Factorization (NMF) to word co-occurrence and PMI (pointwise mutual information) matrices constructed from a corpus, we identified locally rank-one substructures that correspond to interpretable semantic relations. Moreover, these rank-one components were found to appear with overlaps and to form hierarchical patterns. In light of the theoretical developments in Chapters 3 and 4, these empirical findings suggest that the observed rank-one structures arise from the compositional organization of language and correspond to the parameter-invariant minimal building blocks defined in this dissertation as *Minimum Invariant Constraints* (MICs). In particular, as shown in Chapter 4, a PMI matrix can be interpreted as computing second-order cumulants of a cross-moment table obtained by marginalizing a probability tensor to two variables. Thus, the PMI matrix is expected to reflect the network of MICs—local subtensors—that constitute the deep structural

organization of the underlying probability distribution.

Although the full application of the proposed theoretical framework and learning methodology to real corpus data remains future work, the analysis in this chapter demonstrates that PMI matrices, together with probability tensors constructed directly from corpora, exhibit rich mathematical structure. These observations provide a concrete basis for understanding the mechanisms through which such structure emerges.

5.2 Background

Word vector representations are central to natural language processing, as they capture semantic and syntactic features [79]. Their significance has been amplified in recent times, as they are used as input for Transformer-based language models [138], where static embeddings are contextualized. Their effectiveness has been explained by the distributional hypothesis [60] linking similar semantics and similar distribution [69]. However, the interpretability of their dimensions remains an active research topic [126]. [81] found neural word embeddings to be uninterpretable while acknowledging that sparse vectors capture some latent topics. [55], among others, pioneered efforts to interpret dynamic embeddings in GPT-2 [120] by projecting them into the vocabulary space, though a systematic approach to interpret dimensions of embeddings remains an open issue.

Many preceding studies have investigated the semantic properties of word embeddings and revealed that word vectors in a vector space capture relational meanings. The most well-known example is the parallelogram formed in the vector space by the embeddings of words in analogical relations (e.g. *king:queen::man:woman*) [97]. Other semantic relationships also exhibit geometrical counterparts, such as semantic composition with vector addition [96, 99], hypernymy captured by linear projection [51], and polysemy as a linear combination of vectors [4]. Regarding the theoretical analysis of embeddings, [82] suggested that word2vec [95] is equivalent to the factorization of a word co-occurrence matrix. [3] proposed a generative model in which PMI-based word embeddings exhibit linear structures. These related studies collectively hint that the latent structure in the co-occurrence matrix reflects linguistic regularities and is inherently embedded within vector representations. Therefore, understanding the word co-occurrence matrix represents a cornerstone in elucidating the interpretability of word representations.

In this study, we directly address the mathematical structure of a word co-occurrence matrix to uncover underlying linguistic patterns and to interpret the dimensions of word embeddings. We claim that *a formal concept*, as mathematically defined in the matrix, corresponds to human-interpretable categories. We substantiate our claim through the category completion task. Specifically, we used Formal Concept Analysis (FCA), a field of applied mathematics [52], to formally characterize the internal structure of a matrix. We define a group of words as interpretable if it can be descriptively labeled. Furthermore, we demonstrate that a hierarchical structure of formal concepts emerges as a geometric formation in the vector space, which explains why relational meanings are captured by

word embeddings.

Our contributions are threefold. First, we propose two methods that apply FCA to real-valued data: binarization by varying thresholds and fuzzification of FCA. Second, we empirically show that the formal concepts in the co-occurrence matrix coincide with interpretable categories. Third, we present a novel algorithm to detect formal concepts, which is capable of disambiguating polysemous words. To our knowledge, this is the first study to apply FCA to a word-word co-occurrence matrix. Our study offers a new approach to uncover latent linguistic structures in co-occurrence matrices.

5.3 Formal concept analysis of word co-occurrence matrix

5.3.1 Basics of FCA

FCA is related to order theory and abstract algebra. It mathematizes *concepts* and *conceptual hierarchy* [52]. A concept comprises a pair of its extents (objects) and its intents (attributes). Concepts can form a hierarchy known as a *lattice*. FCA has been empirically applied for data mining and ontology [114], especially in bioinformatics [124].

A **formal context** $\mathbb{K} := (G, M, I)$ consists of two sets G, M and a binary relation $I \subseteq G \times M$. The elements of G and M are called **objects** and **attributes**, respectively. For $g \in G$ and $m \in M$, a relation $(g, m) \in I$ means that the object g has the attribute m. We define two derivation operators; $\uparrow : 2^G \to 2^M$ maps a subset of objects to a subset of attributes, and its reverse $\downarrow : 2^M \to 2^G$ maps attributes to objects. For $A \subseteq G, B \subseteq M$,

$$A\uparrow := \{m \in M \mid (g, m) \in I \ (\forall g \in A)\} \tag{5.1}$$

$$B\downarrow := \{g \in G \mid (g, m) \in I \ (\forall m \in B)\} \tag{5.2}$$

$A\uparrow \subseteq M$ is the set of attributes common to all objects in A, whereas $B\downarrow \subseteq G$ is the set of objects that possess all the attributes in B. It can be shown that $A \subseteq B\downarrow \Leftrightarrow B \subseteq A\uparrow$, which is a structure-preserving (order-reversing) correspondence between ordered sets known as a Galois connection [38].

A **formal concept** of the context (G, M, I) is defined as a pair $(A, B) \in 2^G \times 2^M$ where both $A\uparrow = B$ and $B\downarrow = A$ hold. A and B are considered the extent and intent, respectively, of the formal concept (A, B). The compositions of two derivation operators $\uparrow\downarrow : 2^G \to 2^G$ and $\downarrow\uparrow : 2^M \to 2^M$ are closure operators [38], with a formal concept defined as the fixed point of these operations. If a formal context is represented as a binary matrix, it corresponds to a maximal rectangular (submatrix) with all ones in its entries when the rows and columns are appropriately reordered.

A formal concept can also be equated with a maximal **biclique**, i.e., a complete subgraph of a bipartite graph [28]. All elements of A and B are completely connected within that subgraph.

5.3.2 Rational and benefit of using FCA

A word co-occurrence matrix, used as input data to learn word embeddings, is constructed by counting the frequency of a target-context word pair that co-occurs in the neighborhood. By regarding target words as objects and context words as attributes, we can express this co-occurrence as a binary relation. Thus, we can treat a co-occurrence matrix as a formal context.

FCA is effective in analyzing co-occurrence matrices for three reasons. First, it can characterize a local structure within the matrix. Second, formal concepts can capture relations between more than three words, which cannot be represented by individual pairwise relationships, yielding a richer analysis of the structure. Third, we can define (partial) order relation between formal concepts. A semantic relationship such as hypernymy can be formalized by such an order relation. We further demonstrate the function of FCA in Section 5.4.

To apply the crisp (binary) FCA to a real-valued co-occurrence matrix, we tested two approaches. First, we simply binarized the matrix values by thresholds, with a varying threshold method deployed to flexibly locate formal concepts (Section 5.5). Second, we extended the crisp FCA to an FCA built on fuzzy logic (Section 5.6).

5.4 Demonstration using synthetic data

5.4.1 Artificial toy corpus

We examined how FCA handles a word co-occurrence matrix using a toy corpus. We demonstrated that formal concepts capture semantic categories emerging from word usage in the corpus and introduced a **concept lattice** of FCA to illustrate the hierarchical structure of concepts.

The demonstration contains 1) a corpus of 24 synthetic sentences with 17 words (Appendix A.1), 2) a co-occurrence matrix obtained from the corpus, and 3) word vectors acquired from the matrix (Fig. 5.1). The corpus is designed to replicate a geometric formation of the analogy relation. Specifically, we targeted eight words—*king*, *queen*, *man*, *woman*, and their plurals—so that their vectors formed a parallelepiped. The sentences were expressed analogously: E.g., *'king (queen) live in palace"*, whereas *"man (woman) live in house"*. The co-occurrence matrix $X \in \{0, 1\}^{17\times17}$ is binary, where $X_{ij} = 1$ if two words co-occur in a sentence and $X_{ij} = 0$ otherwise. Each row of this matrix represents a word vector. Projected on the 3-dimensional space, the eight word vectors form a parallelepiped (Fig. 5.2).

5.4.2 Detecting formal concepts

We now apply FCA to the matrix X. Although formal concepts can be determined by applying the closure operator $\uparrow\downarrow$, a simplified method is to find a rectangular in the

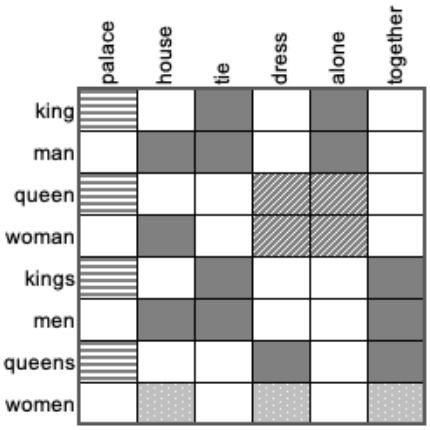


Figure 5.1: Binary co-occurrence (sub)matrix: Each entry is 1 if shaded and 0 otherwise. Each row is a word vector. Three submatrices with shade patterns indicate different formal concepts f, e, v.

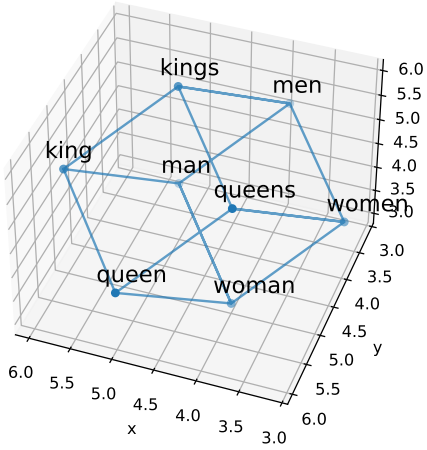


Figure 5.2: A parallelepiped emerges when eight word vectors (rows) are projected onto 3-dimensional space.

matrix. For example, the submatrix of rows $i \in \{1, 3, 4, 7\}$ and columns $j \in \{1\}$ represents a formal concept, as all its entries are 1s and no other rectangular matrix contains it. This concept represents a pair of the extent $\{king, queen, kings, queens\}$ and the intent $\{palace\}$, interpreted as "royal."

There are a total of 28 formal concepts in this matrix (see Appendix A.2 for the list and notation). They are classified into five types, including two trivial ones wherein one element is empty. Examples of the three non-trivial types include the following:

$$f_1 := (\{king, man, kings, men\}, \{tie\}) \tag{5.3}$$

$$e_1 := (\{king, man\}, \{tie, alone\}) \tag{5.4}$$

$$v_1 := (\{king\}, \{tie, palace, alone\}) \tag{5.5}$$

To see hierarchical relations between formal concepts, we first define the order relation. Let $\mathfrak{B}(G, M, I)$ be the set of all concepts of (G, M, I). Given $(A_1, B_1), (A_2, B_2) \in \mathfrak{B}(G, M, I)$,

$$(A_1, B_1) \leq (A_2, B_2) \overset{\text{def}}{\iff} A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2 \tag{5.6}$$

Thus, if the extent A_1 is contained by the extent A_2, then the formal concept (A_1, B_1) is less than or equal to (A_2, B_2). Owing to the Galois connection, $A_1 \subseteq A_2$ holds if and only if $B_1 \supseteq B_2$. Then, $\langle \mathfrak{B}(G, M, I) : \leq \rangle$ is a complete lattice known as a **concept lattice**, a nonempty ordered set where a join and a meet exist for all elements and subsets. Fig. 5.3 visualizes all ordered relations between the formal concepts identified in the matrix X. We

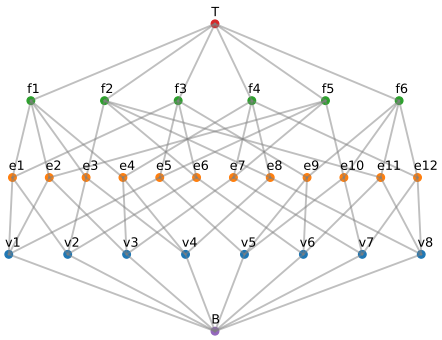


Figure 5.3: Concept lattice. Each node represents a formal concept. They correspond to geometric simplices of the parallelepiped: 8 vertices, 12 edges, 6 faces.

observe that the lattice of formal concepts (Fig. 5.3) corresponds to the parallelepiped (Fig. 5.2). This suggests that geometric relations between word vectors reflect the hierarchical structure latent in the word co-occurrence matrix.

5.4.3 Three implications of FCA

First, FCA allows us to easily interpret the identified formal concepts. For example, f_1 should be labeled as **masculine** from its extent {*king*, *kings*, *man*, *men*}, whereas f_6, with

the extent {*queen, queens, woman, women*}, must be labeled as **feminine**. The other f-type concepts can be labeled as **royal**, **common**, **singular** and **plural**. Thus, formal concepts coincide with semantic categories.

Second, v_1 (*king*) can be seen as the intersection of three others—f_1, f_3, f_5— analogous to a vertex included in three faces. Semantically, *king* is something royal, masculine, and singular. This relation can be algebraically formulated as $v_1 = f_1 \wedge f_3 \wedge f_5$ where $\wedge$ is a *meet* operation.

Third, pairs of opposing faces in the parallelepiped form complementary concepts such as **masculine** vs. **feminine**. Mathematically, we can construct a *formal concept algebra* by defining additional operations as axioms [145]. Using this algebra, the formal concept of **masculine** can be demonstrated to complement that of **feminine**; $\neg f_1 = f_6$ where $\neg$ is a negation. The observation that *king* $v_1 = f_1 \wedge f_3 \wedge f_5$ and *queen* $v_5 = f_6 \wedge f_3 \wedge f_5$ share f_3 and f_5 explains the phenomenon that both synonyms and antonyms appear close to each other in the vector space [137].

In summary, the co-occurrence matrix exhibits the geometrical and algebraic structures formed by interpretable formal concepts.

5.5 Experiment 1: FCA by binarization

We now demonstrate that formal concepts can be defined on actual word co-occurrence data and correspond to both semantic and syntactic categories.

5.5.1 Algorithm to identify formal concepts

We designed a novel algorithm to locate formal concepts through the conversion of two derivation operators (Eq. 5.1 and 5.2). The corresponding pseudo-algorithm is shown in Algorithm 3. Given a co-occurrence matrix X and set of target words S as a seed, the algorithm returns a formal concept $(S\uparrow\downarrow, S\uparrow)$, which is a pair of two subsets of the vocabulary. Here, $S\uparrow\downarrow$ is the closed set of S.

The first derivation operator $\uparrow$ must identify context words that co-occur with all target words in S. In other words, a context word is selected when it has all entry values exceeding the threshold t for the target words in S. Equivalently, any entry value that the seed words have with the context word should not be less than t, meaning that their minimum must be greater than or equal to t. As indicated in Line 3, the algorithm finds the minimum value that the seed words (in rows $\forall i \in I_S$) have against a certain context word (in a column $j \in \{1, \ldots, N\}$), sorts them in descending order (Line 5), and selects the first k context words (columns) $S\uparrow$ (Line 6). The threshold is automatically determined as the kth largest minimum value (Line 8). Next, an inverse operation executes. Given $S\uparrow$, the algorithm finds a minimum value over the context words $S\uparrow$ ($J_{S\uparrow}$ in the column index) against a target word in a row i (Line 11) and selects the target words (rows $I_{S\uparrow\downarrow}$) with minimum values exceeding the threshold (Line 13), which form $S\uparrow\downarrow$.

Algorithm 3 Varying Threshold Method

Input: $X \in \mathbb{R}^{N \times N}, S := \{w_i\}_{i \in I_S}, k \in \mathbb{N}$
Output: $FC := (S\uparrow\downarrow, S\uparrow), t \in \mathbb{R}$

function FINDFORMALCONCEPT(S, k)
  for $j \leftarrow 1$ to N **do**
    $m_j \leftarrow \min_{i \in I_S} X_{ij}$
  end for
  Sort $[m_j]$ in descending order $\leftarrow [m_{p(j)}]$
  $J_{S\uparrow} \leftarrow \{p(j)\}_{j \leq k}$
  $S\uparrow \leftarrow \{w_j\}_{j \in J_{S\uparrow}}$
  $t \leftarrow m_{p(k)}$
  $I_{S\uparrow\downarrow} \leftarrow \emptyset$
  for $i \leftarrow 1$ to N **do**
    $\mu_i \leftarrow \min_{j \in J_{S\uparrow}} X_{ij}$
    if $\mu_i \geq t$ **then**
      $I_{S\uparrow\downarrow} \leftarrow I_{S\uparrow\downarrow} \cup \{i\}$
    end if
  end for
  $S\uparrow\downarrow \leftarrow \{w_i\}_{i \in I_{S\uparrow\downarrow}}$
  return $(S\uparrow\downarrow, S\uparrow), t$
end function

$I_{S\uparrow\downarrow}$ and $J_{S\uparrow}$ are subsets of rows and columns corresponding to $S\uparrow\downarrow$ and $S\uparrow$, respectively. t is the determined threshold. The algorithm ensures that a submatrix $(X_{ij})_{i \in I_{S\uparrow\downarrow}, j \in J_{S\uparrow}}$ satisfies:

$$X_{ij} \geq t \quad (i \in I_{S\uparrow\downarrow},\ j \in J_{S\uparrow}) \tag{5.7}$$

$$X_{ij} < t \quad (\forall j \notin J_{S\uparrow}, \exists i \in I_{S\uparrow\downarrow}) \tag{5.8}$$

$$X_{ij} < t \quad (\forall i \notin I_{S\uparrow\downarrow}, \exists j \in J_{S\uparrow}) \tag{5.9}$$

Note that the submatrix of $I_{S\uparrow\downarrow} \times J_{S\uparrow}$ is discriminated from its neighbouring area. Its inner region has higher values than t (Eq. 5.7), whereas each of its exterior rows and columns horizontally (Eq. 5.8) and vertically (Eq. 5.9) adjacent to the submatrix contains at least one cell below the threshold. In other words, the higher entry values discriminate the submatrix of a formal concept from its neighbors, forming a local plateau-like structure that is not necessarily captured by the cosine similarity.

5.5.2 Category completion test

The experiment was conducted to verify that the formal concepts identified from the co-occurrence matrix coincide with interpretable categories.

Test set We adopted two existing test sets from [83] containing semantic categories: the Battig set [25], comprising 53 categories with 10 words for each, and BLESS [8], which

contains 17 categories with 5-17 words for each. We also compiled two additional sets: Series and Syntactic. The tested categories are listed in Appendix A.3.

Procedure For each category, we systematically furnished the algorithm with all possible word pairs as seeds derived from the category's word set. Next, we identified the optimal seed that yields the most extensive set of accurately classified words. We then assessed how effectively the algorithm retrieves the correct words from the optimal seed for the given category (**Precision, Recall**). Because the word sets are not necessarily exhaustive, we also regarded those missed words as correct, based on our human judgement (**Extended precision)**[1].

Baseline We used a similarity-based approach as a baseline. Specifically, we applied the k-nearest neighbor algorithm with cosine similarity. To ensure a fair comparison, we utilized the identical optimal seeds derived by the FCA method and found the nearest vectors to their mean vectors.

Data The co-occurrence matrix was constructed from the English Wikipedia dump (20171001)[2] (2.9B tokens), counted with a window of 10. We adopted PPMI (positive point-wise mutual information) as it yields the best results in the semantic task [25]. To keep the matrix size manageable, we limited the vocabulary to the 10K most frequent words.

5.5.3 Results

Seed	**Formal Concept** (upper: extents; lower: intents)	**Th.**	**Category**
large, huge	*large, huge, enormous, vast*	3.95	Adjectives of size
	sums, amounts, quantities		
large, small	*large, small*	3.47	Adjectives of scale
	amounts, quantities, intestine		
church, temple, mosque	*chapel, church, mosque, synagogue, temple*	2.85	Religious buildings
	worship, jpg, ruined		
quicker, bigger, warmer	*bigger, brighter, colder, cooler, heavier, hotter, louder, ...*	2.45	Comparatives
	than, considerably, deeper		

Table 5.1: Examples of formal concepts identified from a binarized PPMI matrix. Given seed words, the algorithm returns an extent–intent pair representing a formal concept. The parameter k was set to 3. Th. means threshold.

Qualitative results Table 5.1 presents the output samples produced by the algorithm. When given {*large, huge*} as a seed, the algorithm returned {*large, huge, enormous, vast*} as

[1]The annotation was done by one of the authors, who is non-native but has educational experience in the U.S.

[2]CC BY-SA 3.0; https://dumps.wikimedia.org/legal.html

the extent and {*sums, amounts, quantities*} as the intent, which constitutes a formal concept. All PPMI values within this concept exceeded 3.95. This formal concept can be labeled as "largeness" or Adjective of size, which implies that it is indeed interpretable. Interestingly, another formal concept consisting of {*large, small*} arises from the different seed instead. Similar results held for other seeds.

Quantitative results Table 5.2 shows that 61.5–84.3% of the identified extent words matched the category labels in the test sets (**Extended precision**). Furthermore, 56.3–76.8% of the words in the test sets were retrieved by the algorithm (**Recall**). Semantic categories in Battig, BLESS, and Series were more effectively captured by formal concepts than syntactic categories. We also observed that homogeneous categories (e.g., Country) frequently formed formal concepts. With the exception of the Extended Precision metric for the Syntactic test set, our proposed method consistently achieved higher scores compared to the baseline.

Testset	**Mtd**	**Pr**	**Ext.P**	**Re**	LKH
Battig	FCA	**51.0**	**81.7**	**64.4**	(37.0)
	BL	36.9	67.8	50.5	-
BLESS	FCA	**57.8**	**84.3**	**67.0**	(64.7)
	BL	50.5	74.5	57.5	-
Series	FCA	**62.8**	**82.7**	**76.8**	-
	BL	53.5	75.6	67.5	-
Syntactic	FCA	**57.1**	61.5	**56.3**	-
	BL	53.6	**61.8**	54.6	-

Table 5.2: Average precision (**Pr**), extended precision (**Ext.P**), and recall (**Re**) over the categories ($k = 3$), expressed as percentages. LKH lists % of the categories identified by [83]. BL=baseline

The use of optimal seeds in the evaluation is justified because the objective is to measure the extent to which a mathematically identified formal concept best matches categories provided in the test set. Other non-optimal seeds return different formal concepts, which indicate the heterogeneity of human-made categories in the test set. See Appendix A.5 for performance spread and a further discussion on the roles of seed words.

5.5.4 Analysis

The results suggest that formal concepts overlap with interpretable categories, which are defined as a set of words that human can descriptively label. Furthermore, the FCA method exhibited a notable advantage over the cosine similarity-based approach in concept retrieval. This is because the latter broadly identifies related words, whereas the former delves into specifying the underlying context. For example, given the seed words {*church, chapel*}, FCA additionally retrieves {*cathedral, shrine*}, emphasizing the context of "religious buildings." In contrast, the cosine method returns {*cathedral, catholic*} as output, failing to extract the feature of "buildings."

This advantage of FCA stems from its ability to locate mathematical structures within the matrix. Higher PPMI values discriminate the submatrix of a formal concept from its neighbors, forming a local plateau-like structure that is not necessarily captured by the cosine similarity (Eq. 5.7–5.9). This insight offers a use case for the proposed algorithm.

Disambiguating polysemy A target word can participate in multiple formal concepts. By inputting seed words with different associations, we found that polysemous words such as *tie* and *spring* have multiple formal concepts, as shown in Table 5.3. We observed that separate formal concepts (e.g., clothing, match, fasten) may contain the same word (e.g., *tie*) in their extents. Three separate plateaus may share the same row as visualized in Fig. 5.4.

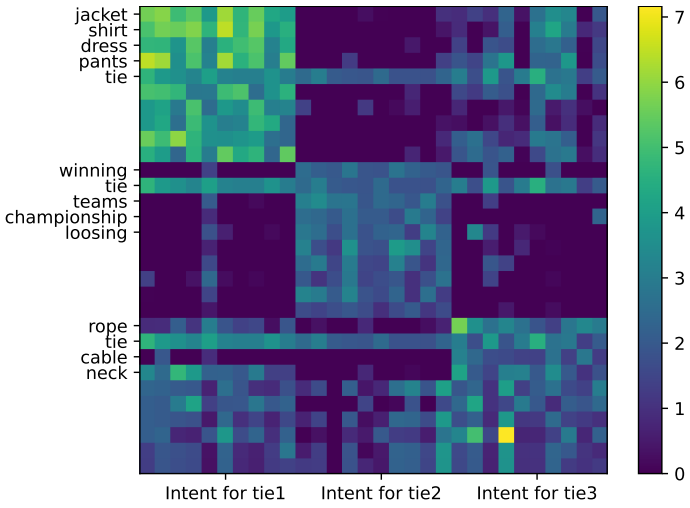


Figure 5.4: PPMI submatrix of three formal concepts containing the same polysemous word *tie*. For ease of visibility, the row for *tie* is presented multiple times.

[4] discovered that the embeddings of polysemous words can be decomposed as linear combinations of sense vectors. Our finding suggests that these vectors reflect separate formal concepts, and that the embeddings inherit the inner structure of the co-occurrence matrix.

Word (sense)	**Seed**	**Extent of Formal concept**
`tie1` (clothing)	*tie, pants, shirt*	*collar, jacket, pants, shirt, tie, wears*
`tie2` (match)	*tie, teams, winning*	*championship, playoffs, teams, tie, winning*
`tie3` (fasten)	*tie, cable, rope*	*cable, loose, neck, rope, tie*
`spring1` (season)	*spring, autumn, month*	*autumn, cold, coldest, cooler, dry, month, rainfall, ...*
`spring2` (metal)	*spring, wheel, suspension*	*fitted, mounted, rear, spring, suspension, wheel, wheels*
`spring3` (water)	*spring, creek, river*	*basin, brook, creek, reservoir, river, spring, stream*

Table 5.3: The extent of multiple formal concepts comprises polysemous words. The proposed algorithm is able to disambiguate these contexts in response to the seeds associated with them. The parameter k was set to 5 except for `spring1` ($k = 10$).

Measuring a similarity in subspace The proposed algorithm generates a byproduct that can be used to investigate the relationship between multiple vectors (the rows of the matrix) in a subspace. Reusing Lines 3–5 in Algorithm 3, we can determine whether target words in a seed share certain context words in limited dimensions and are semantically related in the shared context.

Specifically, we propose the **subspace similarity** $\phi(S)$ defined as

$$\phi(S) := \frac{1}{k} \sum_{i=1}^{k} m_{p(i)} \tag{5.10}$$

for a group of words $S = \{w_i\}_{i \in I_S}$, where $m_j := \min_{i \in I_S} X_{ij}$, $p(i)$ is a permuted index in descending order and k is a hyperparameter for the scope of subspace. The notation is the same as in Algorithm 3. The subspace similarity is the mean of the thresholds t determined over different parameters values up to k. Fig. 5.5 shows the computed values of the subspace similarity for several word groups. Semantically related groups show significantly higher values than the randomly chosen word group. These results indicate that semantically related groups share certain context words locally, even if their cosine similarities are low. Generally, randomly chosen vectors in high-dimensional space tend to be orthogonal, which implies a low chance of detecting correlations in any dimension. In contrast, a higher subspace similarity should suggest that a certain structure can be defined more than incidentally.

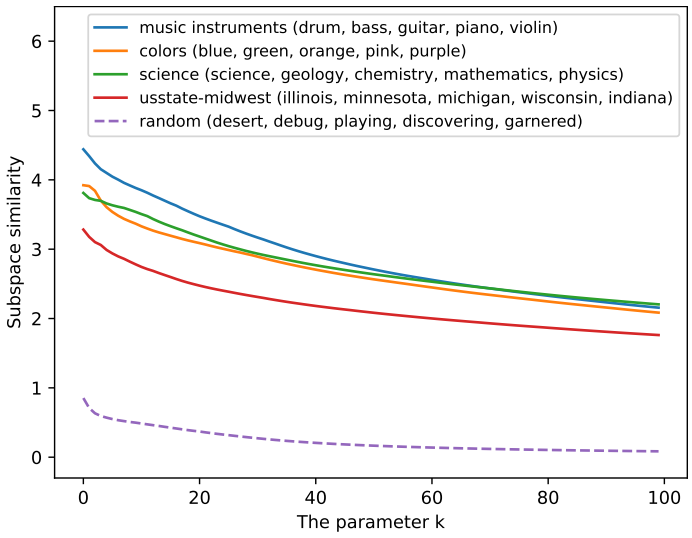


Figure 5.5: Subspace similarity for groups of five words. Semantically related groups exhibit significantly higher values than the randomly selected set.

5.6 Experiment 2: Applying Fuzzy FCA

5.6.1 Fuzzification of FCA

Our second application of FCA to a real-valued matrix involves the fuzzification of the crisp FCA by incorporating fuzzy set theory [10]. A fuzzy set formalizes an ambiguous set, such as "a set of tall people," by assigning a degree of membership to each element. In Appendix A.4, we give the definition of a fuzzy formal concept and show that it is equivalent to a rank-one submatrix under our proposed specification. Thus, the problem of finding fuzzy formal concepts can be regarded as that of identifying nonnegative rank-one submatrices in a PPMI matrix.

Because it is NP-hard to exactly decompose a matrix into nonnegative factors [139], we obtained an approximation by deploying nonnegative matrix factorization (NMF; [78]), as its L_1 regularization is considered effective in making them as sparse as possible. We controlled the sparseness so that the decomposed submatrices became disjoint. NMF decomposes $X \in R_+^{m\times n}$ into two matrices $W \in R_+^{m\times r}$ and $H \in R_+^{n\times r}$ so that $X = WH^T = \sum_{k=1}^{r} \boldsymbol{w}_k \boldsymbol{h}_k^T$, where $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$ are the kth columns of W and H, respectively. The outer product $\boldsymbol{w}_k \boldsymbol{h}_k^T$ is of rank one and preferably sparse, thereby approximating a fuzzy formal concept. The loss function is $\mathcal{L}_\alpha(W,H) = \frac{1}{2}\|X - WH^T\|_F^2 + \alpha\big(n\|W\|_1 + m\|H\|_1\big)$. We recursively applied NMF [3] over three rounds—first to the PPMI matrix as in Section 5.5, then twice to the positive residual matrices resulting from decomposition—factorizing into $r = 300$ components each round. Parameters for the $L1$ norms were set to $\alpha = 5, 3, 1 \times 10^{-4}$ for each round.

5.6.2 Results

We manually labeled 900 rank-one submatrices by reviewing the words corresponding to the largest entries in $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$ (see Appendix A.6.1 for details). We then classified the submatrices among four categories to assess how well the labels describe the words in each formal concept[4] (Table 5.4). Out of 900 acquired formal concepts, 95.7% were labeled

Class	**R1**	**R2**	**R3**	LKH
Descriptive	182	75	73	27
Partial	56	63	48	72
Meaningful	56	150	158	2
Nonsense	6	12	21	11
Total	300	300	300	112

Table 5.4: Decomposed rank-one submatrices in four classes for each round, indicating how the submatrices coincide with labeled categories. Definitions are provided in Appendix A.6.2 and the numbers under LKH are cited from [83].

descriptively or partially descriptively, or at least consisted of meaningfully related words.

[3] NMF from Scikit-learn library: BSD license.
[4] The same as the footnote 1.

5.6.3 Analysis

We found that Fuzzy FCA reveals the same formal concepts as the crisp FCA. For example, all categories listed in Table 5.2 also appear as rank-one submatrices. Of the 108 formal concepts identified in Experiment 1, 89 formal concepts (82.4%) were included by those found by the fuzzy method (Supplemental statistics in Appendix A.6.3). In fact, Fuzzy FCA detected more eligible words (e.g. *immense, massive* for Largeness, *shrine* for Religious Buildings). This observation demonstrates the robustness of FCA, as well as the correlation between the two methods.

Another interesting finding is that two types of rank-one submatrices were discovered: a *clique* type with identical rows and columns, and a *biclique* type with different rows and columns. An example of the latter is ({*explain, describe, discuss, ...*}, {*beliefs, concepts, ideas,...*}), which represents a verb phrase for an act of communication.

5.7 Discussion: Formal Concepts as Projections of MICs

Applying algebraic invariants to PMI formal concepts. Chapters 3 and 4 developed a theoretical and algorithmic framework in which compositional structure in language is represented as algebraic invariants of probability tensors. In the toric model of Chapter 3, symmetries in the combinatorial use of words induce relations among joint probabilities through the duality on the monomial mapping, which is constrained by vanishing binomials—including vanishing 2×2 minors. These binomials are parameter-independent and define *Minimal Invariant Constraints* (MICs): local rank-one structures in the underlying probability distribution that cannot be decomposed further within the given design. Chapter 4 then treated MICs as targets of structure learning, proposing an algorithm that selectively explores higher-order moments starting from lower ones. This demonstrated that pointwise mutual information (PMI) can be interpreted as a cumulant quantity that directly encodes pairwise interactions. Within this framework, the present chapter utilizes PMI matrices as observable projections of the probability tensor. We apply Formal Concept Analysis (FCA) to systematically extract and formalize those rank-one submatrices whose rows and columns form *formal concepts*. Our empirical results show that these formal concepts align with interpretable pairs of word groups, demonstrating that the algebraic invariants (MICs) described in Chapters 3 and 4 are indeed visible as structured components in real co-occurrence data.

Why rank-one PMI blocks align with linguistic structure: The role of MICs. The observation that rank-one blocks in the PMI matrix correspond to coherent linguistic groupings is not trivial; rather, it is a direct consequence of the underlying combinatorial structure of language. A formal concept, which is mathematically equivalent to a biclique in the co-occurrence graph, arises when two sets of words are recurrently used together under a shared context. Our proposed theory, developed in Chapter 3, provides the mathematical

mechanism for this phenomenon: the combinatorial usage of words induces invariant patterns in their joint probabilities. Specifically, if the joint distribution over sentences admits a local factorization—where two sets of words A and B are freely combinable under an appropriate shared context—then the conditional distribution over $(a, b) \in A \times B$ exhibits a rank-one structure. In the toric formulation of Chapter 3, such a factorization is captured by vanishing binomials and corresponds precisely to MICs. When the underlying probability tensor is marginalized to yield bigram distributions $P(w_i, w_j)$, the rank-one property of the MIC is preserved or closely approximated, provided that interactions with other variables factorize compatibly. Upon applying the PMI transformation, the resulting quantity can be viewed as a second-order cumulant under the moment lattice, capturing pure pairwise deviations from independence. Therefore, if an underlying (sub)tensor beneath a block $A \times B$ in $P(w_i, w_j)$ consists of these MICs, the corresponding PMI block naturally inherits a low-dimensional structure, realized as a rank-one (or nearly rank-one) submatrix. As our examples showed (cf., Section 5.5), the underlying structures captured by these formal concepts are not limited to simple *topical proximity*. They broadly reflect *functional proximity* (e.g., comparative constructions), *phrasal proximity* (e.g., verb phrases), and other *paradigmatic patterns* of word usage. Thus, the PMI-based formal concepts derived via FCA offer a powerful window into the inherent algebraic invariants of linguistic structure.

Relation to Generative Models of Co-occurrence. The low-rank structure of PMI matrices has also been explored from the perspective of *random-walk generative models*. Specifically, [3] proposes a model where co-occurrences are generated by shared latent states through a random-walk process, which naturally yields linearly structured embeddings. In their framework, the assumption of randomness under a shared latent context results in the desired linearity observed within PMI-based word embeddings. While our proposed theory also elucidates the emergence of linearity in word embeddings, it distinctively avoids assuming a random-walk process or latent states generated by randomness. Instead, we formalize the phenomenon using algebraic constraints. Our analysis shows that the rank-one blocks in the PMI matrix, which align with linguistic structure, are the consequence of MICs—parameter-independent, local rank-one structures captured by vanishing binomials in the probability tensor. Thus, whereas the generative model approach attributes the observed linearity (low-rank structure) to statistical randomness in latent factors, our framework provides a *deterministic, algebraic explanation*. This difference is significant: our approach rigorously formalizes *why* these structures exist in the language's combinatorial design itself, independent of specific generative assumptions. The alignment of Formal Concepts (bicliques) with interpretable linguistic groupings is therefore seen as a manifestation of these inherent algebraic invariants, rather than the result of a particular stochastic process.

Completeness, scalability, and higher-order structure. From the perspective of algebraic invariants, a fundamental question remains: to what extent do PMI-based formal concepts exhaustively capture the MICs underlying the full probability tensor, and can this process be systematized? While this chapter demonstrated two FCA-based methods for detecting interpretable rank-one submatrices, a complete theory unifying these procedures with the MIC framework of Chapters 3 and 4 is required.

Theoretically, MICs serve as the building blocks of complex invariant structures (Segre varieties) in the full probability tensor. Since fully exploring these structures in the high-dimensional tensor is computationally intractable, Chapter 4 proposed a strategy of marginalization: projecting the tensor onto pairwise co-occurrences to identify lower-order signatures (PMI) first. In this view, PMI formal concepts serve as anchors, allowing us to detect strong rank-one blocks in 2D matrices before selectively constructing higher-order slices to verify if the associated MICs extend to larger structures. To empirically validate this bottom-up verification strategy, we report an additional experiment in Appendix A.7. In this analysis, we reconstructed third-order co-occurrence tensors using word groups defined by the rank-one submatrices obtained via NMF. We confirmed that the slice matrices corresponding to these triplets also exhibit a rank-one property, thereby demonstrating that the invariant structures detected in the pairwise PMI matrix are indeed preserved in the higher-order tensor representation.

However, even within this simplified domain of rank-2 tensors (matrices), applying FCA to large-scale data presents distinct challenges. First, there is a scalability issue due to the combinatorial explosion of formal concepts as the vocabulary size grows ($N > 10,000$), even for pairwise matrices. Second, large corpora exhibit significant *heterogeneity*. In our experiments, interpretable formal concepts were detected at various threshold levels and factorization layers, suggesting that multiple latent formal contexts co-exist as if superposed—generated by distinct processes and producing rank-one submatrices that may be disjoint, overlapping, or hierarchically nested. Extracting such latent structures precisely requires algorithms sensitive to this generative heterogeneity. Consequently, a full characterization of the conditions under which PMI-based FCA efficiently recovers all underlying MICs, considering both the reduction from high-order tensors and the practical heterogeneity of large matrices, remains an open theoretical and practical challenge.

Implications for embeddings and distributional representations. Formal concepts, defined as rank-one submatrices, also appear as fundamental components in matrix factorizations of the form $X = WH^\top$ (Section 5.6). In such a factorization, a column of W corresponds to a (possibly fuzzy) set of words that constitutes a formal concept, whereas a row of W serves as a word embedding vector. A coordinate of a word in a given embedding dimension can thus be interpreted as the degree to which the word participates in the corresponding formal concept. The matrix H can be seen as encoding the attributes or contexts associated with these concepts. Word embeddings obtained by explicit matrix factorization or by implicit factorization via neural models [82] must therefore inherit the

structure of formal concepts. In light of the theory developed in Chapters 3 and 4, this suggests that at least part of the geometric organization observed in word embeddings can be traced back to MICs: invariant rank-one components in the underlying probability tensor that survive marginalization and become visible as rank-one PMI blocks, thereby mathematically justifying the effectiveness of PMI-based heuristics.

Practical benefits of FCA and links to contextualized models. Our experiments further indicate that FCA offers practical benefits over standard similarity-based methods in certain tasks. In the category completion task of Section 5.5, FCA-based methods outperformed cosine similarity. Both approaches capture relationships between words, but they differ in the subspace where similarity is assessed. Cosine similarity operates in the full embedding space and treats vectors as static entities, whereas FCA dynamically narrows the subspace based on a given set of words, identifying subvectors with significantly high co-occurrence. This ability to induce task-specific subspaces makes FCA particularly promising for tasks such as polysemy disambiguation and concept completion (e.g., [127]).

This perspective of treating latent subspaces allows FCA to serve as an analytical tool for interpreting *contextualized embeddings*. [37] showed that the internal representations of GPT-2 can be interpreted by projecting hidden states back into the vocabulary space. They reported actual pairs of words processed in intermediate layers, which resemble the formal concepts identified in our experiments. This suggests that contextualization in large language models might be viewed as performing algebraic operations on structures akin to MICs or formal concepts, although this connection is currently speculative. Hypothetically, the attention mechanisms and circuits within Transformer architectures may be *rediscovering* these latent invariant structures (MICs) encoded in word embeddings and manipulating them to reconstruct specific representations tailored to a context. While the present dissertation does not directly analyze Transformer internals, the results of this chapter indicate that PMI-based formal concept analysis provides a bridge between the algebraic invariants of probability tensors, distributional word representations, and the mechanisms of contextual computation in neural models, thus offering a principled starting point for future work.

5.8 Related studies

Several studies demonstrated that sparse embeddings are interpretable. [105] and [17] applied nonnegative matrix factorization with a sparsity constraint to word-document co-occurrence data and discovered topics. Other studies [47, 108, 65] investigated word embeddings to restore interpretability by using sparsity. We mathematically formalized the latent structure in the word co-occurrence matrix, which prior studies might have empirically detected.

FCA has been applied in linguistics [116], primarily for ontology. [30] applied FCA for the automatic acquisition of taxonomies from a corpus. [104] built a semantic structure

by setting the S-V-C tuples of the annotated corpora as a formal context. [16] used FCA by binarizing sparse word embedding for hypernymy discovery. In contrast to these studies, we deployed FCA to explore the structure of the matrix itself, which revealed the underlying structure of word-word co-occurrence matrices.

[54] delved into the underlying mechanism of word embeddings from a linguistic-philosophical perspective and pointed out simultaneous codetermination or *bi-duality* between terms and contexts as a significant feature of language, which we believe to have successfully formalized via FCA. Our mathematical approach to interpreting co-occurrence data may shed light on the structure of language, as [20] frames language in category theory.

Chapter 6

Conclusion and Future Work

6.1 Overview of the Chapter

This chapter concludes the dissertation and reflects on its broader implications for both linguistic theory and computational modeling. It summarizes the theoretical and empirical findings, clarifies the significance of the mathematical foundation proposed in this study, and explores how the developed framework may guide future research in natural language processing (NLP), large language models (LLMs), and mathematics.

6.2 Summary

Recap of problems and Research questions

- Language models still lack a principled mechanism for algebraic compositionality: although neural architectures combine tokens, they do not guarantee that these combinations correspond to systematic and interpretable operations in meaning space. The core problem is therefore to construct a language model in which compositionality functions as an explicit structural principle rather than an emergent artifact.
- These issues motivate the central research questions of how to formulate a probabilistic model that captures the algebraic and geometric structure present in linguistic data, and how to express linguistic regularities as mathematically defined objects. A further question is whether such a formulation can also account for structural phenomena observed in word embeddings and large language models.

Conclusion by chapters

- In Chapter 2, we survey how compositionality has been treated across symbolic, probabilistic, and neural paradigms, examining the extent to which each realizes structure-preserving mappings between syntax and semantics. The chapter identifies persistent limitations across these approaches and characterizes them as a

representation-theoretic inverse problem, thereby extracting the requirements of explicit structure, probabilistic consistency, and algebraic interpretability.

- In Chapter 3, the dissertation develops the theoretical foundation of the compositional probability model by formulating linguistic structure as an invariance property within a structured probability space. Language is represented as a probability tensor whose algebraic constraints—such as vanishing 2-minors and rank-one factorizations—define the Minimum Invariant Constraint (MIC) as the atomic unit of structure, connecting probabilistic independence with geometric objects such as Segre and Secant varieties. The chapter integrates algebraic statistics, Clifford algebra, and Walsh–Hadamard harmonic analysis to show that independence, factorization, rank-one structure, and vanishing minors are mathematically equivalent expressions of the same invariant relation. It further introduces the log-semiring to unify the linear structure of model parameters with the nonlinear geometry of probability space, thereby reconciling empirical linearities in embedding space with theoretical linearity in toric models. These constructions establish the algebraic and geometric principles that underlie the learning algorithms of Chapter 4 and the PMI-based analyses of Chapter 5.

- In Chapter 4, the dissertation operationalizes the compositional probability model by developing learning procedures that recover Minimum Invariant Constraints (MICs)—local rank-one structures characterized by vanishing 2×2 minors—directly from empirical probability tensors. The chapter introduces two complementary paradigms: a geometric approach that detects MICs through local minor inspection and its moment-based extensions exploiting marginalization, and a harmonic approach that transforms the tensor into the Walsh basis, where invariant relations appear as zero spectral coefficients. While the geometric method offers direct interpretability of local independence patterns, it suffers from combinatorial complexity, whereas the harmonic method scales nearly linearly and reveals global syzygies among invariants. Comparative experiments demonstrate that both approaches uncover consistent MIC patterns corresponding to compositional structure. Together, they establish a correspondence between geometry and algebra by showing that invariants defined as toric vanishing conditions can be efficiently recovered as linear spectral constraints.

- Chapter 5 examines the mathematical structure of the word co-occurrence matrix and shows that formal concepts extracted through Formal Concept Analysis (FCA) correspond to human-interpretable semantic categories and their hierarchical organization. By introducing binarized and fuzzy variants of FCA and validating them through category-completion experiments, the chapter demonstrates that these concepts recover meaningful regularities, including patterns of polysemy. This constitutes the first systematic application of FCA to co-occurrence data and provides

empirical grounding for the latent linguistic structures predicted by the theoretical framework. Crucially, the chapter interprets these rank-one structures as observable projections of Minimal Invariant Constraints (MICs), offering a deterministic algebraic explanation that contrasts with standard random-walk generative models. While this establishes a bridge between algebraic theory and empirical data, the precise characterization of all underlying invariants remains an open direction for future investigation.

Overall conclusion: How we solved for the questions

- **What we solved.** The dissertation establishes that linguistic structure can be represented as *algebraic and geometric invariants* embedded in probability distributions. By introducing the compositional probability model and defining Minimum Invariant Constraints (MICs) as local rank-one components, it demonstrates that independence, factorization, vanishing 2-minors, and Segre/Secant structures are mathematically equivalent descriptions of the same invariant relation. The learning algorithms of Chapter 4 and the empirical PMI analysis of Chapter 5 further show that these invariants can be recovered from real data, providing a principled and operational account of how compositional structure arises within probabilistic representations.

- **What we are yet to solve.** Two limitations remain beyond the scope of the present study. First, structural learning still faces computational challenges: exhaustive minor enumeration is infeasible at scale, and although spectral methods improve efficiency, a fully scalable algorithm for high-dimensional probability tensors has not yet been achieved. Second, the application of the proposed framework to linguistic data is only partially complete: while PMI matrices reveal rank-one patterns consistent with MIC-based invariants, a precise and systematic characterization of these structures—and their integration into large-scale linguistic modeling—remains an open direction for future work.

6.3 Limitation and Remaining Issues

Modeling Limitations and Future Directions. The demonstrative model employed in this dissertation was restricted to a simple $2\times2\times2$ probability tensor. Consequently, it does not yet fully capture the structural richness or dimensional complexity inherent in natural language distributions. A critical remaining issue is the formalization of word order and other syntactic phenomena that exhibit both commutative and non-commutative behaviors. Addressing this requires a more refined algebraic treatment, potentially informed by the geometric structure of Clifford algebra to capture these asymmetries. Furthermore, another open problem concerns the hypothesis that contextual effects underlying meaning behave as hidden random variables. A complete formulation that explicitly incorporates such latent contexts into the design matrix of a toric model remains to be fully developed.

Potential for modeling asymmetry and syntactic structures. This study constructs a probabilistic model based on the joint probability of words to formalize the combinatorial properties of language, which have rarely been treated explicitly in existing language models. By adopting the Toric model framework, where the contribution of each variable to the probability is defined via a configuration matrix, the current formulation naturally results in a bag-of-words representation, abstracting away sequential information. However, from the perspective of modeling linguistic compositionality, this abstraction was a strategic choice to focus on the formulation of combinatorial dependencies. A major contribution of this work is the demonstration that invariant algebraic structures (MICs) emerge purely from combinatorial properties, independent of word order.

Importantly, the Toric model framework does not inherently assume exchangeability among variables; the multivariate distribution is defined over an ordered set of variables, allowing for distinct parameterizations based on position. Consequently, linguistic asymmetry (e.g., the admissibility of VO versus the inadmissibility of OV in English) can be explicitly formulated by introducing non-commutative constraints or specific parameter settings that respect variable order. The challenge lies in distinguishing between exchangeable word combinations and order-sensitive ones within real data, and formulating the order accordingly.

Furthermore, syntactic tree structures can be interpreted as hierarchical partition structures within the multivariate joint probability. For example, a structure such as $(S, (V, O))$ can be modeled by decomposing the interaction terms into a coupling between V and O, combined with S. Mathematically, any such combination of variables or hierarchical constraints can be encoded as specific rows within the configuration matrix. Extracting these structured configuration matrices directly from empirical data remains a promising avenue for future work.

Challenges in Learning and Scalability. Regarding the learning process, the direct application of the proposed invariant-structure learning procedures to large-scale linguistic data remains challenging due to the sparsity, noise, and high dimensionality of empirical probability tensors. There is a pressing need for more efficient structural learning algorithms, as current methods either rely on computationally expensive exhaustive minor enumeration or utilize heuristic moment-based reductions that do not yet guarantee full coverage of MIC structures. Ultimately, the computational complexity of invariant detection remains a major bottleneck; even with the aid of spectral approaches, a fully scalable algorithm capable of handling high-dimensional probability tensors has yet to be achieved. Nevertheless, although the moment-based method currently lags in classification performance, its superior linear scaling makes it the most promising candidate for high-dimensional language modeling, provided that parameter tuning is further refined.

Applications and Interpretability. A complete algebraic translation of the hierarchical structure found in PMI matrices is not yet available, and the current framework only

partially explains how rank-one patterns relate to human-interpretable semantic categories. Consequently, a critical open issue is to refine the interpretation of these rank-one structures and nested bicliques within PMI matrices, specifically to clarify their connection to the semantics encoded in word embeddings within the algebraic framework. Beyond these structural interpretations, another remaining challenge is to bridge the proposed compositional probability model with the internal mechanisms of large language models. In particular, further research is required to determine whether invariant components analogous to MICs explicitly emerge within the attention mechanisms or feed-forward computations of modern neural architectures.

6.4 Future Work

Algebraic Structure in Large Language Models. We hypothesize that the internal representations within Large Language Models (LLMs) exhibit latent algebraic structures that align with the compositional probability framework proposed in this study. A promising direction for future work is to investigate whether the Attention and Feed-Forward layers operate as dual mechanisms for decomposition and reconstruction, effectively mirroring the algebraic principles developed in this dissertation. By analyzing the duality between meaning and context as a compositional isomorphism, we aim to identify these internal algebraic structures, thereby improving the interpretability of these complex models.

Foundational Algebra for Compositional Probability. Extending the mathematical foundations of the proposed framework is essential, particularly by refining the algebraic underpinnings grounded in Clifford algebra and semiring structures. Strengthening these foundations may also enhance the interpretability of machine learning algorithms by revealing the invariant components inherent in their computation. Ultimately, integrating these concepts via category theory offers a route toward a unified compositional formalism that seamlessly links algebraic, geometric, and probabilistic reasoning.

Constructing a Linguistic Clifford Algebra via Algebraic Twists. A central challenge in extending the current framework is to explicitly model the non-commutative structure of natural language, such as word order asymmetry, within the probabilistic formulation. As discussed, these asymmetries can be encoded structurally as specific row constraints within the configuration matrix of the Toric model. In this dissertation, we established a bridge between these probability structures and Clifford algebra by utilizing the Walsh transform. Crucially, the algebraic transition from the commutative algebra of Walsh functions to a non-commutative Clifford algebra is governed by a specific sign convention, mathematically formalized as a *2-cocycle* twist on the group algebra.

Currently, this 2-cocycle is chosen to yield a canonical Clifford algebra, a setting that remains mathematically generic. We hypothesize, however, that the "true" algebra of natural language is not generic but possesses a specific signature dictated by syntactic

constraints. Therefore, a promising direction for future work is to determine the optimal algebraic twist—and consequently, the specific multiplication rules—directly from the asymmetric dependencies observed in linguistic data. By identifying the correspondence between the row structures of the configuration matrix (which encode valid word combinations) and the generators of the algebra, we envision constructing a data-driven "Linguistic Clifford Algebra." This approach would provide a rigorous mathematical foundation where the non-commutativity of the algebra is not an arbitrary assumption, but a natural consequence of the intrinsic geometry of language.

Geometric Neural Networks with Clifford-Inductive Bias. A crucial future direction is to develop neural architectures that explicitly incorporate Clifford-algebraic inductive biases to capture invariant structure more directly. To support this, systematically discovering and cataloguing linguistic invariances is necessary to provide principled inductive biases for such geometric neural models. Once established, these architectures could be integrated into LLMs to improve generalization capabilities, particularly in small-data regimes where structural priors are most beneficial.

Toward a Mathematical Theory of Context. Future research should further develop the duality between row space and kernel space of the proposed model's configuration matrix as a mathematical counterpart to the duality between meaning and context, thereby identifying the hidden subspaces that govern compositional meaning. This perspective treats contexts as structural constraints that mediate between probability and meaning, significantly extending the compositional probability framework. Notably, this duality between context and meaning parallels mechanisms in error-correcting code theory, where the correspondence between the configuration matrix and the kernel in toric models serves as an intriguing algebraic analogue to encoder–decoder duality.

6.5 Concluding Remarks

This dissertation has presented a unified framework that views linguistic structure as a system of algebraic and geometric invariants embedded in probability distributions. Building on this perspective, it proposed a foundational model of language that integrates methods from algebraic statistics, geometric representation, harmonic analysis, and probabilistic modeling, thereby offering a compositional probabilistic architecture that addresses long-standing limitations in existing language models. While several issues remain—particularly regarding scalability, latent contextual structure, and large-scale application—the mathematical formulation developed here establishes a principled basis for analyzing language as a structured probabilistic system. It is our hope that this framework not only serves as a robust analytical tool for future research, but also provides new insights into the nature of linguistic structure across NLP, computational linguistics, and theoretical linguistics.

Appendix A

A.1 Toy corpus

The corpus contains 24 synthetic sentences shown in Table A.1. The target words—*king, queen, man, woman* and their plurals—are subjects of the sentences. Each of the eight words appears with three verbs—*live-in, wear, eat*—once for each. The remaining six words—*palace, house, tie, dress, alone, together*—discriminate the subject words so that they are in the analogical relations of three dimensions.

king live-in palace	*kings live-in palace*
queen live-in palace	*queens live-in palace*
man live-in house	*men live-in house*
woman live-in house	*women live-in house*
king wear tie	*kings wear tie*
queen wear dress	*queens wear dress*
man wear tie	*men wear tie*
woman wear dress	*women wear dress*
king eat alone	*kings eat together*
queen eat alone	*queens eat together*
man eat alone	*men eat together*
woman eat alone	*women eat together*

Table A.1: 24 sentences in the toy corpus

A.2 List of formal concepts

There are 28 formal concepts in the co-occurrence matrix derived from the toy corpus.

Suppose that the set of objects (target words) and the set of attributes (context words) be G, M respectively, defined as:

$$G = \{king, man, queen, queens, kings, men, queens, women\}$$
$$M = \{tie, dress, palace, house, alone, together\}$$

Then, all the formal concepts are identified as below:

$$T = (G, \emptyset)$$
$$f_1 = (\{king, man, kings, men\}, \{tie\})$$
$$f_2 = (\{man, woman, men, women\}, \{house\})$$
$$f_3 = (\{king, queen, man, woman\}, \{alone\})$$
$$f_4 = (\{kings, queens, men, women\}, \{together\})$$
$$f_5 = (\{king, queen, kings, queens\}, \{palace\})$$
$$f_6 = (\{queen, woman, queens, women\}, \{dress\})$$
$$e_1 = (\{king, man\}, \{tie, alone\})$$
$$e_2 = (\{king, kings\}, \{tie, palace\})$$
$$e_3 = (\{man, men\}, \{tie, house\})$$
$$e_4 = (\{kings, men\}, \{tie, together\})$$
$$e_5 = (\{king, queen\}, \{palace, alone\})$$
$$e_6 = (\{man, woman\}, \{house, alone\})$$
$$e_7 = (\{kings, queens\}, \{palace, together\})$$
$$e_8 = (\{men, women\}, \{house, together\})$$
$$e_9 = (\{queen, woman\}, \{dress, alone\})$$
$$e_{10} = (\{queen, queens\}, \{palace, dress\})$$
$$e_{11} = (\{woman, women\}, \{house, dress\})$$
$$e_{12} = (\{queens, women\}, \{dress, togther\})$$
$$v_1 = (\{king\}, \{tie, palace, alone\})$$
$$v_2 = (\{man\}, \{tie, house, alone\})$$
$$v_3 = (\{kings\}, \{tie, palace, together\})$$
$$v_4 = (\{men\}, \{tie, house, together\})$$
$$v_5 = (\{queen\}, \{dress, palace, alone\})$$
$$v_6 = (\{woman\}, \{dress, house, alone\})$$
$$v_7 = (\{queens\}, \{dress, palace, together\})$$
$$v_8 = (\{women\}, \{dress, house, together\})$$
$$B = (\emptyset, M)$$

A.3 Category completion test

We used the four test sets for the category completion test: Battig, BLESS, Series and Syntactic.

Battig test [25], originated from [9], contains 53 categories with 10 words for each category, of which we used 44 categories in the experiments, since the others have less than two words in our vocabulary of the co-occurrence matrix.

BLESS [8] contains 17 categories with 5-17, of which we used 12 categories for the same reason.

Both of Series and Syntactic are developed by the authors to supplement Battig and BLESS, which contain only common nouns. Series is hinted by [61] that proposed the series completion task (*penny:nickel:dime:?*) for word embeddings. Syntactic is motivated by our early finding that comparative adjectives such as *quicker, faster, ...* emerge as a salient formal concept with a high threshold in the binary FCA experiment. In both test sets, each category consists of 4 to 5 words, which are manually selected by one of the authors. In the development process, we partly use AI assistance[1] to generate a list of candidates for a category and its word set, by prompting with an example "Direction: *north, east, south, west*".

Examples of a category in each test set are shown below (Table A.2)

Test set	Category	Word set
Battig	Metal	*gold, iron, lead, steel,...*
BLESS	Fruit	*apple, banana, pear,...*
Series	Direction	*north, east, south, west*
Syntactic	Verb (go)	*go, goes, went, gone*

Table A.2: Examples of test sets

We used only the categories that contain more than or equal to three words in our vocabulary, which are listed in Table A.3.

A.4 Fuzzification of FCA

Formally, a fuzzy set A is a function $A : X \longrightarrow L$ where X is a ground set and $L = [0, 1]$, which assigns the value to each member of X. A subsumption relation $A \subseteq B$ holds if and only if $A(x) \leq B(x)$ for all $x \in X$. In Fuzzy FCA, a formal concept is $\mathbb{K} := (G, M, I, L)$. We consider two fuzzy sets $A \in L^G, B \in L^M$ as objects and attributes and a fuzzy relation $I \in L^{G \times M}$. Mathematically, L can be generalized to a *residuated lattice* that includes $[0, 1]$ as its special case. Similar to the crisp setting, two fuzzy derivation operators $\uparrow : L^G \longrightarrow L^M$

[1] https://chat.openai.com/

Battig	BLESS	Series	Syntactic
Disease	Ground mammal	Emotion	Demonstrative adverb
Metal	Furniture	Season	Comparative adjective
Carpenter's tool	Tool	Sea	Preposition
Crime	Container	Great lakes	Verb conjugation
Substance for flavoring food	Fruit	Direction	Manner adverb
Elective Office	Vehicle	Art form	Adverb of frequency
Toy	Appliance	Part of a tree	Personal pronoun
Weapon	Weapon	Book part	Linking verb
Member of clergy	Musical instrument	Continent	Demonstrative determiner
Four-footed animal	Building	Movie genre	Coordinating conjunction
Nonalcoholic beverages	Clothing	Number	Adjective of taste
Building for religious services	Bird	US president	Possessive pronoun
Precious stone		Stage of life	Frequency adverb
Part of human body		Planet	Quantitative determiner
Fruit		Weekday	Subordinating conjunction
Sport		Music genre	Action verb
Part of a building		Natural disaster	Modal auxiliary
Male's first name		Decathlon	Total pronoun
Relative		Family	Adjective of size
Human dwelling		Ocean	Interrogative pronoun
Insect		Adverb of time	Article
Type of fuel		Month	Totality adverb
Music instrument		Communication act	Verb conjugation
Furniture		Match	
Ship		Religion	
Kind of money		Time of day	
Color		Writing	
Kind of cloth		Style of architecture	
Unit of distance		Midwest U.S. state	
Type of music			
City			
Country			
Reading material			
Military title			
Natural earth formation			
Unit of time			
Part of speech			
Kitchen utensil			
Vehicle			
Science			
Weather phenomenon			
Occupation or profession			
Bird			

Table A.3: Used categories of the test sets

and $\downarrow : L^M \to L^G$ are defined as follows: For all $m \in M$ and $g \in G$,

$$A\uparrow(m) := \bigwedge_{g \in G} \big(A(g) \to I(g,m)\big) \in L \tag{A.1}$$

$$B\downarrow(g) := \bigwedge_{m \in M} \big(B(m) \to I(g,m)\big) \in L \tag{A.2}$$

Note that $A\uparrow \in L^M, B\downarrow \in L^G$ and $(\to) : L \times L \to L$, which is a binary operation defined on L. In plain English, the degree to which an object g belongs to the fuzzy set A should imply the level of co-occurrence between g and an attribute m, which retrospectively should determine the degree to which the attribute m belongs to another fuzzy set $A\uparrow$. Then, fuzzy formal concepts are defined as a pair of fuzzy sets (A, B) where $A\uparrow = B$ and $B\downarrow = A$ hold as in the crisp FCA.

We need to specify operations such as $(\to)$ to numerically compute them. Three specifications, named as Lukasiewicz, Gödel and Goguen, have already been proposed [11], but instead we propose our own specification tailored to the analysis of a word co-occurrence matrix.

$$a \to b := \begin{cases} b/a & \text{if } a > 0 \\ \top & \text{if } a = 0 \end{cases} \tag{A.3}$$

where $\top$ is the greatest element in L. This specification is a slight modification of the one proposed by Goguen. The meet $\wedge$ is numerically calculated as a minimum.

Our specification is equivalent to defining $(A, A\uparrow)$ and $(B\downarrow, B)$ as a rank-one submatrix. Recall that the fuzzy set $A \in L^G$ assigns a value $x \in L$ to the element g. Similarly, the fuzzy set $A\uparrow \in L^M$ assigns a value $y \in L$ to the element m. Thus, the specification in Eq. (A.3) ensures that $y = xy/x$ for $x > 0$. This means that nonnegative entries in both fuzzy sets $A, A\uparrow$ constitute a rank-one submatrix.

A.5 Role of seed words and performance spread

In Algorithm 3 (Section 5.5), seed words are required as input to the algorithm since a formal concept is detected as a closed set containing those seed words. A closed set is the fixed point of a closure operator. In this algorithm, any set of words can be a seed. If seed words are randomly chosen, then the algorithm will find and return a formal concept in an unsupervised way. Different sets of words return the same formal concept if all of the used words belong to the same closed set, and derive different formal concepts otherwise due to the mathematical property of a closure operator and closure system.

In the category completion test in Experiment 1, all possible pairs from the word list of the same category are used as seeds (2 words) to derive a formal concept. Therefore, different formal concepts can be identified by a chosen pair of words from the test set.

Table A.4 shows the statistics of the distribution of recalls calculated for all combinations of two words, reflecting the degree to which the FCA retrieved elements from the test set based on seed choices. To verify, cross-reference the numbers in the Max column with the corresponding recalls reported in Table 5.2 where the optimal seed pairs were selected. In cases where a word set within a specific category in the test set comprises 10 words, there exist ${}_{10}C_2 = 45$ possible pairs. We computed the minimum, maximum, and median values for each category, subsequently averaging them between categories for each test set and the entire dataset.

These statistics suggest that employing a "right seed" (optimal pair) results in the formal concept covering 56.3% (Syntactic) to 76.8% (Series). On the contrary, the use of a different seed may yield a distinct formal concept. This divergence can be attributed to the non-cohesive nature of word groups within the test set.

For instance, the "Occupation or profession" category of Battig comprises words such as *doctor, lawyer, teacher, engineer, professor, carpenter, salesman, nurse, psychologist* (with one word omitted due to limited vocabularies in the matrix). Notably, the FCA found that the maximum formal concept within this category is only four words: *lawyer, nurse, psychologist, teacher*, which seem to represent the "profession" part of the category.

Test set	Min	Max	Median
Battig	33.8	64.4	37.3
BLESS	36.5	67.0	43.9
Series	49.2	76.8	53.6
Syntactic	46.1	56.3	48.6

Table A.4: Spread % of Recall over different choice of seeds

A.6 Decomposition by NMF

A.6.1 Decomposed submatrices by NMF

We applied NMF recursively in three rounds. In the first round, we decomposed the PPMI matrix as in $X_0 \approx W_1 H_1^T$ into 300 components ($\alpha = 0.0005$). In the second round, we applied NMF to the positive residual matrix after the first decomposition: $X_1 := \max(X_0 - W_1 H_1^T, 0)$ as decomposed as in $X_1 \approx W_2 H_2^T$ ($\alpha = 0.0003$). In the third round, the residual matrix $X_2 := \max(X_1 - W_2 H_2^T, 0)$ was decomposed into $X_2 \approx W_3 H_3^T$ ($\alpha = 0.0001$). Note that each component (rank-one matrix) $\boldsymbol{w}_k \boldsymbol{h}_k^T$ was forced to be sparse by L_1 regularization. Thus, their nonnegative rows and columns make a nonnegative rank-one submatrix, which we regard as a fuzzy formal concept.

The components derived in the first round were indexed from 1 to 300. Similarly, those in the second round were indexed from 301 to 600, and ones in the third round were indexed from 601 to 900. We ordered each component according to the Frobenius norm within each round. Therefore, the smaller ID number implies that the submatrix has a greater norm in each round.

Samples of the components are presented in Table A.8. The class was evaluated by one of the authors according to the definition given in Appendix A.6.2. The author also labeled a category from the words that comprise the submatrix $\boldsymbol{w}_k \boldsymbol{h}_k^T$. More specifically, for each vector $\boldsymbol{w}_k$ and $\boldsymbol{h}_k$, we picked 20 words that correspond to the largest elements in the vectors, respectively. In Table A.8, the only four top words are presented for both $\boldsymbol{w}_k$ as extents and $\boldsymbol{h}_k$ as intents. For ease of visibility, categories were labeled with more general expressions, although they could be labeled with more focused category names.

Table A.5 shows a supplemental analysis of the type of relatedness between words participating in each submatrix.

Proximity	**R1**	**R2**	**R3**
Categorical	74	64	53
Contextual	171	147	148
Combinatorial	41	59	62
Syntactic	9	18	19
None	5	12	18
Total	300	300	300

Table A.5: Proximity types of word relations in each NMF-decomposed component. Categorical: words are in the same category, Contextual: words are related in a shared context, Combinatorial: words are a part of possible phrases, i.e., paradigmatic, Syntactic: words are in the same syntactic category.

A.6.2 Types of qualitative classes

The set of words corresponding to the largest dimensions within each component is classified into four qualitative classes, as in the below definition (Table A.7), following

lindh2015exploratory . These classes indicate how well an identified formal concept (a rank-one matrix) is interpretable as a category.

A.6.3 Overlap of two FCA methods

In addition, we performed an analysis of set overlap at the word level. For each of the 89 groups, we calculated the set overlap using the Jaccard index, which is defined as the number of words in the intersection divided by the number in the union. The results are presented in Table A.6 as percentages.

Min	Max	Mean	Median
1.4	64.5	23.7	20.0

Table A.6: Jaccard index between the corresponding formal concepts of Binarization method and Fuzzy method over 89 categories

Class	Description
Descriptive	Words are related in some way, and the majority of the labels given are as descriptive as possible of the words in the set.
Partial	Words are related in some way, and the majority label is somewhat descriptive, but a more descriptive account can be easily given.
Meaningful	Words are related, but no majority label describes the words.
Nonsense	There is no majority label, nor is there any perceived relation between the words in the set.

Table A.7: Definition of qualitative classes assessing how well the labels describe the words in each formal concept. [83]

A.7 Relationship between PMI rank-one structures and MICs

While the rank-one structure in the Pointwise Mutual Information (PMI) matrix suggests the existence of Minimum Invariant Constraints (MICs) as proposed in Chapter 3, we conducted an additional analysis using data from a real corpus. The purpose of this analysis is to demonstrate more directly that the PMI matrix serves as a manifestation of the marginalization of MICs.

A.7.1 Objective

The objective of this analysis is to empirically demonstrate that the rank-one submatrices observed in the PMI matrix correspond to marginalization of rank-one structures residing within higher-order (specifically, 3rd-order) tensors.

A.7.2 Experimental Procedure

Data and Preprocessing We utilized the English Wikipedia Dump dated 20171001, consistent with the dataset used in Sections 5.5 and 5.6.

ID	Class	Category	**Extents** (top 4 words)	**Intent** (top 4 words)
2	D	Geography	*iran, kerman, khorasan, province*	*iran, kerman, khorasan, province*
5	N	None	*pineapples, tasteful, lilongwe, unimpressive*	*dawn, windsor, batting, relegation*
8	D	Music	*chart, charts, billboard, singles*	*chart, charts, billboard, singles*
14	D	Sports	*discus, javelin, jump, hurdles*	*discus, javelin, jump, hurdles*
22	D	Education	*degree, bachelor, doctorate, laude*	*degree, bachelor, doctorate, laude*
35	D	Diplomacy	*embassy, ambassador, diplomatic, relations*	*turkmenistan, tajikistan, kyrgyzstan, uzbekistan*
46	D	Sports	*baseman, pitcher, outfielder, shortstop*	*baseman, pitcher, outfielder, shortstop*
89	D	Religion	*rabbi, yeshiva, synagogue, hebrew*	*rabbi, yeshiva, synagogue, hebrew*
90	D	US states	*idaho, montana, dakota, wyoming*	*idaho, montana, dakota, wyoming*
95	D	Climates	*cyclone, hurricane, storm, typhoon*	*cyclone, hurricane, storm, typhoon*
98	D	Politics	*polling, votes, voters, vote*	*polling, votes, voters, vote*
102	D	Phrases	*increases, decreases, decrease, increase*	*temperature, concentrations, accuracy, velocity*
104	D	Politics	*incumbent, reelection, democrat, republican*	*incumbent, reelection, democrat, republican*
116	P	Medical	*ligament, knee, ankle, injury*	*ligament, knee, ankle, injury*
125	P	Career	*postdoctoral, professor, adjunct, emeritus*	*postdoctoral, professor, adjunct, emeritus*
137	P	TV show	*starring, roommate, daughters, actress*	*jennifer, laura, jessica, nicole*
146	P	Legal	*convicted, guilty, sentenced, imprisonment*	*convicted, guilty, sentenced, imprisonment*
147	P	History	*nazi, nazis, deported, camps*	*nazi, nazis, deported, camps*
159	P	Geography	*mountain, peaks, summit, mountains*	*mountain, peaks, summit, mountains*
160	M	Expression	*acclaim, garnered, reviews critical*	*garnered, acclaim, reviews, critical*
165	P	Expression	*regain, recover, conquer, attract*	*trying, attempting, attempt, attempts*
181	M	Expression	*tasked, thereby, prevented, intention*	*securing, obtaining, capturing, creating*
184	M	Expression	*lied, intentions, poisoned, whereabouts*	*reveals, realizes, believing, realises*
192	D	Music	*punk, hop, hip, folk*	*punk, hop, hip, folk*
210	D	Religion	*christianity, catholicism, islam, beliefs*	*christianity, catholicism, islam, beliefs*
212	M	Expression	*you, think, really, know*	*you, think, really, know*
214	M	Adjective	*various, numerous, several, these*	*genera, disciplines, locations, dialects*
237	D	Comparative	*faster, stronger, heavier, than*	*faster, stronger, heavier, than*
239	D	Politics	*obama, barack, reagan, clinton*	*obama, barack, reagan, clinton*
313	D	Unit	*quantities, amounts, sums, amassed*	*enormous, huge, immense, considerable*
329	P	Time	*spends, spend, spent, spending*	*summers, much, time, remainder*
341	P	Geography	*maui, oahu, hawaii, honolulu*	*maui, oahu, hawaii, honolulu*
370	D	Unit	*millions, billions, million, billion*	*millions, billions, million, dollars*
405	M	Linguistics	*vowel, vowels, stressed, accent*	*vowel, vowels, stressed, accent*
408	P	Travel	*immigration, nationals, emigration, citizen*	*immigration, nationals, emigration, citizen*
419	D	Expression	*proposal, offer, invitation, plea*	*rejected, accepted, rejects, accepting*
431	M	Adverb	*poorly, properly, carefully, fully*	*handled, treated, understood, trained*
435	D	Buildings	*housed, built, constructed, build*	*synagogue, mosque, mansion, convent*
484	D	Auxiliary	*did, does, doesn, didn*	*speak, exist, suffer, appear*
507	D	Movement	*down, forth, out, into*	*fell, put, falling, fallen*
514	D	Weapon	*pistol, revolver, magnum, rifle*	*pistol, revolver, magnum, rifle*
517	M	Plants	*botanical, zoological, garden, gardens*	*botanical, zoological, garden, gardens*
577	P	Adverb	*totally, completely, virtually, almost*	*totally, virtually, completely, vanished*
605	D	Number	*vii, ix, viii, xiii*	*fantasy, corps, intensity, chapter*
626	P	Accounting	*collect, collecting, exception, collected*	*taxes, debt, debts, fees*
645	D	Month	*june, july, august, september*	*premiered, consecrated, baptised, inaugurated*
667	D	Expression	*taking, take, taken, takes*	*hostage, advantage, seriously, refuge*
669	D	Geography	*gaza, palestinians, palestinian, israeli*	*strip, gaza, rockets, barrier*
679	D	Geography	*colombian, venezuelan, peruvian, chilean*	*peso, divisi, primera, aut*
774	P	IT	*java, server, windows, software*	*java, server, windows, software*
781	D	Expression	*bought, purchased, buying, buy*	*shares, stake, tickets, tracts*
784	M	Marketing	*advertising, commercials, campaigns, marketing*	*advertising, commercials, campaigns, marketing*
804	M	Expression	*about, detail, matters, topics*	*discuss, discussed, discussing, discusses*
855	M	Expression	*heavily, originally, by, recently*	*influenced, inspired, invented, borrowed*
864	D	Expression	*currently, presently, still, today*	*currently, resides, owns, produces*
874	P	Expression	*launching, pursued, launched, developed*	*ventures, venture, scheme, initiative*

Table A.8: Samples of decomposed submatrices labeled with a category name. Classes are abbreviated; D:Descriptive, P:Partial, M:Meaningful, N:Nonsense

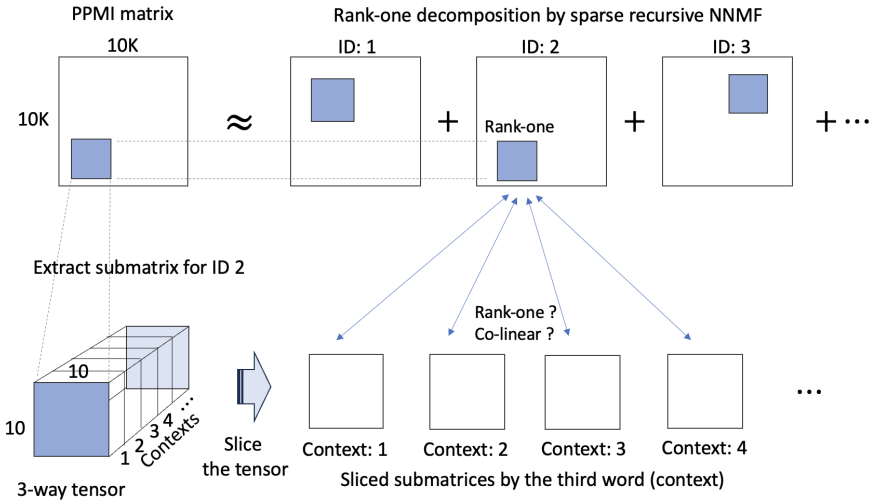


Figure A.1: Construction of 3rd order tensors.

Construction of 3rd-Order Tensors The procedure for constructing the tensors is as follows (See Fig.A.1):

1. **Selection of Word Pairs:** Following the methodology in Section 5.6, we used the rank-one matrices extracted via Sparse Recursive Non-negative Matrix Factorization (NMF), denoted as $X \approx WH^T$. For each column vector of W and H, we identified the corresponding word groups by selecting the top 10 components or components accounting for a cumulative 80% of the weight (up to a maximum of 10 words). This process yielded a set of word pairs, with a maximum of $10 \times 10 = 100$ pairs per rank-one component.

2. **Triplet Extraction:** For each of the 100 word pairs, we formed triplets by counting the co-occurrences of a third word within a window of up to 15 words surrounding the pair. This expansion is consistent with the original window size of 5, accounting for the combined span required to encompass three words.

3. **Tensor Construction:** We constructed a 3rd-order tensor using these triplets as elements. To determine the third dimension (context words) of the tensor, we calculated the sum of all elements within each 10×10 slice corresponding to a specific context word. We then selected the top 100 context words with the largest sums. This resulted in a tensor of size $10 \times 10 \times 100$.

Sparse Recursive NMF Decomposition The NMF rank was set to $r = 1000$, resulting in the analysis of 1000 tensors of size $10 \times 10 \times 100$. We employed a sparse recursive NMF approach, wherein the decomposition was performed recursively in 10 partitions. Specifically, the initial matrix X_1 was decomposed into $r = 100$ components ($X_1 \approx W_1 H_1^T$) using an initial L_1 sparsity parameter. The residual was then updated as $X_2 = \max(X_1 - W_1 H_1^T, 0)$

where the max applies elementwise, and the NMF process was applied iteratively to X_2 with a gradually decreasing L_1 parameter value, repeating this procedure for subsequent steps.

We adopted this sparse recursive approach because preliminary experiments indicated it extracted the most interpretable word groups. This observation suggests that the rank-one submatrices constituting the global structure exist at varying scales of magnitude, requiring an iterative approach to capture both dominant and subtle signals effectively.

A.7.3 Analysis and Results

Qualitative Analysis As demonstrated in Section 5.6, the top-10 words identified via NMF of the PPMI matrix for the first two modes (referred to as u-words and v-words) in each component form interpretable clusters sharing a common concept or context. We extend this analysis by incorporating the third mode, referred to as w-words (context words relative to the two word pairs), to examine the structure of word triplets (u, v, w). Focusing on the top-100 w-words for each component, we construct a sub-tensor of size $10 \times 10 \times 100$, which consists of 100 slice matrices of size 10×10. From these candidate slices, we selected the most representative slices based on the sum of the 3rd-order PMI values within each slice matrix. To conserve space, we display w-words that are distinct from the u- and v-words in the subsequent results.

Tables A.9 and A.10 present representative examples of the triplet groups extracted via this 3rd-order tensor decomposition. Table A.9 illustrates paradigmatic relationships, where the sets of u-words and v-words are identical or semantically equivalent. In contrast, Table A.10 displays syntagmatic relationships, where u-words and v-words differ, capturing co-occurrence patterns or grammatical structures (e.g., verb-object or subject-verb dependencies).

Table A.9: Examples of Triplets: Paradigmatic Relations

	ID: 0004	ID: 0580
u-words	*techniques, methods, modeling, testing, tools,...*	*immigration, asylum, refugee, refugees, migration,...*
v-words	*techniques, methods, modeling, testing, tools,...*	*immigration, asylum, refugee, refugees, migration,...*
w-words	*analysis, software, tools, data, evaluation,...*	*detentions, camps, illegal, persons, ...*

Table A.10: Examples of Triplets: Syntagmatic Relations.

	ID: 0127	ID: 0270
u-words	*marry, sell, join, abandon, settle, ...*	*lasted, lengthy, subsequent, interrupted, during,...*
v-words	*decided, persuaded, tried, attempted, decides...*	*negotiations, feud, excavations, deployment, filming,...*
w-words	*to, him, but, he, her,...*	*during, after, the, in, and,....*

These context words (w-words) provide contextual grounding for the u-v pairs. In syntagmatic relations, specifically, the triplets constitute phrases that share identical syntactic structures and exhibit coherent semantic affinities. These observations corroborate the consistency of our findings, indicating that the 3rd-order rank-one tensor components

capture semantically interpretable relations, just as observed in the 2nd-order PMI matrix analysis. Furthermore, this structural alignment between the 3rd- and 2nd-order rank-one components supports our hypothesis that the algebraic invariant constraints derived from MICs are preserved through marginalization and manifest in the lower-order matrix.

Quantitative Analysis To quantitatively evaluate the relationship between the rank-one structure in the 2nd-order matrix and in the 3rd-order tensor, we conducted a comprehensive analysis using the entire dataset of slices. Specifically, we aggregated $99,209$ slice samples (derived from up to 100 slice matrices for each of the $1,000$ global components). For each slice, we computed two metrics: (1) the *Sigma Ratio*, which measures the dominance of the primary singular value (representing the degree of rank-one structure) of the slice matrix, and (2) the *Global Slice Similarity*, defined as the cosine similarity between the slice's principal singular vectors ($\mathbf{u}, \mathbf{v}$) and the corresponding global factors ($\mathbf{w}, \mathbf{h}$) obtained from the NMF decomposition of the PMI matrix. Hereafter, we refer to ($\mathbf{w}, \mathbf{h}$) as the *global* factors to distinguish them from the *local* singular vectors of the individual slices.

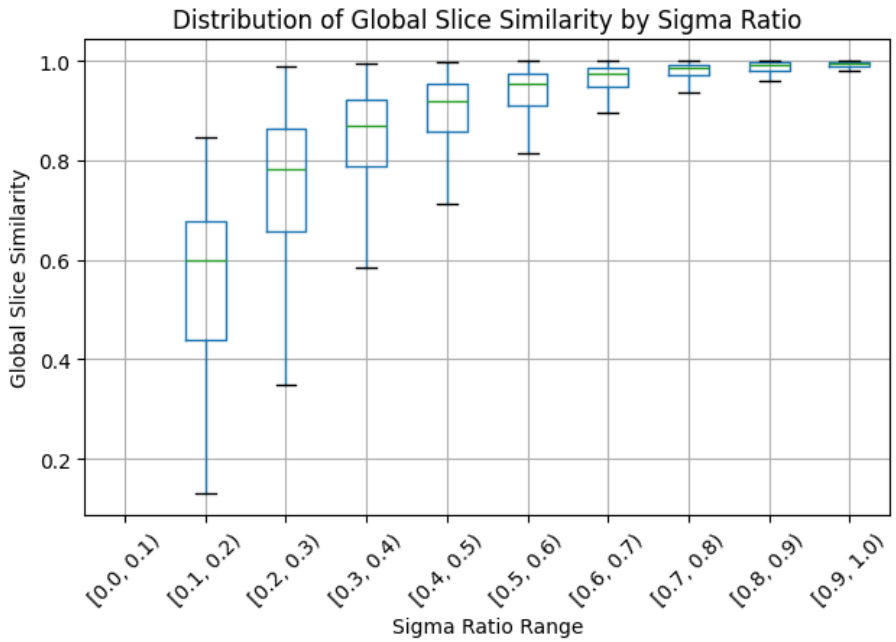


Figure A.2: **Distribution of Global Slice Similarity conditioned on Sigma Ratio.** The box plots illustrate the distribution of cosine similarities between local slice components and global factors, grouped by the dominance of the slice's first singular value (Sigma Ratio). The trend demonstrates that as the local structure approaches a pure rank-one state (Sigma Ratio $\rightarrow 1.0$), the local components consistently align with the global semantic vectors, supporting the hypothesis of structural consistency across local and global scales.

Figure A.2 reveals a significant monotonic relationship between local structural purity and global alignment. As illustrated, slices with a high Sigma Ratio (indicating a near rank-one structure) exhibit a Global Slice Similarity converging to 1.0 with vanishing variance. This result implies that when a local context slice is decomposed into a rank-one matrix

$\mathbf{u}_k \mathbf{v}_k^T$ with high spectral dominance, its components $\mathbf{u}_k$ and $\mathbf{v}_k$ align almost perfectly with the global factors $\mathbf{w}$ and $\mathbf{h}$ derived from the decomposition of the PMI matrix.

This finding suggests that the global rank-one component $\mathbf{w}\mathbf{h}^T$ is not merely an abstract approximation but is substantiated by locally coherent structures. In other words, the global semantic structure effectively *emerges* from the marginalization of these rigorous local rank-one patterns. Consequently, the decomposition of the PMI matrix can be interpreted as an aggregation of context-specific rank-one structures that share a consistent directionality in the semantic space.

A.7.4 Discussion

The analysis demonstrated that the rank-one submatrices observed in the PMI matrix are underpinned by identical rank-one structures in their tensor slices. By recursively extending this logic, it is theoretically hypothesized—via inductive implication—that Minimum Invariant Constraints (MICs), defined as irreducible vanishing binomial relations, exist within the original probability tensor. These empirical results align with the theoretical model proposed in Chapter 3, suggesting that rank-one structures appear in the PMI matrix as a result of marginalizing these stacked MIC structures.

Conversely, cases where marginalizing different slices results in rank-one structures (potentially reflecting linguistic phenomena such as polysemy) present an intriguing avenue for analysis but remain a subject for future research.

Bibliography

[1] Rafal Ablamowicz. On the structure theorem of clifford algebras. *arXiv preprint arXiv:1610.02418*, 2016.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to PMI-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

[4] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495, 2018.

[5] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International conference on learning representations*, 2017.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] Marco Baroni, Raffaella Bernardi, Roberto Zamparelli, et al. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in language technology*, 9:241–346, 2014.

[8] Marco Baroni and Alessandro Lenci. How we BLESSed distributional semantic evaluation. In Sebastian Pado and Yves Peirsman, editors, *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 1–10, Edinburgh, UK, July 2011. Association for Computational Linguistics.

[9] William F Battig and William E Montague. Category norms of verbal items in 56 categories a replication and extension of the connecticut category norms. *Journal of experimental Psychology*, 80(3p2):1, 1969.

[10] Radim Belohlavek. A note on variable threshold concept lattices: threshold-based operators are reducible to classical concept-forming operators. *Information Sciences*, 177(15):3186–3191, 2007.

[11] Radim Belohlavek and Vilem Vychodil. Formal concept analysis and linguistic hedges. *International Journal of General Systems*, 41(5):503–532, 2012.

[12] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[13] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[14] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[15] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[16] Gábor Berend, Márton Makrai, and Péter Földiák. 300-sparsans at SemEval-2018 task 9: Hypernymy as interaction of sparse attributes. In Marianna Apidianaki, Saif M. Mohammad, Jonathan May, Ekaterina Shutova, Steven Bethard, and Marine Carpuat, editors, *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 928–934, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[17] Michael Biggs, Ali Ghodsi, and Stephen Vavasis. Nonnegative matrix factorization via rank-one downdate. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 64–71, New York, NY, USA, 2008. Association for Computing Machinery.

[18] William Blacoe and Mirella Lapata. A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 546–556, 2012.

[19] Craig Boutilier, Nir Friedman, Moises Goldszmidt, and Daphne Koller. Context-specific independence in bayesian networks. *arXiv preprint arXiv:1302.3562*, 2013.

[20] Tai-Danae Bradley, Juan Luis Gastaldi, and John Terilla. The structure of meaning in language: parallel narratives in linear algebra and category theory. *Notices of the American Mathematical Society*, 71(2), 2023.

[21] Richard Brauer and Hermann Weyl. Spinors in n dimensions. *American Journal of Mathematics*, 57(2):425–449, 1935.

[22] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[24] Richard A Brualdi, Herbert John Ryser, et al. *Combinatorial matrix theory*, volume 39. Springer, 1991.

[25] John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods*, 44:890–907, 2012.

[26] Eugene Charniak. Statistical parsing with a context-free grammar and word statistics. *AAAI/IAAI*, 2005(598-603):18, 1997.

[27] Wei-Chen Cheng, Stanley Kok, Hoai Vu Pham, Hai Leong Chieu, and Kian Ming Adam Chai. Language modeling with sum-product networks. In *Interspeech*, volume 2014, pages 2098–2102, 2014.

[28] Giampiero Chiaselotti, Davide Ciucci, and Tommaso Gentile. Simple undirected graphs as formal contexts. *Formal Concept Analysis: 13th International Conference, ICFCA 2015, Nerja, Spain, June 23-26, 2015, Proceedings 13*, pages 287–302, 2015.

[29] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.

[30] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of artificial intelligence research*, 24:305–339, 2005.

[31] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała,

Yonatan Belinkov, and Dieuwke Hupkes, editors, *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[32] Stephen Clark. Vector space models of lexical meaning. *The Handbook of Contemporary semantic theory*, pages 493–522, 2015.

[33] Stephen Clark and James R Curran. Wide-coverage efficient statistical parsing with ccg and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

[34] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*, 2010.

[35] Michael Collins. Head-driven statistical models for natural language parsing. *Computational linguistics*, 29(4):589–637, 2003.

[36] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[37] Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170, Toronto, Canada, July 2023. Association for Computational Linguistics.

[38] Brian A Davey and Hilary A Priestley. *Introduction to lattices and order*. Cambridge university press, 2002.

[39] A Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 1979.

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[41] Persi Diaconis and Bernd Sturmfels. Algebraic algorithms for sampling from conditional distributions. *The Annals of statistics*, 26(1):363–397, 1998.

[42] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*. Springer, 2009.

[43] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

[44] Heinz Werner Engl, Martin Hanke, and A Neubauer. *Regularization of Inverse Problems*, volume 375. Springer Science & Business Media, 1996.

[45] Pavel I Etingof, Oleg Golberg, Sebastian Hensel, Tiankai Liu, Alex Schwendner, Dmitry Vaintrob, and Elena Yudovina. *Introduction to representation theory*, volume 59. American Mathematical Soc., 2011.

[46] Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. Assessing composition in sentence vector representations. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[47] Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. Sparse overcomplete word vector representations. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China, July 2015. Association for Computational Linguistics.

[48] Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-Jussà. A primer on the inner workings of transformer-based language models. *arXiv preprint arXiv:2405.00208*, 2024.

[49] Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[50] Gottlob Frege. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50, 1892.

[51] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. Learning semantic hierarchies via word embeddings. In Kristina Toutanova and Hua Wu, editors, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1199–1209, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[52] Bernhard Ganter and Rudolf Wille. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[53] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. W. H. Freeman, first edition edition, 1979.

[54] Juan Luis Gastaldi. Why can computers understand natural language? the structuralist image of language behind word embeddings. *Philosophy & Technology*, 34(1):149–214, 2021.

[55] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[56] Edward Grefenstette and Mehrnoosh Sadrzadeh. Concrete models and empirical evaluations for the categorical compositional distributional model of meaning. *Computational Linguistics*, 41(1):71–118, 2015.

[57] Per-Erik Hagmark and Pertti Lounesto. Walsh functions, clifford algebras and cayley-dickson process. In *Clifford Algebras and Their Applications in Mathematical Physics*, pages 531–540. Springer, 1986.

[58] Sungjun Han and Sebastian Padó. Towards understanding the relationship between in-context learning and compositional generalization. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16664–16679, Torino, Italia, May 2024. ELRA and ICCL.

[59] Joe Harris. *Algebraic geometry: a first course*, volume 133. Springer Science & Business Media, 2013.

[60] Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[61] Tatsunori B. Hashimoto, David Alvarez-Melis, and Tommi S. Jaakkola. Word embeddings as metric recovery in semantic spaces. *Transactions of the Association for Computational Linguistics*, 4:273–286, 2016.

[62] Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[63] Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. Compositionality decomposed: How do neural networks generalise? *Journal of Artificial Intelligence Research*, 67:757–795, 2020.

[64] Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174, 2023.

[65] Kyoung-Rok Jang and Sung-Hyon Myaeng. Elucidating conceptual properties from word embeddings. In Jose Camacho-Collados and Mohammad Taher Pilehvar, editors, *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 91–95, Valencia, Spain, April 2017. Association for Computational Linguistics.

[66] Theo M. V. Janssen and Thomas Ede Zimmermann. Montague Semantics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2025 edition, 2025.

[67] Theo MV Janssen and Barbara H Partee. Compositionality. In *Handbook of logic and language*, pages 417–473. Elsevier, 1997.

[68] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[69] Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA, 2009.

[70] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[71] Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *ICLR*. OpenReview.net, 2020.

[72] Najoung Kim and Tal Linzen. COGS: A compositional generalization challenge based on semantic interpretation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online, November 2020. Association for Computational Linguistics.

[73] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184 vol.1, 1995.

[74] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[75] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[76] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

[77] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[78] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[79] Alessandro Lenci. Distributional models of word meaning. *Annual review of Linguistics*, 4(1):151–171, 2018.

[80] Alessandro Lenci and Magnus Sahlgren. *Distributional semantics*. Cambridge University Press, 2023.

[81] Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, 2014.

[82] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27, 2014.

[83] Tiina Lindh-Knuutila and Timo Honkela. Exploratory analysis of semantic categories: comparing data-driven and human similarity judgments. *Computational Cognitive Science*, 1:1–25, 2015.

[84] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

[85] Michel Loève. Probability theory: foundations, random sequences. *(No Title)*, 1955.

[86] Pertti Lounesto. Clifford algebras and spinors. In *Clifford algebras and their applications in mathematical physics*, pages 25–37. Springer, 2001.

[87] Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. CRC Press, 2018.

[88] Alan Macdonald. *Linear and geometric algebra*. Alan Macdonald Nottingham, 2010.

[89] Diane Maclagan and Bernd Sturmfels. *Introduction to tropical geometry*, volume 161. American Mathematical Soc., 2015.

[90] Akihiro Maeda, Takuma Torii, and Shohei Hidaka. Decomposing co-occurrence matrices into interpretable components as formal concepts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4683–4700, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[91] Akihiro Maeda, Takuma Torii, Shohei Hidaka, Naoya Inoue, and Yohei Oseki. Interpreting internal representations of transformer language models via pseudo-orthogonality of subspaces. In *The 31st Annual Meeting of the Association of Natural Language Processing*, 2025.

[92] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[93] Kate McCurdy, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao. Toward compositional behavior in neural models: A survey of current views. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9323–9339, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[94] Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 523–530, 2005.

[95] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *International Conference on Learning Representations*, 2013.

[96] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[97] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[98] Kanishka Misra and Kyle Mahowald. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on*

Empirical Methods in Natural Language Processing, pages 913–929, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[99] Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *Proceedings of ACL-08: HLT*, pages 236–244, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[100] Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429, 2010.

[101] Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648, 2007.

[102] Daichi Mochihashi. *Statistical Text Models: A Bayesian Approach to Language. (in Japanese)*. Iwanami Shoten, Publishers, 2025.

[103] Richard Montague. The proper treatment of quantification in ordinary english. In Jaakko Hintikka, Julius Moravcsik, and Patrick Suppes, editors, *Approaches to Natural Language*, pages 221–242. D. Reidel Publishing Company, Dordrecht, 1973.

[104] Sílvia Moraes and Vera Lima. Combining formal concept analysis and semantic information for building ontological structures from texts : an exploratory study. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3653–3660, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[105] Brian Murphy, Partha Talukdar, and Tom Mitchell. Learning effective and interpretable semantic models using non-negative sparse embedding. In *Proceedings of COLING 2012: Technical Papers*, pages 1933–1950, Mumbai, 2012.

[106] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023.

[107] Miyu Oba, Yohei Oseki, Akiyo Fukatsu, Akari Haga, Hiroki Ouchi, Taro Watanabe, and Saku Sugawara. Can language models induce grammatical knowledge from indirect evidence? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20591–20603, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[108] Sungjoon Park, JinYeong Bak, and Alice Oh. Rotated word vector representations and their interpretability. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 401–411, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[109] Barbara Partee et al. Compositionality. *Varieties of formal semantics*, 3:281–311, 1984.

[110] J PEARL. Graphoids: a graph-based logic for reasoning about relevance relations. *Advances in Artificial Intelligence*, pages 357–363, 1987.

[111] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[112] Steven T Piantadosi. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, 15:353–414, 2023.

[113] T.A. Plate. Holographic reduced representations. *IEEE Transactions on Neural Networks*, 6(3):623–641, 1995.

[114] Jonas Poelmans, Dmitry I. Ignatov, Sergei O. Kuznetsov, and Guido Dedene. Formal concept analysis in knowledge processing: A survey on applications. *Expert Systems with Applications*, 40(16):6538–6560, 2013.

[115] Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690. IEEE, 2011.

[116] Uta Priss. Linguistic applications of formal concept analysis. In *Formal Concept Analysis: Foundations and Applications*, page 149–160, Berlin, Heidelberg, 2005. Springer-Verlag.

[117] James Pustejovsky. *The generative lexicon*. MIT press, 1998.

[118] Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernandez Astudillo. Structural guidance for transformer language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3735–3745, Online, August 2021. Association for Computational Linguistics.

[119] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

[120] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[121] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics.

[122] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020.

[123] Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

[124] Sarah Roscoe, Minal Khatri, Adam Voshall, Surinder Batra, Sukhwinder Kaur, and Jitender Deogun. Formal concept analysis applications in bioinformatics. *ACM Comput. Surv.*, 55(8), dec 2022.

[125] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[126] Lütfi Kerem Şenel, Ihsan Utlu, Veysel Yücesoy, Aykut Koc, and Tolga Cukur. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779, 2018.

[127] Chen Shani, Jilles Vreeken, and Dafna Shahaf. Towards concept-aware large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13158–13170, Singapore, December 2023. Association for Computational Linguistics.

[128] Claude E Shannon. The redundancy of english. In *Cybernetics; Transactions of the 7th Conference, New York: Josiah Macy, Jr. Foundation*, pages 248–272, 1951.

[129] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, 2025.

[130] Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, 46(1-2):159–216, 1990.

[131] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211, 2012.

[132] Milan Studenỳ. *On Probabilistic Conditional Independence Structures*. Springer, 2005.

[133] Seth Sullivant. *Algebraic statistics*, volume 194. American Mathematical Society, 2023.

[134] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

[135] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[136] Takuma Torii, Akihiro Maeda, and Shohei Hidaka. Distributional hypothesis as isomorphism between word-word co-occurrence and analogical parallelograms. *PloS one*, 19(10):e0312151, 2024.

[137] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

[138] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[139] Stephen A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2010.

[140] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[141] Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors, *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore, December 2023. Association for Computational Linguistics.

[142] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. Survey Certification.

[143] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[144] John Wieting and Douwe Kiela. No training required: Exploring random encoders for sentence classification. In *International Conference on Learning Representations*, 2019.

[145] Rudolf Wille. Preconcept algebras and generalized double boolean algebras. In *Concept Lattices: Second International Conference on Formal Concept Analysis, ICFCA 2004, Sydney, Australia, February 23-26, 2004. Proceedings 2*, pages 1–13. Springer, 2004.

[146] Hitomi Yanaka and Koji Mineshima. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284, 2022.

[147] Ryo Yoshida, Taiga Someya, and Yohei Oseki. Tree-planted transformers: Unidirectional transformer language models with implicit syntactic supervision. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5120–5134, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Curriculum Vitae

Publications

∗: Items satisfy the publication requirements for the degree of Doctor of Philosophy.
(Ch. X): Indicates the chapter(s) in this dissertation based on the respective publication.

Peer-Reviewed Journal Articles

- ∗Akihiro Maeda, Takuma Torii, Yohei Oseki, Shohei Hidaka. *Mathematical Foundations of Compositional Language Models: Learning as an Inverse Problem in Representation Theory for Compositional Generalization*. Transactions of the Japanese Society for Artificial Intelligence, Vol.41, No.4. (In press). (in Japanese). **Lead author**. (**Ch.2**)
- Takuma Torii, Akihiro Maeda, Shohei Hidaka. *Distributional hypothesis as isomorphism between word-word co-occurrence and analogical parallelograms*. PLOS ONE, 19(10) e0312151-e0312151, Oct 21, 2024.

International Conferences (Peer-Reviewed)

- ∗Akihiro Maeda, Takuma Torii, Shohei Hidaka. *Decomposing Co-occurrence Matrices into Interpretable Components as Formal Concepts*. Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand, pp. 4683–4700, Aug 2024. **Lead author**. (**Ch.5**)
- Akihiro Maeda, Takuma Torii, Shohei Hidaka. *Interpretability of Formal Concepts in Word Co-occurrence Matrices*. 1st International Joint Conference on Conceptual Knowledge Structures (CONCEPTS 2024), p 33, Sep 2024. **Lead author**.
- Takuma Torii, Akihiro Maeda, Shohei Hidaka. *Embedding parallelepiped in co-occurrence matrix: simulation and empirical evidence*. Joint Conference on Language Evolution (JCoLE), pp. 725–732, Sep 2022.

Domestic Conferences

- Akihiro Maeda, Takuma Torii, Shohei Hidaka. *Developing Structural Learning Methods for Language Models Using Algebraic Statistical Approaches*. The JSAI Annual Conference

(JSAI2025), 2L5-GS-1-04, Jun 2025. **Lead author**. (**Ch.4**)

- Akihiro Maeda, Takuma Torii, Shohei Hidaka. *A Preliminary Study on Algebraic Statistical Approaches for the Mathematical Analysis of Language Structures*. The 31st Annual Meeting of the Association of Natural Language Processing, pp. 1675–1680, March 2025. **Lead author**. (**Ch.3**)

- Akihiro Maeda, Takuma Torii, Shohei Hidaka, Naoya Inoue, Yohei Oseki. *Interpreting Internal Representations of Transformer Language Models via Pseudo-Orthogonality of Subspaces*. The 31st Annual Meeting of the Association of Natural Language Processing, pp. 651–656, Mar 2025. **Lead author**.

- Akihiro Maeda, Takuma Torii, Shohei Hidaka, Yohei Oseki. *Extracting Features in Attention Heads of Transformer by Projection onto Subspaces*. The 30th Annual Meeting of the Association for Natural Language Processing, March, pp. 2714–2719, 2024. **Lead author**.

- Akihiro Maeda, Takuma Torii, Shohei Hidaka. *Geometric distance measurement for a parallelogram of word vectors*. The 29th Annual Meeting of the Association for Natural Language Processing, pp. 2853–2858, March 2023 **Lead author**

- Akihiro Maeda, Takuma Torii, Shohei Hidaka. *Modeling Order Effects with Context-Reflecting Word Embeddings*. Japanese Cognitive Science Society, 41st Annual Conference, pp. 325–328, Oct 2024. **Peer-reviewed, Lead author**.

- Akihiro Maeda, Takuma Torii, Shohei Hidaka *Why does the algebraic structure emerge in word distributed representations?*. Proceedings of the Annual Meeting of the Japanese Cognitive Science Society, 40th, pp. 99–102, Sep 2023. **Peer-reviewed, Lead author**.

- Akihiro Maeda, Takuma Torii, Shohei Hidaka. *Constructive approach to study the internal structure of word co-occurrence matrix*. Proceedings of the Annual Meeting of the Japanese Cognitive Science Society, 39th, pp. 302–306, 2022. **Peer-reviewed, Lead author**.

Theses / Unpublished Works

- Akihiro Maeda. *Analysis of Geometric Arrangements of Distributed Representation of Words and Internal Structure of Word Co-occurrence Matrix* [Master's Thesis, Japan Advanced Institute of Science and Technology]. Mar 2023.

Awards

- March 2025: **Outstanding Paper Award** The 31st Annual Meeting of the Association of Natural Language Processing: Akihiro Maeda, Takuma Torii, Shohei Hidaka, Naoya Inoue, Yohei Oseki. *Interpreting Internal Representations of Transformer Language Models via Pseudo-Orthogonality of Subspaces*. **Lead author**

- March 2023: **Committee Special Award** The 29th Annual Meeting of the Association for Natural Language Processing: Akihiro Maeda, Takuma Torii, Shohei Hidaka. *Geometric distance measurement for a parallelogram of word vectors*. **Lead author**