言語構造の数理分析のための代数統計的アプローチ試論

前田晃弘 [1,3]　鳥居拓馬 [2]　日髙 昇平 [1]
[1] 北陸先端科学技術大学院大学　[2] 東京電機大学　[3] 日本学術振興会特別研究員
{akihiro.maeda,shhidaka}@jaist.ac.jp　tak.torii@mail.dendai.ac.jp

概要

代数統計は代数幾何と統計の融合領域であり，構造を備えた確率モデルを高次元空間における多様体として捉える新たな視点を提供する．本研究では，言語の確率モデルを文を構成する単語の同時確率分布として定式化し，その確率ベクトルが言語の構造を反映した多様体をなすことを示す．これにより，代数幾何のツールである多項式イデアルを用いて，単語の同時確率を制約する多様体構造から言語の構造を抽出するための新たな手法を検討する．

1　はじめに

1.1　単語ベクトルと言語の数理構造

単語ベクトルの解釈可能性研究として，著者らは単語共起行列を数理的に分析してきた．[1] では，共起行列の内部に形式概念と呼ばれる数学的構造 [2] が存在し，それが解釈可能な意味的カテゴリに対応することを明らかにした．さらに共起行列の成分のうちランク 1 となる部分行列が形式概念とみなせる成分を構成していることを示した．単語ベクトルは共起行列の分解であり [3]，その数理構造を反映していると考えられる．分布仮説が単語ベクトルの理論的根拠を与えるとされるが，[1] で示された数理構造を十分説明するものではないと考えられる．

1.2　言語の生成系から確率空間への対応

一般に確率行列がランク 1 であることは二つの確率変数が独立であることと同値である．単語共起行列の部分行列がランク 1 となる現象の背景には，これらの単語群が条件付き独立を構成していることが予想される．すなわち，ある文脈を共有している単語群が，その文脈を条件として独立に生起しているのであり，そのため共起行列中でランク 1 の部分行列を構成する単語群が文脈を共有する意味的カテゴリとして観察されたと考えられる．本研究の関心は，こうした数理的・構造的対応がどのようにして現れたかを明らかにすることである．言語の生成系から単語の確率分布への対応を数理的に分析するため，構造的制約を持つ確率モデルを扱うことができる代数統計の適用を試みる．

2　代数統計的アプローチ

2.1　代数統計の背景

代数統計は代数幾何学的ツールを統計に応用したものであり，計算生物学における遺伝子系統分析のために展開された融合研究領域である [4, 5]. 代数幾何は，多項式方程式で記述される代数多様体の幾何学的構造を研究する数学であり [6, 7]，構造的な確率分布を記述する枠組みを提供する．確率モデルをパラメータ空間から確率空間への写像として記述し，この写像の像が確率空間内で代数多様体を構成することを用いて，確率モデルの推定や検定を行うことができる．こうした特性から，言語データにおける条件付き独立性や階層的構造といった複雑な制約を自然にモデル化することが期待できる．

2.2　階層的対数線形モデル

代数統計の代表的な確率モデルを説明する前に**単体的複体**を定義する．集合 V の部分集合の族 Δ が $A \in \Delta$ かつ $B \subset A$ ならば $B \in \Delta$ を満たすとき，Δ は単体的複体である．Δ の要素のうち包含を順序関係とした極大要素をファセットという．V を m 個の離散確率変数 $X_1, \ldots, X_m$ の集合として，V がなす単体的複体を Δ，ファセットの族を $\mathfrak{D}$ とする．X_i の実現値の集合を I_i とし，$X_1, \ldots, X_m$ の実現値の組を $\boldsymbol{i} = (i_1, \ldots, i_m) \in I_1 \times \cdots \times I_m$ とするとき，次の形の同時確率 $p(\boldsymbol{i})$ を**階層的対数線形モデル**と呼ぶ [5].

$$\log p(\boldsymbol{i}) = \sum_{D \in \mathfrak{D}} \mu_D(\boldsymbol{i}_D) \tag{1}$$

ただし，$\boldsymbol{i}_D = (i_k)_{k \in D} \in \prod_{k \in D} I_k$ は D が含む変数のみの実現値の組（D-周辺セル），$\mu_D : I_D \to \mathbb{R}$ は D-周

辺セルに対応するパラメータであり，マルコフ確率場のポテンシャル [8] にあたる．$\mu_D(\boldsymbol{i}_D)$ は，$|D| = 1$ の時は一変数の主効果を，$|D| \geq 2$ の時は複数変数間の交互作用を表す自由パラメータである [5].

対数線形モデルで表された同時確率関数を単項式 (monomial) の形に表す．式（1）において $\boldsymbol{i}_D$ にインデックス i を，$\boldsymbol{i}$ に j を対応させ同時確率を $p_j = p(\boldsymbol{i})$ とし，$\mu_D(\boldsymbol{i}_D) = q_j a_{ij}$ とおき，両辺の指数をとる．

$$p_j \propto \prod_{i=1}^{d} (\exp q_i)^{a_{ij}} =: \prod_{i=1}^{d} \theta_i^{a_{ij}} \tag{2}$$

$$= \theta_1^{a_{1j}} \theta_2^{a_{2j}} \cdots \theta_d^{a_{dj}} =: \boldsymbol{\theta}^{\boldsymbol{a}_j} \ (j = 1, \ldots, N) \tag{3}$$

ただし，$N = |\boldsymbol{i}|$ は実現値の組の総数，$d = \sum_{D \in \mathcal{D}} |\boldsymbol{i}_D|$ はパラメータの総数を表す．$A := (a_{ij}) \in \mathbb{N}^{d \times N}$ を配置行列 (design matrix)，式（3）で表す確率モデルをトーリックモデルと呼ぶ．

2.3 代数多様体とイデアルの導出

トーリックモデルは，配置行列 A により定まる単項写像 $\phi_A : \theta^d \to p^N, \boldsymbol{\theta} \mapsto (\boldsymbol{\theta}^{\boldsymbol{a}_1}, \boldsymbol{\theta}^{\boldsymbol{a}_1}, \ldots, \theta^{\boldsymbol{a}_N})$ である．一般に $d \ll N$ であるので，写像 ϕ_A の像は，高次元空間の中に低次元構造を与える．代数幾何的には，その閉包（Zariski 閉包と呼ばれる）$\overline{\phi_A(\mathbb{C}^d)}$ が代数多様体となる．代数多様体は，多項式方程式の集合（多項式環イデアル）の解の集合である．すなわちトーリックモデルに対応する代数多様体上のいかなる点（確率ベクトル）もイデアルに含まれる方程式を零にする（消失する）．これはイデアルが同時確率間の関係を制約していることを意味する．

マルコフ基底の基本定理 [4] を用いて，写像 ϕ_A に付随する配置行列 A からイデアルを導出することができる．手順としては，まず A の整数カーネル $z \in \ker_{\mathbb{Q}} A$ を $z = z^+ - z^-$ のように非負成分のみを持つベクトル z^+ と非正成分を正負反転したベクトル z^- に分割すると，$\boldsymbol{p}^{z^+} - \boldsymbol{p}^{z^-}$ が生成イデアル（マルコフ基底）となる．（詳細は付録 A を参照）

2.4 具体例：2×2 独立モデル

2 つの二値確率変数を $X_1, X_2 \in \{1, 2\}$ として，その同時確率が $p_{X_1 X_2} = \theta_1^{(X_1)} \theta_2^{(X_2)}$ であるとする．これはファセットを $\mathcal{D} = \{\{1\}, \{2\}\}$ とする単体的複体の階層的線形モデルであり，その配置行列は次のように表される．

$$A = \begin{array}{c} \\ \theta_1^{(1)} \\ \theta_1^{(2)} \\ \theta_2^{(1)} \\ \theta_2^{(2)} \end{array} \begin{array}{c} \begin{array}{cccc} p_{11} & p_{12} & p_{21} & p_{22} \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \end{array} \tag{4}$$

列は同時確率に，行はパラメータに対応し，行列 A はパラメータと同時確率の関係を与えている．A の整数カーネルに含まれる $z = [1, -1, -1, 1]^T$ を，$z^+ = [1, 0, 0, 1]^T$ と $z^- = [0, 1, 1, 0]^T$ に分解し，確率ベクトル $\boldsymbol{p} = [p_{11}, p_{12}, p_{21}, p_{22}]$ にマルコフ基底の基本定理を適用すると，次式の生成イデアルをえる．

$$\boldsymbol{p}^{z^+} - \boldsymbol{p}^{z^-} = p_{11} p_{22} - p_{12} p_{21} = 0 \tag{5}$$

2 変数が独立の場合，同時確率 p_{ij} は式（5）を満たす．同時確率を表す 2×2 の行列がランク 1 であり，その行列式 $\begin{vmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{vmatrix}$ は零である．図 1 の編みかけ

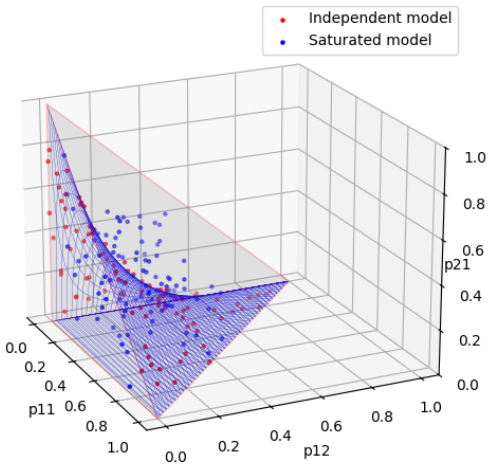

図 1 3 次元シンプレックス上の代数多様体

部分は，イデアル（式 5）の代数多様体を示す．4 次元の確率ベクトルを 3 次元シンプレックス上に示している．併せて，独立確率モデルからの生成データ（赤い点）と，制約のない四変数の多項分布からの生成データ（青い点）それぞれ 100 回試行したものを示す．前者は代数多様体上に位置している．

3 代数統計的言語モデリング

3.1 文の生起確率

単語共起分布は単語対が同一の文（文脈）で生起することに由来し，文の統語的・意味的な制約を反

映している．この対応を明示的に分析するため，文に着目しその生起確率をモデル化する．具体的には，文構造を構成する品詞（part-of-speech）を確率変数と見做し，文の生起確率を多変数の離散同時確率により定式化する．

トイモデルとして，3 つの確率変数 S,V,O を持つ同時確率分布を考える．簡単化のために二値確率変数として各々の実現値を $\{s_1,s_2\},\{v_1,v_2\},\{o_1,o_2\}$ とする．(s_1,v_1,o_1) は SVO の形をもつ文を表し，その同時確率 $Pr(s_1,v_1,o_1)=p_{111}$ は文の生起確率を表す．トイモデルには 8 文のみが含まれ，その確率分布は 3 階テンソル $(p_{ijk})_{i,j,k=1,2}$ により表現される．

3.2 完全独立モデル

S,V,O のいずれの 2 変数間でも独立性が成り立つ場合を考える[1)]．ファセットを $\mathfrak{D}=\{\{S\},\{V\},\{O\}\}$ とする階層的対数線形モデルであり，文の生起確率は $p_{ijk}=\theta_S^{(i)}\theta_V^{(j)}\theta_O^{(k)}$ $(i,j,k=1,2)$ で与えられる[2)]．トーリックモデルのイデアルを導出すると（付録 A），$p_{111}p_{122}-p_{112}p_{121}=0$ の形をした 2 次の小行列式 (2-minor と呼ぶ）12 本の式 (8)–(19) が生成イデアルとなる．これは，8 文に対応する同時確率 $p_{111},p_{112},\ldots,p_{222}$ の間には 12 本の方程式の全てを零とするような関係があることを意味する．また，確率ベクトルの総和が 1 である．

$$\sum_{i,j,k=1,2} p_{ijk}=1 \tag{6}$$

これらの方程式は，8 次元空間 (7 次元シンプレックス）上の代数多様体に対応しており，このモデルから生成される確率ベクトルは常にその代数多様体上に位置する．

3.3 条件付き独立モデル

確率変数 S を条件として V,O が独立となる条件付き確率モデルを示す[3)]．階層的対数線形モデルのファセットは $\mathfrak{D}=\{\{S,V\},\{S,O\}\}$ となる．このモデルから生成される同時確率ベクトルは，イデアルとして導出される二つの式 $p_{111}p_{122}-p_{112}p_{121}=0$ と $p_{211}p_{222}-p_{212}p_{221}=0$ と確率総和 1 の制約式（6) を満たす．これは 3 階テンソル (p_{ijk}) を S 軸に直交するようにスライスして得られる 2 つの 2×2 行列の行列式が零であることを意味する．

なお，条件付き独立モデルより弱い独立性を持つモデルである Context-specific independence も階層的対数線形モデルによりモデル化できる [5]．例えば，$S=s_1$ の時のみ V,O が独立となるモデルでは，イデアルは $p_{111}p_{122}-p_{112}p_{121}=0$ のみとなる．

3.4 モデル間の階層構造

階層的対数線形モデルは付随する単体的複体のファセットによって特定される．3 変数の場合ファセットの組み合わせは 9 通りである[4)]ので，9 個の確率モデルが考えられる．それぞれの確率モデルから生成された同時確率ベクトル（データ）は，高次元の確率空間中の多様体上に位置する．

9 個のモデルのうち，完全独立モデル（確率変数 S,V,O を X_1,X_2,X_3 と呼びかえてファセットを添字で表すと $\mathfrak{D}=\{\{1\},\{2\},\{3\}\}$）が最も制約の強い独立性を備えているのに対して，最も弱い独立性を持つのが式（6）のみを制約として持つ確率モデルであり飽和モデルと呼ばれる．そのファセットは $\mathfrak{D}=\{\{123\}\}$ であり，3 つの変数全てが交絡するモデルである．飽和モデルのクラスは 8 次元空間上にある 7 次元単体を被覆する．

これらの 9 個のモデルは，イデアルの包含関係を通じて階層構造にある．図 2 は，9 つのモデルに

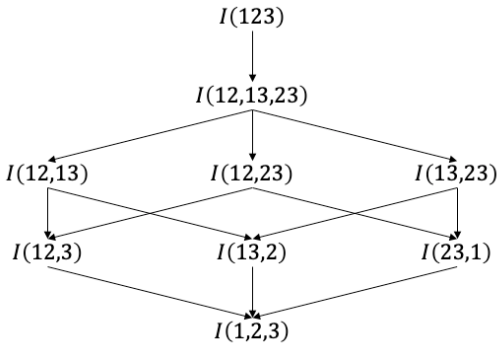

図 2 モデル選択のためのイデアルの階層（束）構造

対応するイデアルの包含を順序関係（図中の矢印の方向に「含まれる」関係）とする束（lattice）を示す．最上位にある $I(123)$ が最も構造のない飽和モデル（満たすべき式は (6) のみ）であり，最下位にある $I(1,2,3)$ が最も構造の多い（満たすべき方程式が 13 個の）完全独立モデルに対応する．12 個ある

1) 例えば，$S=\{man,woman\},V=\{likes,hates\},O=\{fish,pork\}$ として，性別間で嗜好の偏りがない場合，各変数が独立となることは自然であると考えられる．独立を構成する単語群は Syntactic equivalence を成している．

2) $\theta_S^{(i)},\theta_V^{(j)},\theta_O^{(k)}(i,j,k=1,2)$ は 6 個のパラメータ

3) $S=\{Andy,Bob\},V=\{likes,hates\},O=\{fish,pork\}$ として，主体により食の嗜好が異なるケースに当たる．

4) 確率変数の添字のみで表現すれば $\{123\},\{12,13,23\},\{12,13\},\{12,23\},\{13,23\},\{12,3\},\{13,2\},\{23,1\},\{1,2,3\}$

2-minor の全てが完全独立モデルのイデアルに含まれるのに対し，飽和モデルはそのいずれも含まれず（イデアルは空集合），束上その二つの間にある確率モデルはその部分集合のみをイデアルとする[5)]．

4　イデアルを用いた構造学習

4.1　モデル選択のための数値実験

構造が異なる確率モデルの間にイデアルの階層構造があることを用いてモデル選択を行い，データから未知のモデルの構造を学習する手法を検討する．

考え方 イデアルの候補となる方程式（2-minor）をあらかじめ列挙しておき，未知のモデルから生成されたデータがいずれの 2-minor を零とするかを調べ，その充足状況から対応する確率モデルを選択する．トイモデル（3 節）の場合，3 階テンソル中に現れる 12 個の 2-minor（式 8–19）を特徴量として，異なるモデルを識別できることを数値実験により示す．

実験手順 9 個のファセットに対応した確率モデルから 1 つを未知モデルとして選択して，ファセットに割り当てるパラメータを $[0,1]$ の一様分布により抽出して同時確率（確率テンソル）を計算したものをデータとする．1000 回施行して 2-minor の値を算出しその分布を得る．

実験結果 図 3 はファセットを $\{\{1,2\},\{1,3\}\}$ とする条件付き独立モデルを例として，2-minor の値の分布を示す．二式（8,13）において値が零に一致しており，分散は零である．その他の確率モデルに対応する 2-minor の分布図を付録 B に示す．

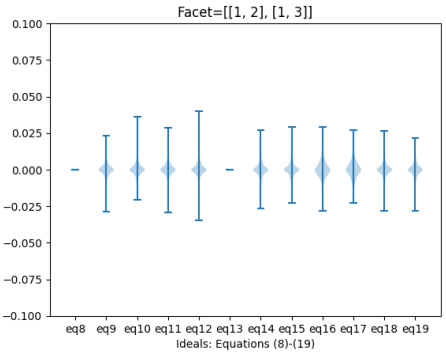

図 3　2-minor の分布：条件付き独立モデル

図 4 は，9 つのモデル（y 軸）がいずれの方程式（x 軸）を零とするか（消失イデアルか）をヒートマップで表している（閾値 $\epsilon = 10^{-10}$ として，白いセルが絶対値 ϵ 以下）．図より明らかなように，未知の確率モデルから生成されたデータが与えられた時，消失イデアルを特徴量としたモデル選択が可能と考えられる．データが与えられた時，それぞれのモデルの最大尤度を計算をすれば良いが，図 2 に示したようにモデル間の包含関係を踏まえ，赤池情報量基準 (AIC) またはベイズ情報量基準 (BIC) を用いることが考えられ，今後の研究課題である．

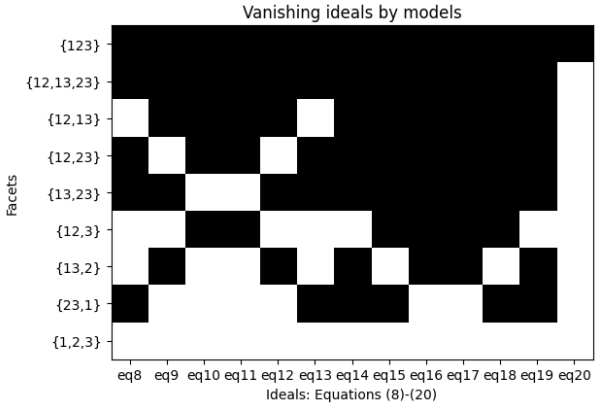

図 4　各モデルと消失イデアル（ヒートマップ）

4.2　課題: 計算量と周辺化

2-minor を用いた構造学習には，変数と状態数が増加すると組合せ爆発により特徴量とする方程式の数が階乗の速度で大きくなるという計算上の課題がある．そのため，確率テンソルの 2-minor を直接計算する代わりに，周辺化した確率行列の 2-minor を調べる方法が考えられる．ある条件のもとでは周辺化によって特定の変数間の条件付き独立の構造が保存される．また，イデアルの束に沿って局所的に探索する方法も考えられる．

単語共起行列は文脈における単語の同時確率を共起ペアのもとで周辺化した確率を表す．共起行列中にランク 1 の構造が現れる事実は，テンソル中のセルの構造が保存されて周辺化された行列で表面化したものと考えることができる．共起行列に対するこうした構造的探索の開発は今後の課題である．

5　まとめ

本研究では，言語を離散変数の同時確率として階層的対数線形モデルにより定式化し，イデアルを用いる代数統計的手法の適用を検討した．特に，条件付き独立モデルの構造学習を可能とする枠組みを示し，数値実験によりその実行可能性を示した．今後の課題として，計算量を考慮したアルゴリズムの開発と統計的モデル選択手法の導入が挙げられる．

5)　$I(12,13,23)$ は 4 次式（20）をイデアルとする．

謝辞

本研究は科研費基盤研究 B(一般) JP23H0369，JST さきがけ JPMJPR20C9，JST CREST JPMJCR23P4，JSPS KAKENHI 24KJ1202 の助成を受けて行われた．

参考文献

[1] Akihiro Maeda, Takuma Torii, and Shohei Hidaka. Decomposing co-occurrence matrices into interpretable components as formal concepts. In **Findings of the Association for Computational Linguistics: ACL 2024**, 2024.

[2] Bernhard Ganter and Rudolf Wille. **Formal Concept Analysis: Mathematical Foundations**. Springer, 1999.

[3] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In **Advances in Neural Information Processing Systems**, 2014.

[4] Mathias Drton, Bernd Sturmfels, and Seth Sullivant. **Lectures on algebraic statistics**, Vol. 39. Springer Science & Business Media, 2008.

[5] 青木敏, 竹村彰通, 原尚幸. 代数的統計モデル. 共立出版, 2019.

[6] David Cox, John Little, Donal O'shea, and Moss Sweedler. **Ideals, varieties, and algorithms**, Vol. 3. Springer, 1997.

[7] 川又雄二郎. 射影空間の幾何学. 朝倉書店, 2001.

[8] D. Koller and N. Friedman. **Probabilistic Graphical Models: Principles and Techniques**. MIT Press, 2009.

A　マルコフ基底の基本定理によるイデアル導出

第 3.2 節で取り上げた 3 変数の完全独立モデルを例に，イデアルの導出手続きを詳説する．完全独立モデルはファセットの族 $\mathcal{D} = \{\{S\}, \{V\}, \{O\}\}$ を持つ対数線形モデルで表され，それぞれ 2 状態を持つので，$\theta_S^{(i)}, \theta_V^{(j)}, \theta_O^{(k)} (i, j, k = 1, 2)$ の 6 個のパラメータを持つ．同時確率 $p_{ijk} = \theta_S^{(i)}\theta_V^{(j)}\theta_O^{(k)}$ は次の配置行列 A に対応する．$A \in \mathbb{N}^{d\times N}$ で $d = 6, N = 2^3 = 8$.

$$A = \begin{array}{c} \\ \theta_S^{(1)} \\ \theta_S^{(2)} \\ \theta_V^{(1)} \\ \theta_V^{(2)} \\ \theta_O^{(1)} \\ \theta_O^{(2)} \end{array} \begin{array}{c} \begin{array}{cccccccc} p_{111} & p_{112} & p_{121} & p_{122} & p_{211} & p_{212} & p_{221} & p_{222} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix} \end{array} \tag{7}$$

行列 A の整数カーネルは $\ker_{\mathbb{Z}} A = \{z \in \mathbb{R}^8 | Az = \mathbf{0}\}$ と定義され，例えば，$z = [1, -1, -1, 1, 0, 0, 0, 0]^T \in \ker_{\mathbb{Z}} A$ である．この時，正負の成分を持つ二つのベクトルに分解して $z = z^- - z^-$，$z^+ = [1, 0, 0, 1, 0, 0, 0, 0]^T, z^- = [0, 1, 1, 0, 0, 0, 0, 0]^T$ とするとイデアルは次のように得られる．基本定理は，この手順により得られる式の集合が生成イデアルであることを保証する．

$$\boldsymbol{p}^{z^+} - \boldsymbol{p}^{z^-} = p_{111}^1 p_{112}^0 p_{121}^0 p_{122}^1 p_{211}^0 p_{212}^0 p_{221}^0 p_{222}^0 - p_{111}^0 p_{112}^1 p_{121}^1 p_{122}^0 p_{211}^0 p_{212}^0 p_{221}^0 p_{222}^0 = p_{111}p_{122} - p_{112}p_{121} \tag{8}$$

同様に $\ker_{\mathbb{Z}} A$ に含まれるその他のベクトルから下記のイデアルを求める．6 つの式（8–13）は，確率テンソルを $2\times2\times2$ の立方体と見た時の 6 面に現れる 2×2 の行列式に対応し，式（14–19）は 4 つの対角線の組合わせ $({}_4C_2)$ に対応する．式（20）はファセット $\{\{1,2\},\{1,3\},\{2,3\}\}$ に対応する 4 次式である．

$$p_{111}p_{212} - p_{112}p_{211} \tag{9}$$
$$p_{111}p_{221} - p_{121}p_{211} \tag{10}$$
$$p_{112}p_{222} - p_{122}p_{212} \tag{11}$$
$$p_{121}p_{222} - p_{122}p_{221} \tag{12}$$
$$p_{211}p_{222} - p_{212}p_{221} \tag{13}$$
$$p_{111}p_{222} - p_{112}p_{221} \tag{14}$$
$$p_{111}p_{222} - p_{121}p_{212} \tag{15}$$
$$p_{111}p_{222} - p_{211}p_{122} \tag{16}$$
$$p_{112}p_{221} - p_{121}p_{212} \tag{17}$$
$$p_{112}p_{221} - p_{211}p_{122} \tag{18}$$
$$p_{121}p_{212} - p_{211}p_{122} \tag{19}$$
$$p_{111}p_{122}p_{212}p_{221} - p_{112}p_{121}p_{211}p_{222} \tag{20}$$

B　確率モデル間の 2-minor の分布比較

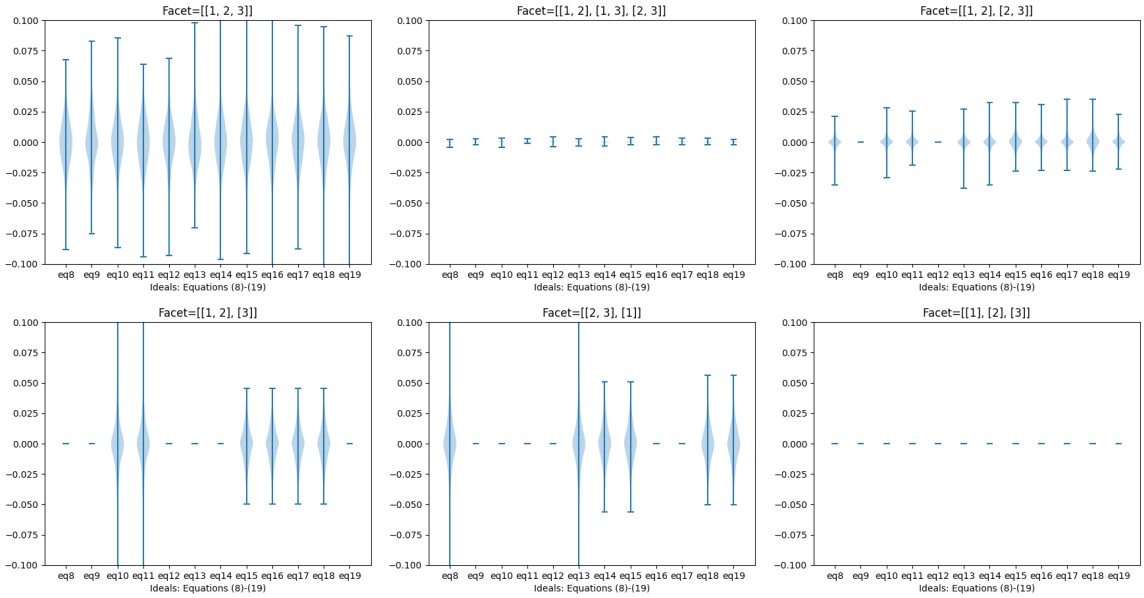

図 5　2-minor の分布：9 つのモデルのうち 6 モデルを示す．残り 3 つは上記いずれかと同型