BINARY EXPANSION GROUP INTERSECTION NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional independence is fundamental to modern statistics and machine learning, yet beyond special parametric families it rarely admits an exact covariance-based characterization. We introduce the binary expansion group intersection network (BEGIN), a distribution-free graphical framework for representing conditional independence in multivariate binary data and bit-encoded multinomial variables. We show that for arbitrary binary random vectors, conditional independence is equivalent to a zero cross block of a generalized Schur complement. Its entries are probability-weighted sums of conditional covariances, with signs determined by conditioning interactions. The representation encompasses Ising models, higher-order binary models, and categorical graphical models, including distributions with structural zeros. We characterize the rank of the interaction covariance and establish Markov properties and conditions for concatenating BEGIN graphs. We test conditional independence using the maximum absolute value of the estimated entries in the BEGIN cross block, and we show that this test attains the optimal sample order under sparse interaction alternatives. Simulations show the power gain under sparsity and the computational efficiency of the interaction scan. An analysis of categorical Mushroom records shows how the maximizing binary interactions explain detected conditional associations.

1 INTRODUCTION

Conditional independence is central to statistical reasoning. It supports the interpretation of multivariate associations, graphical modeling, and many procedures in causal inference, variable selection, and structure learning. In classical settings, it is often studied through parametric models whose covariance structures have direct probabilistic interpretations. Modern data, however, are often heterogeneous, high dimensional, or only weakly modeled, making exact parametric assumptions hard to justify.

This tension is especially acute in distribution-free inference. Conditional independence is a natural way to formalize "no direct association after adjustment," yet fully distribution-free conditional-independence testing is fundamentally impossible without additional structure; see, for example, Shah & Peters (2020). The challenge is therefore to identify structure that is mathematically exact while remaining as assumption-lean as possible; see, for example, Christgau et al. (2023).

We approach this problem through binary expansion statistics (Zhang, 2019; Zhang et al., 2021; Brown et al., 2025). Two binary variables are independent if and only if they are uncorrelated, and every function of finitely many bits is a linear combination of their interactions. This finite structure permits an exact covariance characterization. Multinomial variables can also be encoded by bits. In this case, bit combinations that represent no category have probability zero. These *structural zeros* make the full interaction covariance matrix singular. BEGIN allows these support restrictions.

BEGIN represents conditioning information by the intersection of two binary interaction groups. Its generalized Schur complement records probability-weighted conditional covariances through an explicit Hadamard transform. Section 3 establishes the characterization, its rank and support properties, and the conditions for combining local structures into global Markov models. Nested binary expansions extend the representation to general variables: under the continuity conditions of Theorem 3.7, conditional independence is equivalent to convergence of the BEGIN cross-block maximum norm to zero as the resolution increases.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The same entries connect testing, computation, and interpretation. Max BEGIN attains the optimal sample order for sparse interactions under a specified design. Binary products permit an efficient scan and identify which category groups contribute to a rejection. Simulations compare power and computation; a Mushroom example illustrates the interpretation.

2 Related work

Graphical models describe conditional independence through separation and factorization (Lauritzen, 1996; Wainwright & Jordan, 2008). Ising neighborhood selection uses logistic regression under pairwise interactions (Ravikumar et al., 2010). The generalized covariance approach of Loh & Wainwright (2012) relates graphs to inverse covariances of indicator statistics augmented over cliques. BEGIN instead uses complete binary interactions and a generalized Schur complement that allows singular encodings.

Residual covariance criteria characterize conditional independence over complete function classes (Daudin, 1980). Theorem 3(ii) of Fukumizu et al. (2008) augments one tested variable with the conditioning variable. Kernel conditional independence testing (KCI) aggregates kernel residual features (Zhang et al., 2011). The generalized covariance measure (GCM) and its weighted version (WGCM) use residual products and weights, including a maximum over fixed weights (Shah & Peters, 2020; Scheidegger et al., 2022). BEGIN specifies binary product weights, sparse alternatives, and bitwise computation within its graphical representation. Its power theory builds on sparse interaction testing (Zhang, 2019; Arias-Castro et al., 2011). Conditional randomization gives exact calibration when one conditional law is known (Candès et al., 2018).

3 Conditional independence through binary interactions

Notation. For $\boldsymbol{X} = (X_1, \dots, X_p) \in \{\pm 1\}^p$, we call a product $X_\Lambda = \prod_j X_j^{\Lambda_j}$ a *binary interaction*, where $\Lambda \in \{0,1\}^p$ records which bits appear. We call a collection a *group* under multiplication when products stay in the collection, multiplication is associative, an identity is present, and every element has an inverse. The 2^p formal interactions form the group $\langle \boldsymbol{X} \rangle$: its identity is 1, and each interaction is its own inverse because $X_j^2 = 1$.

We refer to the rule $0 \oplus 0 = 1 \oplus 1 = 0$ and $0 \oplus 1 = 1 \oplus 0 = 1$ as *addition modulo two*. We write $\mathbb{Z}_2 = \{0,1\}$ for these symbols with addition modulo two and ordinary multiplication. Applied coordinatewise to exponent vectors, it gives $X_\Lambda X_\Gamma = X_{\Lambda \oplus \Gamma}$. Distinct labels remain distinct indices even if their evaluations agree almost surely, so the covariance retains deterministic relations. We call a subset a *subgroup* when it is itself a group under the same multiplication. Group intersections use the same coordinate labels. Appendix K illustrates these definitions and deterministic relations.

Write $\boldsymbol{X}_M$ for the subvector indexed by M. The *power vector* collects all interactions:

$$\boldsymbol{X}^{\otimes} = \begin{pmatrix} 1 \\ X_1 \end{pmatrix} \otimes \cdots \otimes \begin{pmatrix} 1 \\ X_p \end{pmatrix}. \tag{3.1}$$

Here $\otimes$ denotes the Kronecker product. Order the interaction labels $\Lambda \in \{0,1\}^p$ lexicographically, with the last coordinate varying fastest, and order the states $\boldsymbol{x}$ by their codewords $\mathbf{s}$, where $s_j = (1 - x_j)/2$, in the same order. The possible power vectors are the columns of the symmetric Hadamard matrix $\mathbf{H}_p$, with $(\mathbf{H}_p)_{\Lambda, \boldsymbol{x}} = (-1)^{\Lambda^\top \mathbf{s}} = x_\Lambda$ and $\mathbf{H}_p^2 = 2^p \mathbf{I}$. For a matrix $\mathbf{M}$, $\mathbf{M}^+$ is its Moore–Penrose inverse, $\mathbf{M}[\mathcal{R}, \mathcal{C}]$ is the indicated submatrix, and $\|\mathbf{M}\|_{\max} = \max_{i,j} |M_{ij}|$. The *support* is $\mathrm{Supp}(\boldsymbol{X}) = \{\boldsymbol{x} : \mathbf{P}(\boldsymbol{X} = \boldsymbol{x}) > 0\}$.

3.1 Rank and support

Binary encodings of categorical variables can contain unused patterns. The following result shows exactly how these support restrictions affect the interaction covariance.

Theorem 3.1 (Rank and support). *Let $\boldsymbol{\Sigma}_{\langle \boldsymbol{X} \rangle \setminus \{1\}}$ be the covariance of the $2^p - 1$ nonidentity interactions of $\boldsymbol{X} \in \{\pm 1\}^p$. Then*

$$\mathrm{rank}\big(\boldsymbol{\Sigma}_{\langle \boldsymbol{X} \rangle \setminus \{1\}}\big) = |\,\mathrm{Supp}(\boldsymbol{X})| - 1.$$

Its null space consists of coefficients whose interaction combinations are constant on the support.

For example, six categories encoded by three bits give seven nonidentity interactions with covariance rank five. The singularity records the support restrictions and does not remove information from the full interaction representation. This motivates the generalized inverse in BEGIN.

3.2 THE BEGIN CHARACTERIZATION

Let $\boldsymbol{A} \in \{\pm 1\}^{r_A}$, $\boldsymbol{B} \in \{\pm 1\}^{r_B}$, and $\boldsymbol{C} \in \{\pm 1\}^{r_C}$, with distinct coordinate labels, $r_A, r_C \geq 1$, and $r_B \geq 0$. We call $\boldsymbol{A}$ and $\boldsymbol{C}$ the *endpoints*, the two vectors whose conditional independence is considered. They are *conditionally independent given* $\boldsymbol{B}$, written $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B}$, when their conditional joint pmf factors at every supported value of $\boldsymbol{B}$. Define

$$\mathcal{B} = \langle \boldsymbol{B} \rangle \setminus \{1\}, \qquad \mathcal{L} = \langle \boldsymbol{A}, \boldsymbol{B} \rangle \setminus \langle \boldsymbol{B} \rangle, \qquad \mathcal{R} = \langle \boldsymbol{B}, \boldsymbol{C} \rangle \setminus \langle \boldsymbol{B} \rangle.$$

Let $\boldsymbol{\Sigma}$ be the covariance of the interactions ordered as $(\mathcal{B}, \mathcal{L}, \mathcal{R})$, and use subscripts for principal blocks. We refer to the following residual covariance matrix as its *generalized Schur complement*:

$$\mathbf{S} = \boldsymbol{\Sigma}[\mathcal{L} \cup \mathcal{R}, \mathcal{L} \cup \mathcal{R}] - \boldsymbol{\Sigma}[\mathcal{L} \cup \mathcal{R}, \mathcal{B}] \boldsymbol{\Sigma}_{\mathcal{B}}^{+} \boldsymbol{\Sigma}[\mathcal{B}, \mathcal{L} \cup \mathcal{R}]. \tag{3.2}$$

For empty $\boldsymbol{B}$, the correction is zero. The *cross block* $\mathbf{S}[\mathcal{L}, \mathcal{R}]$ contains the residual covariances between the left and right interaction blocks.

Theorem 3.2 (BEGIN)**.** *For any joint pmf, the following statements are equivalent.*

(a) $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B}$.

(b) For every $\Lambda_1 \in \{0,1\}^{r_A}$ *and* $\Lambda_2 \in \{0,1\}^{r_C}$*, almost surely,*

$$\begin{aligned} \mathbf{E}[A_{\Lambda_1} \mid \boldsymbol{B}, \boldsymbol{C}] &= \mathbf{E}[A_{\Lambda_1} \mid \boldsymbol{B}], \\ \mathbf{E}[C_{\Lambda_2} \mid \boldsymbol{A}, \boldsymbol{B}] &= \mathbf{E}[C_{\Lambda_2} \mid \boldsymbol{B}]. \end{aligned} \tag{3.3}$$

(c) $\mathbf{S}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$.

Either family of conditional-expectation identities in (b) alone is sufficient.

Each conditional mean in (b) is a linear combination of interactions of $\boldsymbol{B}$, giving the BELIEF representation (Brown et al., 2025). The intersection $\langle \boldsymbol{A}, \boldsymbol{B} \rangle \cap \langle \boldsymbol{B}, \boldsymbol{C} \rangle = \langle \boldsymbol{B} \rangle$ identifies the conditioning group. Its nonidentity block may be reduced if the retained interactions and the constant still span all functions on the support; arbitrary deletion can invalidate either implication (Appendix J).

For nonzero endpoint labels Λ_1, Λ_2 and supported conditioning states $\boldsymbol{b}$, define

$$\mathfrak{c}_{\Lambda_1\Lambda_2}(\boldsymbol{b}) = \mathrm{Cov}(A_{\Lambda_1}, C_{\Lambda_2} \mid \boldsymbol{B} = \boldsymbol{b}), \qquad w_{\Lambda_1\Lambda_2}(\boldsymbol{b}) = \mathbf{P}(\boldsymbol{B} = \boldsymbol{b}) \mathfrak{c}_{\Lambda_1\Lambda_2}(\boldsymbol{b}),$$

where $w_{\Lambda_1\Lambda_2}(\boldsymbol{b}) = 0$ off the support. Let $\boldsymbol{w}_{\Lambda_1\Lambda_2}$ collect these values over the binary cube.

Proposition 3.3 (Explicit cross block)**.** *For nonzero* Λ_1, Λ_2 *and* $\Gamma, \Gamma' \in \{0,1\}^{r_B}$*,*

$$\mathbf{S}[A_{\Lambda_1} B_{\Gamma}, B_{\Gamma'} C_{\Lambda_2}] = \mathbf{E}[B_{\Gamma \oplus \Gamma'} \mathfrak{c}_{\Lambda_1\Lambda_2}(\boldsymbol{B})] = (\mathbf{H}_{r_B} \boldsymbol{w}_{\Lambda_1\Lambda_2})_{\Gamma \oplus \Gamma'}. \tag{3.4}$$

Consequently, with maxima over nonzero endpoint labels,

$$\max_{\Lambda_1, \Lambda_2} \|\boldsymbol{w}_{\Lambda_1\Lambda_2}\|_2 \leq \|\mathbf{S}[\mathcal{L}, \mathcal{R}]\|_{\max} \leq \max_{\Lambda_1, \Lambda_2} \mathbf{E}|\mathfrak{c}_{\Lambda_1\Lambda_2}(\boldsymbol{B})|. \tag{3.5}$$

In (3.4), each conditional covariance has weight $\mathbf{P}(\boldsymbol{B} = \boldsymbol{b}) B_{\Gamma \oplus \Gamma'}(\boldsymbol{b})$. The conditioning product supplies its sign; the constant gives the average conditional covariance.

Proposition 3.3 gives a short proof of Theorem 3.2. Since $\mathbf{H}_{r_B}$ is invertible, the cross block is zero if and only if every $\boldsymbol{w}_{\Lambda_1\Lambda_2}$ is zero, or equivalently, all conditional covariances between endpoint interactions vanish on $\mathrm{Supp}(\boldsymbol{B})$. Binary interactions and the constant span the endpoint state indicators, so this is equivalent to factorization of the conditional joint pmf. The same indicator expansion proves that either family in (b) is equivalent to (a). Appendices A and B give complete proofs. Appendix I gives finite-sample bounds and residual scores for searches over a prespecified finite family of conditioning subgroups.

3.3 EXAMPLES

The following examples represent conditional independence by group intersections.

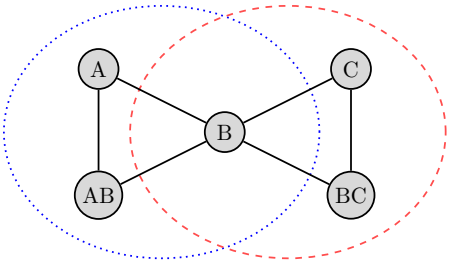
1. *Three binary variables.* In the simplest case $r_A = r_B = r_C = 1$, Figure 1(a) represents $A \perp\!\!\!\perp C \mid B$ through
$$\langle A, B\rangle \cap \langle B, C\rangle = \langle B\rangle.$$
Removing the common group gives $\mathcal{L} = \{A, AB\}$ and $\mathcal{R} = \{C, BC\}$, and Proposition 3.3 yields
$$\mathbf{S}[\mathcal{L}, \mathcal{R}] = \sum_{b\in\{\pm 1\}} \mathbf{P}(B = b)\,\mathrm{Cov}(A, C \mid B = b) \begin{pmatrix} 1 & b \\ b & 1 \end{pmatrix}.$$
The diagonal entries average the conditional covariances, while the off-diagonal entries weight them by B. Together they determine whether the conditional covariances vanish on the support of B, even when their probability-weighted average is zero.

2. *A binary first-order Markov chain.* Let (A_1, A_2, A_3, A_4) be a first-order Markov chain, not necessarily stationary. The adjacent pair groups satisfy
$$\langle A_1, A_2\rangle \cap \langle A_2, A_3\rangle = \langle A_2\rangle, \qquad \langle A_2, A_3\rangle \cap \langle A_3, A_4\rangle = \langle A_3\rangle.$$
Figure 1(b) joins these groups through their intersections. The Markov property also gives $A_4 \perp\!\!\!\perp (A_1, A_2) \mid A_3$, whose conditioning group is $\langle A_1, A_2, A_3\rangle \cap \langle A_3, A_4\rangle = \langle A_3\rangle$. Section 3.4 states the conditions under which such local structures concatenate into a global Markov graph.

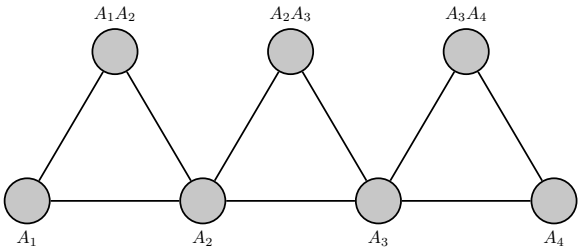
3. *Conditioning on interactions.* The example of Brown et al. (2025) in Figure 1(c) represents
$$B \perp\!\!\!\perp (A_1, A_2, A_3) \mid (A_1A_2, A_2A_3, A_1A_3)$$
through the intersection
$$\langle B, A_1A_2, A_1A_3\rangle \cap \langle A_1, A_2, A_3\rangle = \{1, A_1A_2, A_1A_3, A_2A_3\}.$$
Only two conditioning products are needed, since $A_2A_3 = (A_1A_2)(A_1A_3)$. Together with $A_1A_2A_3$, they determine all three original signs. Thus Theorem 3.2 applies with endpoints $B, A_1A_2A_3$ and conditioning variables A_1A_2, A_1A_3. The group intersection represents conditioning information that is not a subset of the original variables. Appendix K gives the recovery identities.

Remark 3.4 (Existing models as special cases)**.** *Theorem 3.2 gives a common covariance representation for the Ising model, which retains single-variable and pairwise terms in the log pmf, and general binary log-linear models with higher-order interactions (Bishop et al., 1975; Ravikumar et al., 2010; Wainwright & Jordan, 2008). It also covers bit-encoded categorical Markov random fields and pmfs with structural zeros. Appendix K gives the expansion and a higher-order example; Appendix E illustrates an Ising cycle.*

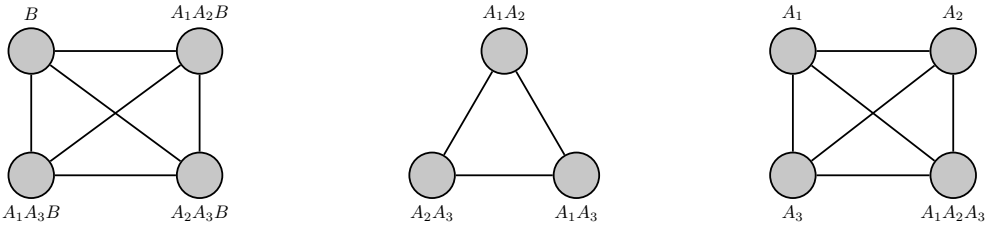
Remark 3.5 (Reduction of Dimension under a Markov condition)**.** *For each j, fix $M_j \subseteq \{1, \dots, j-1\}$ with $|M_j| \le r_{\max}$. Suppose each bit X_j is conditionally independent of the earlier bits outside M_j given $\boldsymbol{X}_{M_j}$. By the standard Bayesian-network factorization (Koller & Friedman, 2009), the positive model has*
$$D = \sum_{j=1}^{p} 2^{|M_j|} \le \sum_{j=1}^{p} 2^{\min(r_{\max}, j-1)} \tag{3.6}$$
free conditional probabilities, or the same number of conditional-mean interaction coefficients. The positive, possibly nonstationary order-three Markov model has $D = 8p - 17$ for $p \ge 3$: twenty bits require 143 parameters instead of $2^{20} - 1 = 1{,}048{,}575$. This reduces the model dimension; under positivity, the full interaction covariance still has rank $2^p - 1$. Separately, a BEGIN comparison of two scalar bits given r_B bits uses $3 \cdot 2^{r_B} - 1$ features, at most 23 for $r_B \le 3$. Appendix D gives the short chain-rule argument, including structural zeros.



(a) BEGIN for $A \perp\!\!\!\perp C \mid B$ as $\langle A, B\rangle \cap \langle B, C\rangle$.



(b) BEGIN for a Markov chain, joining $\langle A_1, A_2\rangle \cap \langle A_2, A_3\rangle$ and $\langle A_2, A_3\rangle \cap \langle A_3, A_4\rangle$.



(c) Left wing, center, and right wing of BEGIN for $B \perp\!\!\!\perp (A_1, A_2, A_3) \mid (A_1A_2, A_2A_3, A_1A_3)$, corresponding to $\langle A_1A_2, A_1A_3, B\rangle \cap \langle A_1, A_2, A_3\rangle$.

Figure 1: Examples of BEGIN, where conditional independence is represented through intersections of multiplicative groups of binary interactions. The constant interaction is omitted from the drawings; each panel represents conditional independence when its BEGIN cross block is zero.

3.4 Graphical interpretation and concatenation

The zero cross block has a graphical interpretation: in the graph of the Schur–Banachiewicz generalized inverse, the conditioning block separates the endpoint blocks. This inverse preserves the separation for singular covariances; the ordinary Moore–Penrose inverse of the full covariance need not do so. Appendix C gives the precise construction and a counterexample.

Let $G = (V, E)$ be a finite undirected graph with a binary vector $\boldsymbol{X}_v$ at each vertex, and let $\boldsymbol{X}_U$ collect the variables in $U \subseteq V$. Write $N(v)$ for the neighbors of v. We refer to the requirement

$$\boldsymbol{X}_v \perp\!\!\!\perp \boldsymbol{X}_{V\setminus(\{v\}\cup N(v))} \mid \boldsymbol{X}_{N(v)}, \qquad v \in V,$$

as the *local Markov property*. It states that the neighbors of a vertex contain all information needed to separate its variable from the nonneighbors. We call a set S a *separator* of I and J if every path from I to J meets S, where I, J, S are pairwise disjoint subsets of V. We refer to the requirement

$$\boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_S \quad \text{for every such separation}$$

as the *global Markov property*. We refer to conditional independence of every nonadjacent pair given all remaining variables as the *pairwise Markov property*.

Global implies local, and local implies pairwise. The three properties are equivalent when the pmf is positive on $\prod_{v\in V} \mathrm{Supp}(\boldsymbol{X}_v)$ (Lauritzen, 1996, Chapter 3). This permits unused bit patterns within a categorical vertex. Appendix E gives the proof, an example distinguishing the properties under structural zeros, and the reduction for chordal graphs. Write $\mathbf{S}_{I,J|S}$ for the BEGIN cross block formed from $(\boldsymbol{X}_I, \boldsymbol{X}_S, \boldsymbol{X}_J)$. Theorem 3.2 characterizes each required conditional independence by $\mathbf{S}_{I,J|S} = \mathbf{0}$.

We now ask when two such graphs can be concatenated. We call a vertex set *complete* if every two distinct vertices in it are adjacent; empty and singleton sets are complete. We write $G[U]$ for the *induced subgraph*, consisting of the vertices in U and all edges of G between them.

Theorem 3.6 (Concatenation of two BEGIN graphs)**.** *Let $G_i = (V_i, E_i)$, $i = 1, 2$, have a common vertex set $S = V_1 \cap V_2$ that is complete in both. Put $A = V_1 \setminus S$, $C = V_2 \setminus S$, and $G = (V_1 \cup V_2, E_1 \cup E_2)$.*

(a) A joint pmf p is globally Markov on G if and only if its marginals p_{V_i} are globally Markov on G_i and $\mathbf{S}_{A,C|S} = \mathbf{0}$. The zero-block condition is equivalent to

$$p(\boldsymbol{x}) = \begin{cases} p_{V_1}(\boldsymbol{x}_{V_1}) p_{V_2}(\boldsymbol{x}_{V_2}) / p_S(\boldsymbol{x}_S), & p_S(\boldsymbol{x}_S) > 0, \\ 0, & p_S(\boldsymbol{x}_S) = 0. \end{cases} \tag{3.7}$$

(b) Given pmfs q_i globally Markov on G_i, a globally Markov extension to G exists if and only if their marginals on S agree. It is unique among globally Markov extensions and is given by (3.7) *with $p_{V_i} = q_i$.*

Structural zeros are allowed in Theorem 3.6. Empty overlaps have marginal one, and empty endpoints impose no conditions. The shared variables correspond to the group intersection $\langle \boldsymbol{X}_{V_1} \rangle \cap \langle \boldsymbol{X}_{V_2} \rangle = \langle \boldsymbol{X}_S \rangle$. Part (a) checks an existing joint law; part (b) constructs one from compatible marginal laws. A tree of subgraphs gives the same conclusion when every vertex occurs along a connected part of the tree and adjacent overlaps are complete. Agreement on the overlaps then determines a unique globally Markov extension. Corollary F.2 states the precise conditions and product formula.

For the chain in Figure 1(b), agreement of adjacent pair pmfs on their shared one-variable marginals determines a unique globally Markov extension. For an existing joint law, the additional BEGIN conditions are $A_3 \perp\!\!\!\perp A_1 \mid A_2$ and $A_4 \perp\!\!\!\perp (A_1, A_2) \mid A_3$. Appendix F gives the general tree construction and proof.

Figure 1(c) also illustrates how the concatenation of two BEGIN conditions yields an Ising model. Write

$$(X_1, X_2, X_3, X_4) = (B, A_1 A_2, A_1 A_2 A_3, A_1 A_3).$$

The BEGIN condition in this panel is $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4)$. By Theorem 3.6(b), marginal laws on (X_1, X_2, X_4) and (X_2, X_3, X_4) that agree on the distribution of (X_2, X_4) determine a unique joint law satisfying this condition. The resulting graph consists of the four-cycle $1, 2, 3, 4, 1$ and the additional edge $\{2, 4\}$. If the same law also satisfies $X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3)$, this edge can be removed. Together, these two BEGIN conditions characterize the global Markov property on the four-cycle, even when the pmf has structural zeros. Under strict positivity on all sixteen states, they are equivalent to an Ising model in (X_1, X_2, X_3, X_4), with single-variable terms and pairwise terms along the cycle. Appendix F gives the construction and proof.

3.5 Binary approximation

For general variables, the binary representation gives increasingly detailed approximations. After a coordinate-wise invertible transformation, if needed, let $(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \in [-1, 1]^{r_U} \times [-1, 1]^{r_V} \times [-1, 1]^{r_W}$. For a scalar $x \in [-1, 1]$, let $A_k(x) \in \{\pm 1\}$ be fixed binary sign digits, with a consistent choice at dyadic endpoints, and write

$$x = \sum_{k=1}^{\infty} A_k(x) 2^{-k}, \qquad T_d(x) = \sum_{k=1}^{d} A_k(x) 2^{-k}.$$

At depth d, the *dyadic cells* partition $[-1, 1]$ into 2^d intervals of length 2^{1-d}. We use their midpoints for T_d and assign $x = 1$ to the last cell. Appendix H gives the explicit convention and proofs. Apply T_d coordinatewise and form $\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]$ from the first d bits per coordinate, giving endpoint lengths dr_U, dr_W and conditioning length dr_V.

Theorem 3.7 (BEGIN under dyadic approximation)**.** *For the nested binary expansions above:*

(a) If $\liminf_{d\to\infty} \|\mathbf{S}^{(d)}[\mathcal{L},\mathcal{R}]\|_{\max} = 0$*, then* $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$.

(b) Let $\mathcal{P}$ *denote probability law and write* $\mathrm{TV}(P_1,P_2) = \sup_E |P_1(E) - P_2(E)|$ *for probability laws* P_1, P_2*. Suppose versions of the conditional marginal laws satisfy, for all* $\boldsymbol{v}, \boldsymbol{v}' \in [-1,1]^{r_V}$ *and constants* $0 < \kappa \le 1$ *and* $0 \le L_{\boldsymbol{U}}, L_{\boldsymbol{W}} < \infty$*,*

$$\mathrm{TV}\{\mathcal{P}(\boldsymbol{U} \mid \boldsymbol{V} = \boldsymbol{v}), \mathcal{P}(\boldsymbol{U} \mid \boldsymbol{V} = \boldsymbol{v}')\} \le L_{\boldsymbol{U}} \|\boldsymbol{v} - \boldsymbol{v}'\|_2^{\kappa},$$

and the same condition for $\boldsymbol{W}$ *with constant* $L_{\boldsymbol{W}}$*. If* $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$*, then*

$$\|\mathbf{S}^{(d)}[\mathcal{L},\mathcal{R}]\|_{\max} \le L_{\boldsymbol{U}} L_{\boldsymbol{W}} r_V^{\kappa} 2^{2\kappa(1-d)}. \tag{3.8}$$

Under these continuity conditions, conditional independence is equivalent to convergence of the cross-block maximum norm to zero.

Part (a) follows by retaining each fixed interaction at finer resolutions and applying martingale convergence. Part (b) bounds the dependence introduced by discretizing the conditioning variable. Exact zero blocks at each finite resolution are not required. Appendix H proves both directions and shows why continuity is needed for (b).

4 Testing sparse conditional interactions

Proposition 3.3 suggests a test combining observations across conditioning states. For scalar signs A, C, write $m_A(\boldsymbol{B}) = \mathbf{E}[A \mid \boldsymbol{B}]$ and $\mathfrak{c}(\boldsymbol{B}) = \mathrm{Cov}(A, C \mid \boldsymbol{B})$. Then

$$\beta_\Gamma := \mathbf{S}[AB_\Gamma, C] = \mathbf{E}[B_\Gamma \mathfrak{c}(\boldsymbol{B})] = \mathbf{E}[\{A - m_A(\boldsymbol{B})\} C B_\Gamma]. \tag{4.1}$$

The full collection in (4.1) vanishes if and only if $\mathfrak{c}(\boldsymbol{B}) = 0$ on the support of $\boldsymbol{B}$, because the products B_Γ span every state indicator. For scalar signs, this is equivalent to conditional independence. Restricting the collection tests only the selected weighted sums and requires an assumption on the alternative for a power guarantee.

Suppose we observe n independent copies of $(A, \boldsymbol{B}, C)$ and know the conditional law of A given $\boldsymbol{B}$, as in a randomized binary assignment. Fix a set $\mathcal{T} \subseteq \{0,1\}^{r_B}$ of interaction labels before observing the data, and let $M = |\mathcal{T}| \ge 2$. The label $\mathbf{0}$ denotes the constant product and may be included. We call the test based on

$$\widehat{\beta}_\Gamma = \frac{1}{n}\sum_{i=1}^{n} \{A_i - m_A(\boldsymbol{B}_i)\} C_i B_{i,\Gamma}, \qquad T_{\max} = \max_{\Gamma \in \mathcal{T}} |\widehat{\beta}_\Gamma| \tag{4.2}$$

the *Max BEGIN test*. It estimates entries of the original BEGIN block without fitting the conditional mean of C. For calibration, hold $(\boldsymbol{B}_i, C_i)$ fixed, redraw A_i from its known conditional law, and recompute the maximum. This gives the *conditional randomization test* (CRT) of Candès et al. (2018). Appendix L gives its p value, finite-sample validity, and power bounds.

When $B_1, \ldots, B_{r_B}$ are mutually independent and *fair*, meaning that each sign equals $+1$ with probability $1/2$, their products are orthonormal and β_Γ are the expansion coefficients of $\mathfrak{c}(\boldsymbol{B})$. We call this function *h-sparse* in $\mathcal{T}$ if its expansion contains at most h nonzero terms, all indexed by $\mathcal{T}$. Its squared signal is $\mathbf{E}[\mathfrak{c}(\boldsymbol{B})^2]$. The conditional law of A may still depend on $\boldsymbol{B}$.

Theorem 4.1 (Optimal sample order under sparse interactions)**.** *Rejecting when* $T_{\max} > \{2\log(2M/\alpha)/n\}^{1/2}$ *has level at most* α *for* $A \perp\!\!\!\perp C \mid \boldsymbol{B}$*, for every conditioning distribution. For independent fair conditioning bits, consider all* h*-sparse alternatives with* $\mathbf{E}[\mathfrak{c}(\boldsymbol{B})^2] \ge \rho^2$*. For fixed* $\alpha, \delta \in (0,1)$ *with* $\alpha + \delta < 1$ *and fixed* $\varepsilon \in (0, 1/2)$*, Max BEGIN attains power* $1-\delta$ *uniformly over this class at the optimal sample order*

$$n \asymp \frac{h \log M}{\rho^2}, \qquad 1 \le h \le M^{1/2-\varepsilon}, \quad 0 < \rho\sqrt{h} \le 1.$$

The lower bound applies to every level-α *test and already holds for the known law* $\mathbf{P}(A = 1 \mid \boldsymbol{B}) = 1/2$.

The comparison constants depend only on $\alpha, \delta, \varepsilon$, and the test does not require h. With a fixed number R of conditional redraws satisfying $1/(R+1) \le \alpha$, its conditional-randomization version has the same sparse sample order (Appendix L). For interactions involving at most k conditioning bits, $M = \sum_{j=0}^{k} \binom{r_B}{j}$. Forming these features costs $O(nkM)$ and scanning them costs $O(nM)$. Sparse interactions therefore permit testing across many conditioning states without estimating 2^{r_B} separate conditional covariances. The maximizing interaction identifies the product weight attaining the largest absolute empirical sum in (4.2). Products of signs are computed by addition modulo two of their encoded bits; packed bit operations further reduce computation (Appendix M).

5 Power and Computation

We generate scalar signs A, C and eight conditioning signs $\boldsymbol{B} = (B_1, \dots, B_8)$. The nine signs $A, B_1, \dots, B_8$ are mutually independent and fair. Given $(A, \boldsymbol{B})$, the response C is a sign with conditional mean $A\mathfrak{c}(\boldsymbol{B})$. Thus $\mathbf{E}[C \mid \boldsymbol{B}] = 0$ and $\mathrm{Cov}(A, C \mid \boldsymbol{B}) = \mathfrak{c}(\boldsymbol{B})$; the null $A \perp\!\!\!\perp C \mid \boldsymbol{B}$ has $\mathfrak{c} = 0$. We scan 1, the eight individual bits, and all $\binom{8}{2} = 28$ products of two distinct bits, for 37 conditioning interactions. In the sparse alternatives, $\mathfrak{c}(\boldsymbol{B}) = 0.4h^{-1/2} \sum_{j=1}^{h} B_{2j-1}B_{2j}$ for $h = 1, 2, 4$. The dense alternative has 16 nonzero cross-pair products B_iB_j, with $i \in \{1, 2, 3, 4\}$ and $j \in \{5, 6, 7, 8\}$; their coefficients are $+0.1$ or -0.1 according to the orthogonal sign matrix specified in Appendix N. In every alternative, $\mathbf{E}[\mathfrak{c}(\boldsymbol{B})^2] = 0.16$ and $|\mathfrak{c}(\boldsymbol{B})| \le 0.8$. The comparison is between a conditional covariance with one active pairwise interaction and one with sixteen. Each setting uses 200 datasets per sample size, level .05, and 199 shared redraws from the known conditional law of A. We compare Max BEGIN with the conditional likelihood-ratio statistic G^2 (Bishop et al., 1975; Agresti, 2013), the Gaussian kernel statistic of Zhang et al. (2011), and a forest-weighted residual statistic motivated by the weighted generalized covariance measure of Scheidegger et al. (2022). All use CRT calibration (Candès et al., 2018), giving the names KCI-CRT and forest-weighted CRT for the latter two comparisons. The forest learns its weight on half the sample and tests on the other half; the other methods use all observations. Appendix N gives the statistics and implementations.

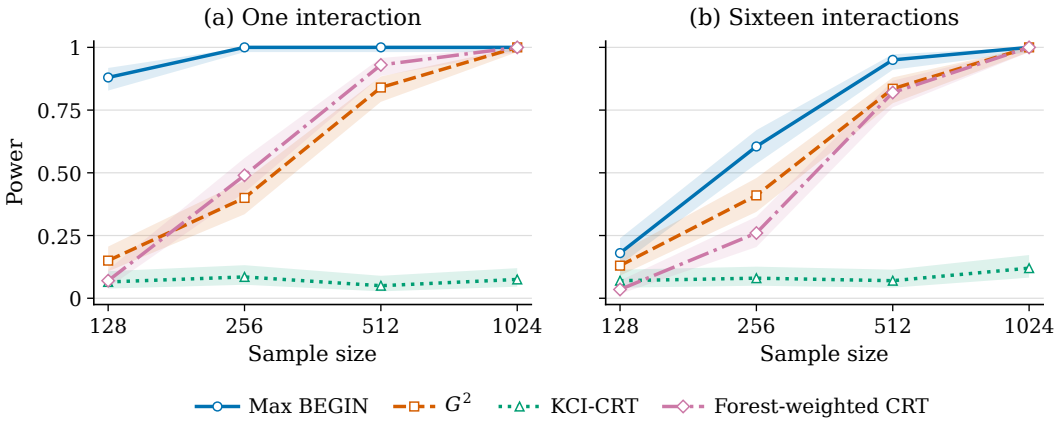


Figure 2: Power with the same squared signal distributed over one or sixteen interactions. Shaded bands are 95% Wilson intervals from 200 replications. All methods receive the same total sample size. Appendix N gives the two- and four-interaction settings, null frequencies, and timings.

Figure 2 shows the power advantage of Max BEGIN for a sparse conditional covariance. At $n = 128$ with one active interaction, its power is .880, compared with .150 for G^2, .065 for KCI-CRT, and .070 for forest-weighted CRT. Spreading the same squared signal over sixteen interactions reduces the power of Max BEGIN to .180; at $n = 512$ it reaches .950. At $n = 512$, packed Max BEGIN takes 2.40 ms, optimized G^2 takes 7.13 ms, and KCI-CRT takes 69.46 ms, including randomizations. Appendix N reports all settings, null frequencies, and timings.

6 Data analysis

The UCI Mushroom data contain hypothetical field-guide records with categorical attributes and an edibility class. We analyze 7,288 records and 18 variables, encoded with the minimal number of bits. Categorical Max BEGIN maximizes the absolute empirical cross-block entries over all endpoint and conditioning interactions. Records with identical conditioning values form a *conditioning cell*. Cell means replace the known mean in (4.2); permuting one whole category label within each cell gives exact calibration under independent, identically distributed rows.

We computed Max BEGIN for 385,101 statements from all subsets of two to six variables and bounded permutation p values for 63 statements in seven candidate subsets. Figure 3(a) shows the selected five-variable subset, each pair conditioned on the other three. Appendix O explains the detailed settings, search count, and bounds. Multiplication by $\binom{18}{2}2^{16} = 10{,}027{,}008$ accounts for selection over every pair and conditioning subset. Seven pairs reject at familywise level .05, bounding the probability of any false rejection under the sampling model. An edge records rejection; an absent edge records nonrejection.

The class is conditionally associated with odor and stalk surface above the ring. For the latter edge, Max BEGIN is .0115, with an adjusted p value bound below 10^{-28}. Only three conditioning cells contribute: odorless records with an evanescent ring and a flat, knobbed, or convex cap. All 24 silky-stalk records in these cells are labeled poisonous; all 816 fibrous-or-smooth records are labeled edible. The maximizing stalk interaction assigns opposite signs to these two groups, while the conditioning product is constant. Figure 3(b) displays the comparison and counts.

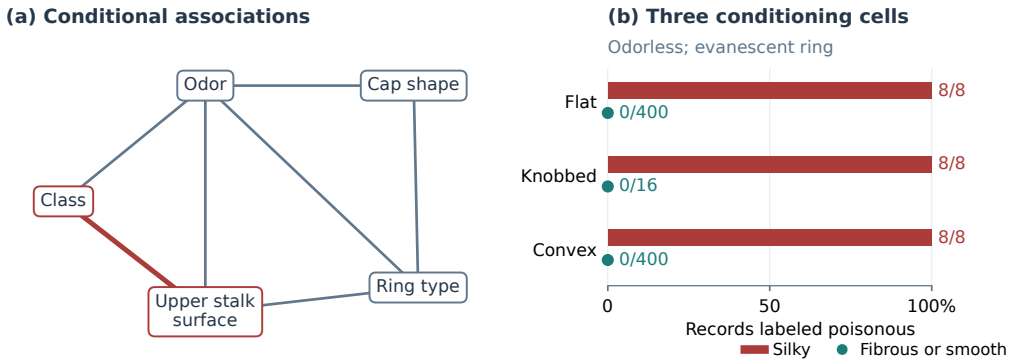


Figure 3: Mushroom analysis. (a) Seven rejected statements, each pair conditioned on the other three variables, at familywise level .05 after correction over 10,027,008 statements. An absent edge is a nonrejection. (b) Category groups explaining the red edge between class and upper stalk surface. Labels give poisonous records divided by all records.

7 Discussion

This paper establishes an exact covariance-based graphical characterization of conditional independence for arbitrary multivariate binary data, including singular multinomial encodings. Group intersections identify the conditioning information, and the generalized Schur complement records the remaining dependence for every joint pmf. Nested binary expansions extend the characterization to general variables under Theorem 3.7. Max BEGIN uses these entries for testing and interpretation.

These results suggest several directions for future work. The Schur complement and BELIEF representation suggest regression and regularization procedures for learning interaction groups. The optimal sparse sample order remains to be extended to dependent conditioning bits, estimated conditional laws, and approximately sparse alternatives. Scientific interpretation also requires further study of binary resolution and the information retained when adjusting for selected interactions. The bit-level perspective suggests resolution-dependent notions of adjustment, mediation, and causal interpretation.

AI USE STATEMENT

The foundational ideas of BEGIN, including its binary-interaction formulation, originated entirely from the authors and were developed without any AI involvement. The authors independently formulated the foundational theoretical results. These contributions include the support-based rank characterization of the interaction covariance; the equivalence between conditional independence, conditional-expectation representations, and a zero cross block of the generalized Schur complement, together with the associated covariance-block identities; the graphical interpretation using a generalized inverse; the Hadamard-prism identities; and the sparse-power arguments. The authors also established, without AI assistance, the dyadic approximation results showing that conditional independence at every binary resolution implies population conditional independence and bounding discretization error under Hölder continuity, together with counterexamples demonstrating the failure of arbitrary reductions of the conditioning block.

Generative AI tools assisted with designing and providing feedback on experiments, generating synthetic datasets for simulations, writing code to apply the methodology, and cleaning and reformatting datasets. Further uses included improving mathematical proofs, preparing figures, and improving the manuscript's wording, references, and LaTeX formatting. The authors take full responsibility for the final manuscript and accompanying materials, including all mathematical claims, proofs, numerical results, references, code, and content produced with AI assistance.

ETHICS STATEMENT

The analysis uses public, hypothetical Mushroom field-guide records. They contain no human-subject data. The class labels are analyzed as recorded and are not intended to guide mushroom consumption.

REPRODUCIBILITY STATEMENT

The appendix contains complete proofs, the bitwise algorithm, data definitions, calibration rules, and full numerical comparisons. The accompanying analysis archive contains the source data, fixed protocols, code, raw outcomes, and scripts that reproduce the tables and figures. All methods in each power comparison use the same total sample size and shared conditional randomizations.

REFERENCES

Alan Agresti. *Categorical Data Analysis*. John Wiley & Sons, 3rd edition, 2013. ISBN 978-0-470-46363-5.

Ery Arias-Castro, Emmanuel J. Candès, and Yaniv Plan. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556, 2011. doi: 10.1214/11-AOS910.

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.

Benjamin Brown, Kai Zhang, and Xiao-Li Meng. BELIEF in dependence: Leveraging atomic linearity in data bits for rethinking generalized linear models. *The Annals of Statistics*, 53(3): 1068–1094, 2025.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577, 2018. doi: 10.1111/rssb.12265.

Alexander Mangulad Christgau, Lasse Petersen, and Niels Richard Hansen. Nonparametric conditional local independence testing. *The Annals of Statistics*, 51(5):2116–2144, 2023.

Jean-Jacques Daudin. Partial association measures and an application to qualitative regression. *Biometrika*, 67(3):581–590, 1980. doi: 10.1093/biomet/67.3.581.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems*, volume 20, pp. 489–496, 2008. URL https://proceedings.neurips.cc/paper/2007/hash/3a0772443a0739141292a5429b952fe6-Abstract.html.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

Steffen L. Lauritzen. *Graphical models*. Clarendon Press, 1996.

Po-Ling Loh and Martin J Wainwright. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses. *Advances in Neural Information Processing Systems*, 25, 2012.

Mushroom [dataset]. UCI Machine Learning Repository, 1981. URL https://doi.org/10.24432/C5959T.

Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014.

Diane Valerie Ouellette. Schur complements and statistics. *Linear Algebra and its Applications*, 36: 187–295, 1981.

Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68, 2022. URL https://jmlr.org/papers/v23/21-1328.html.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. doi: 10.1214/19-AOS1857.

Audrey Terras. *Fourier analysis on finite groups and applications*. Number 43 in London Mathematical Society Student Texts. Cambridge University Press, 1999.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Kai Zhang. BET on independence. *Journal of the American Statistical Association*, 114(528): 1620–1637, 2019. doi: 10.1080/01621459.2018.1537921.

Kai Zhang, Wan Zhang, Zhigen Zhao, and Wen Zhou. BEAUTY powered BEAST. *arXiv preprint arXiv:2103.00674*, 2021. doi: 10.48550/arXiv.2103.00674.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 804–813. AUAI Press, 2011. URL https://arxiv.org/abs/1202.3775.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in Python. *Journal of Machine Learning Research*, 25(60):1–8, 2024. URL https://jmlr.org/papers/v25/23-0970.html.

APPENDIX

Appendices A–K give the representation proofs and supporting examples. Appendix L proves the sparse power and calibration results. Later appendices cover implementation and data analysis.

A PROOF OF THEOREM 3.1

Let $\boldsymbol{p} = (p_{\boldsymbol{x}})_{\boldsymbol{x}\in\{\pm1\}^p}$ be the pmf of $\boldsymbol{X}$, ordered to match the columns of $\mathbf{H}_p$. Those columns are precisely the vectors $\boldsymbol{x}^{\otimes}$, so the mean vector $\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{X}^{\otimes}]$ satisfies $\boldsymbol{\mu} = \mathbf{H}_p\boldsymbol{p}$. The covariance of the full interaction vector is therefore

$$\boldsymbol{\Sigma}_{\langle \boldsymbol{X}\rangle} = \mathbf{H}_p \mathrm{diag}(\boldsymbol{p})\mathbf{H}_p - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{A.1}$$

The second-moment identity underlying this formula is proved in Lemma G.2(c) in Appendix G. Using $\mathbf{H}_p^2 = 2^p\mathbf{I}$ gives

$$\mathbf{H}_p\boldsymbol{\Sigma}_{\langle \boldsymbol{X}\rangle}\mathbf{H}_p = 2^{2p}\big(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\big). \tag{A.2}$$

Thus $\boldsymbol{\Sigma}_{\langle \boldsymbol{X}\rangle}$ and $\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$ have the same rank.

Let $k = |\mathrm{Supp}(\boldsymbol{X})|$, and let $\boldsymbol{q} \in \mathbb{R}^k$ collect the positive masses. Deleting the zero rows and columns corresponding to unsupported states reduces the second matrix to $\mathrm{diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\top$. For every vector $\boldsymbol{a} \in \mathbb{R}^k$,

$$\boldsymbol{a}^\top\big(\mathrm{diag}(\boldsymbol{q}) - \boldsymbol{q}\boldsymbol{q}^\top\big)\boldsymbol{a} = \sum_{j=1}^{k} q_j a_j^2 - \left(\sum_{j=1}^{k} q_j a_j\right)^2.$$

This is the variance of the values $a_1, \dots, a_k$ under their positive probabilities. It vanishes exactly when all values are equal. The null space is therefore one-dimensional, and the rank is $k - 1$, including $k = 1$.

The formal identity is the first coordinate of $\boldsymbol{X}^{\otimes}$ and is identically 1, so its covariance row and column are zero. Deleting this coordinate does not change the rank. Hence

$$\mathrm{rank}\big(\boldsymbol{\Sigma}_{\langle \boldsymbol{X}\rangle\setminus\{1\}}\big) = \mathrm{rank}\big(\boldsymbol{\Sigma}_{\langle \boldsymbol{X}\rangle}\big) = |\mathrm{Supp}(\boldsymbol{X})| - 1.$$

Other formal interaction coordinates are retained even when their evaluations coincide or are constant under the distribution.

For the null space, $\boldsymbol{w}^\top\boldsymbol{\Sigma}_{\langle \boldsymbol{X}\rangle\setminus\{1\}}\boldsymbol{w} = \mathrm{Var}\big(\sum_{\Lambda\neq\mathbf{0}} w_\Lambda X_\Lambda\big)$. Since the matrix is positive semidefinite, $\boldsymbol{w}$ lies in its null space if and only if this variance vanishes, that is, if and only if $\sum_{\Lambda\neq\mathbf{0}} w_\Lambda X_\Lambda$ is almost surely constant, or equivalently the function $\boldsymbol{x} \mapsto \sum_{\Lambda\neq\mathbf{0}} w_\Lambda x_\Lambda$ is constant on $\mathrm{Supp}(\boldsymbol{X})$. □

B PROOFS OF THEOREM 3.2 AND PROPOSITION 3.3

All interaction coordinates retain their formal labels, including coordinates whose evaluations coincide almost surely. Conditional expectations below are equalities almost surely.

For any binary vector $\boldsymbol{X} \in \{\pm1\}^p$ and state $\boldsymbol{x} \in \{\pm1\}^p$,

$$\mathbf{1}_{\{\boldsymbol{X}=\boldsymbol{x}\}} = 2^{-p}\prod_{j=1}^{p}(1 + x_j X_j). \tag{B.1}$$

Expanding the product shows that the full interaction vector spans every real function of $\boldsymbol{X}$. This holds on any support, without a positivity assumption.

$(a) \iff (b)$. Conditional independence gives $\mathbf{E}[A_{\Lambda_1} \mid \boldsymbol{B}, \boldsymbol{C}] = \mathbf{E}[A_{\Lambda_1} \mid \boldsymbol{B}]$ and the symmetric equality for C_{Λ_2}. This proves (b). Each conditional expectation given $\boldsymbol{B}$ is a function of $\boldsymbol{B}$, so (B.1) expresses it as $\boldsymbol{\alpha}^\top\boldsymbol{B}^{\otimes}$ for some $\boldsymbol{\alpha} \in \mathbb{R}^{2^{r_B}}$. Conversely, either family of identities in (b) alone implies (a). Applying, say, the first family to the interaction expansion of every endpoint state indicator gives

$$\mathbf{P}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{B}, \boldsymbol{C}) = \mathbf{P}(\boldsymbol{A} = \boldsymbol{a} \mid \boldsymbol{B})$$

for every $\boldsymbol{a}$. Multiplying by $\mathbf{1}_{\{\boldsymbol{C}=\boldsymbol{c}\}}$ and conditioning on $\boldsymbol{B}$ yields the statewise conditional factorization in (a).

Conditional projection and residual covariance. Let $\boldsymbol{X}_{\mathcal{B}}, \boldsymbol{X}_{\mathcal{L}}, \boldsymbol{X}_{\mathcal{R}}$ be the interaction vectors indexed by the three blocks. Since the constant and $\boldsymbol{X}_{\mathcal{B}}$ span every function of $\boldsymbol{B}$, any square-integrable random variable Y satisfies

$$\mathbf{E}[Y \mid \boldsymbol{B}] = \mathbf{E}[Y] + \mathrm{Cov}(Y, \boldsymbol{X}_{\mathcal{B}})\boldsymbol{\Sigma}_{\mathcal{B}}^{+}(\boldsymbol{X}_{\mathcal{B}} - \mathbf{E}[\boldsymbol{X}_{\mathcal{B}}]). \tag{B.2}$$

To justify this formula when $\boldsymbol{\Sigma}_{\mathcal{B}}$ is singular, write $\mathbf{E}[Y \mid \boldsymbol{B}] = a + \boldsymbol{\beta}^{\top}\boldsymbol{X}_{\mathcal{B}}$ using the spanning identity. Taking expectations and covariances gives

$$a = \mathbf{E}[Y] - \boldsymbol{\beta}^{\top}\mathbf{E}[\boldsymbol{X}_{\mathcal{B}}], \qquad \boldsymbol{\Sigma}_{\mathcal{B}}\boldsymbol{\beta} = \mathrm{Cov}(\boldsymbol{X}_{\mathcal{B}}, Y).$$

The coefficient vector $\boldsymbol{\Sigma}_{\mathcal{B}}^{+}\,\mathrm{Cov}(\boldsymbol{X}_{\mathcal{B}}, Y)$ solves the same equation. The difference of any two solutions lies in the null space of $\boldsymbol{\Sigma}_{\mathcal{B}}$, so its linear combination of the centered features has variance zero. Thus all solutions give the same evaluated conditional expectation, proving (B.2).

Applying this identity to every left and right interaction and expanding the covariance shows that the cross block of the generalized Schur complement is

$$\mathbf{S}[\mathcal{L}, \mathcal{R}] = \mathbf{E}\Big[\big(\boldsymbol{X}_{\mathcal{L}} - \mathbf{E}[\boldsymbol{X}_{\mathcal{L}} \mid \boldsymbol{B}]\big)\big(\boldsymbol{X}_{\mathcal{R}} - \mathbf{E}[\boldsymbol{X}_{\mathcal{R}} \mid \boldsymbol{B}]\big)^{\top}\Big]. \tag{B.3}$$

Proof of Proposition 3.3. Fix nonzero Λ_1, Λ_2 and $\Gamma, \Gamma' \in \{0,1\}^{r_B}$, and write $\xi = A_{\Lambda_1}$ and $\zeta = C_{\Lambda_2}$. Because B_{Γ} is a function of $\boldsymbol{B}$,

$$\xi B_{\Gamma} - \mathbf{E}[\xi B_{\Gamma} \mid \boldsymbol{B}] = B_{\Gamma}\big(\xi - \mathbf{E}[\xi \mid \boldsymbol{B}]\big), \qquad B_{\Gamma'}\zeta - \mathbf{E}[B_{\Gamma'}\zeta \mid \boldsymbol{B}] = B_{\Gamma'}\big(\zeta - \mathbf{E}[\zeta \mid \boldsymbol{B}]\big).$$

By (B.3) and the tower property, the $(\xi B_{\Gamma}, B_{\Gamma'}\zeta)$ entry of $\mathbf{S}[\mathcal{L}, \mathcal{R}]$ is therefore

$$\mathbf{E}\Big[B_{\Gamma}B_{\Gamma'}\,\mathbf{E}\big\{\big(\xi - \mathbf{E}[\xi \mid \boldsymbol{B}]\big)\big(\zeta - \mathbf{E}[\zeta \mid \boldsymbol{B}]\big) \mid \boldsymbol{B}\big\}\Big] = \mathbf{E}\big[B_{\Gamma\oplus\Gamma'}\,\mathfrak{c}_{\Lambda_1\Lambda_2}(\boldsymbol{B})\big].$$

Writing this expectation as a sum over $\mathrm{Supp}(\boldsymbol{B})$ and using $(\mathbf{H}_{r_B})_{\Theta,\boldsymbol{b}} = b_{\Theta}$ gives (3.4). Every $\Theta \in \{0,1\}^{r_B}$ arises as $\Gamma \oplus \Gamma'$, for instance with $\Gamma = \Theta$ and $\Gamma' = \mathbf{0}$. Write $\boldsymbol{w} = \boldsymbol{w}_{\Lambda_1\Lambda_2}$. The upper bound in (3.5) follows from $|(\mathbf{H}_{r_B}\boldsymbol{w})_{\Theta}| \le \sum_{\boldsymbol{b}} |w_{\boldsymbol{b}}| = \mathbf{E}|\mathfrak{c}_{\Lambda_1\Lambda_2}(\boldsymbol{B})|$. For the lower bound, $\mathbf{H}_{r_B}^{\top}\mathbf{H}_{r_B} = 2^{r_B}\mathbf{I}$ gives $\sum_{\Theta}(\mathbf{H}_{r_B}\boldsymbol{w})_{\Theta}^2 = 2^{r_B}\|\boldsymbol{w}\|_2^2$, so the largest of these 2^{r_B} squared entries is at least $\|\boldsymbol{w}\|_2^2$. Maximizing over (Λ_1, Λ_2) completes the proof.

$(a) \iff (c)$**.** For a supported state $\boldsymbol{b}$, expanding the state indicators of $\boldsymbol{A}$ and $\boldsymbol{C}$ by (B.1) writes the conditional pmf of $(\boldsymbol{A}, \boldsymbol{C})$ given $\boldsymbol{B} = \boldsymbol{b}$ through the conditional means $\mathbf{E}[A_{\Lambda_1}C_{\Lambda_2} \mid \boldsymbol{B} = \boldsymbol{b}]$, and the product of its marginals through the products $\mathbf{E}[A_{\Lambda_1} \mid \boldsymbol{B} = \boldsymbol{b}]\,\mathbf{E}[C_{\Lambda_2} \mid \boldsymbol{B} = \boldsymbol{b}]$. Since the Walsh–Hadamard transform is invertible, the conditional pmf factorizes if and only if these agree for all Λ_1, Λ_2; agreement is automatic when Λ_1 or Λ_2 is zero. Hence (a) holds if and only if $\mathfrak{c}_{\Lambda_1\Lambda_2}(\boldsymbol{b}) = 0$ for all nonzero Λ_1, Λ_2 and all supported $\boldsymbol{b}$, that is, if and only if every $\boldsymbol{w}_{\Lambda_1\Lambda_2}$ vanishes. Since $\mathbf{H}_{r_B}$ is invertible, Proposition 3.3 shows that this is equivalent to $\mathbf{S}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$, which is (c). The augmented left block supplies the products $A_{\Lambda_1}B_{\Gamma}$ that localize the covariance to each separator state.

Covariance range identity. The required covariance range identity holds without conditional independence. If $\boldsymbol{v}$ lies in the null space of $\boldsymbol{\Sigma}_{\mathcal{B}}$, then $\boldsymbol{v}^{\top}(\boldsymbol{X}_{\mathcal{B}} - \mathbf{E}[\boldsymbol{X}_{\mathcal{B}}]) = 0$ almost surely, so its covariance with every other interaction is zero. Thus some matrices $\mathbf{M}_1, \mathbf{M}_2$ satisfy

$$\boldsymbol{\Sigma}[\mathcal{L}, \mathcal{B}] = \mathbf{M}_1\boldsymbol{\Sigma}_{\mathcal{B}}, \qquad \boldsymbol{\Sigma}[\mathcal{R}, \mathcal{B}] = \mathbf{M}_2\boldsymbol{\Sigma}_{\mathcal{B}}. \tag{B.4}$$

One choice is $\mathbf{M}_1 = \boldsymbol{\Sigma}[\mathcal{L}, \mathcal{B}]\boldsymbol{\Sigma}_{\mathcal{B}}^{+}$ and $\mathbf{M}_2 = \boldsymbol{\Sigma}[\mathcal{R}, \mathcal{B}]\boldsymbol{\Sigma}_{\mathcal{B}}^{+}$. For any matrices satisfying (B.4), the Moore–Penrose identity gives

$$\begin{aligned}\mathbf{S}[\mathcal{L}, \mathcal{R}] &= \boldsymbol{\Sigma}[\mathcal{L}, \mathcal{R}] - \mathbf{M}_1\boldsymbol{\Sigma}_{\mathcal{B}}\boldsymbol{\Sigma}_{\mathcal{B}}^{+}\boldsymbol{\Sigma}_{\mathcal{B}}\mathbf{M}_2^{\top} \\ &= \boldsymbol{\Sigma}[\mathcal{L}, \mathcal{R}] - \mathbf{M}_1\boldsymbol{\Sigma}_{\mathcal{B}}\mathbf{M}_2^{\top}.\end{aligned}$$

In particular, $\mathbf{M}_1\boldsymbol{\Sigma}_{\mathcal{B}}\mathbf{M}_2^{\top} = \boldsymbol{\Sigma}[\mathcal{L}, \mathcal{B}]\boldsymbol{\Sigma}_{\mathcal{B}}^{+}\boldsymbol{\Sigma}[\mathcal{B}, \mathcal{R}]$ does not depend on the choice of $\mathbf{M}_1$ and $\mathbf{M}_2$. The covariance identity in the definition of $\mathbf{S}$ therefore remains valid for all such choices, including singular conditioning blocks. □

C GRAPH REPRESENTATION AND GENERALIZED INVERSES

To obtain an undirected graph, write $\mathbf{F} = \mathbf{\Sigma}[\mathcal{B}, \mathcal{L} \cup \mathcal{R}]$ and define the Schur–Banachiewicz generalized inverse (Ouellette, 1981) by

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Sigma}_{\mathcal{B}}^{+} + \mathbf{\Sigma}_{\mathcal{B}}^{+}\mathbf{F}\mathbf{S}^{+}\mathbf{F}^{\top}\mathbf{\Sigma}_{\mathcal{B}}^{+} & -\mathbf{\Sigma}_{\mathcal{B}}^{+}\mathbf{F}\mathbf{S}^{+} \\ -\mathbf{S}^{+}\mathbf{F}^{\top}\mathbf{\Sigma}_{\mathcal{B}}^{+} & \mathbf{S}^{+} \end{pmatrix}. \tag{C.1}$$

Its lower-right block is exactly $\mathbf{S}^{+}$, which preserves the required separation even when the covariance is singular.

Corollary C.1 (Graph representation)**.** *The Schur–Banachiewicz inverse $\mathbf{\Omega}$ is a symmetric reflexive generalized inverse of $\mathbf{\Sigma}$: $\mathbf{\Sigma}\,\mathbf{\Omega}\,\mathbf{\Sigma} = \mathbf{\Sigma}$ and $\mathbf{\Omega}\,\mathbf{\Sigma}\,\mathbf{\Omega} = \mathbf{\Omega}$. Moreover, the following statements are equivalent.*

(a) $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B}$.

(b) $\mathbf{\Omega}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$.

(c) In the undirected graph with vertex set $\mathcal{B} \cup \mathcal{L} \cup \mathcal{R}$ and an edge between two distinct vertices whenever the corresponding entry of $\mathbf{\Omega}$ is nonzero, the set $\mathcal{B}$ separates $\mathcal{L}$ from $\mathcal{R}$.

Because the vertex set is exactly $\mathcal{B} \cup \mathcal{L} \cup \mathcal{R}$, part (c) restates (b) in graph language: $\mathcal{B}$ separates the endpoint blocks precisely when no edge joins them. The choice matters. When $\mathbf{\Sigma}$ is singular, the Moore–Penrose inverse $\mathbf{\Sigma}^{+}$ need not have a zero $(\mathcal{L}, \mathcal{R})$ block under conditional independence. For example, let B be a fair sign, $A = B$, and let C be independent of B with $\mu = \mathbf{E}[C] \in (-1, 1)\setminus\{0\}$. Then $A \perp\!\!\!\perp C \mid B$ holds trivially, yet $\mathbf{\Sigma}^{+}[A, BC] = -\mu/\{2(1-\mu^2)\} \neq 0$, as calculated below. The Schur–Banachiewicz form avoids this because it inverts the conditioning block and the Schur complement separately.

For three scalar signs, the blocks are $\mathcal{B} = \{B\}$, $\mathcal{L} = \{A, AB\}$, and $\mathcal{R} = \{C, BC\}$. The condition $A \perp\!\!\!\perp C \mid B$ is equivalent to the absence of edges between the two endpoint blocks. Edges within either block or incident to B depend on the distribution. Each such graph represents one specified conditional-independence statement; different statements can require different covariance matrices. The linear coefficients representing the conditional means in Theorem 3.2(b) can be nonunique on a restricted support, while the conditional expectations remain unique almost surely.

PROOF OF COROLLARY C.1

Write $\mathbf{F} = \mathbf{\Sigma}[\mathcal{B}, \mathcal{L} \cup \mathcal{R}]$ and $\mathbf{G} = \mathbf{\Sigma}[\mathcal{L} \cup \mathcal{R}, \mathcal{L} \cup \mathcal{R}]$, so

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{\mathcal{B}} & \mathbf{F} \\ \mathbf{F}^{\top} & \mathbf{G} \end{pmatrix}, \qquad \mathbf{S} = \mathbf{G} - \mathbf{F}^{\top}\mathbf{\Sigma}_{\mathcal{B}}^{+}\mathbf{F}.$$

By (B.4), for $\mathbf{M} = \mathbf{F}^{\top}\mathbf{\Sigma}_{\mathcal{B}}^{+}$,

$$\mathbf{F}^{\top} = \mathbf{M}\mathbf{\Sigma}_{\mathcal{B}}, \qquad \mathbf{F} = \mathbf{\Sigma}_{\mathcal{B}}\mathbf{M}^{\top}.$$

This gives the factorization

$$\mathbf{\Sigma} = \mathbf{T}\begin{pmatrix} \mathbf{\Sigma}_{\mathcal{B}} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}\mathbf{T}^{\top}, \qquad \mathbf{T} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{M} & \mathbf{I} \end{pmatrix}.$$

By the definition in (C.1),

$$\mathbf{\Omega} = \mathbf{T}^{-\top}\begin{pmatrix} \mathbf{\Sigma}_{\mathcal{B}}^{+} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{+} \end{pmatrix}\mathbf{T}^{-1}.$$

The middle matrix is symmetric, so $\mathbf{\Omega}$ is symmetric. Multiplying the two factorizations and using $\mathbf{\Sigma}_{\mathcal{B}}\mathbf{\Sigma}_{\mathcal{B}}^{+}\mathbf{\Sigma}_{\mathcal{B}} = \mathbf{\Sigma}_{\mathcal{B}}$ and $\mathbf{S}\mathbf{S}^{+}\mathbf{S} = \mathbf{S}$ gives

$$\mathbf{\Sigma}\mathbf{\Omega}\mathbf{\Sigma} = \mathbf{T}\begin{pmatrix} \mathbf{\Sigma}_{\mathcal{B}} & \mathbf{0} \\ \mathbf{0} & \mathbf{S} \end{pmatrix}\mathbf{T}^{\top} = \mathbf{\Sigma}.$$

Similarly, $\mathbf{\Sigma}_{\mathcal{B}}^{+}\mathbf{\Sigma}_{\mathcal{B}}\mathbf{\Sigma}_{\mathcal{B}}^{+} = \mathbf{\Sigma}_{\mathcal{B}}^{+}$ and $\mathbf{S}^{+}\mathbf{S}\mathbf{S}^{+} = \mathbf{S}^{+}$ give

$$\mathbf{\Omega}\mathbf{\Sigma}\mathbf{\Omega} = \mathbf{T}^{-\top}\begin{pmatrix} \mathbf{\Sigma}_{\mathcal{B}}^{+} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^{+} \end{pmatrix}\mathbf{T}^{-1} = \mathbf{\Omega},$$

so $\mathbf{\Omega}$ is a reflexive generalized inverse. It generally differs from $\mathbf{\Sigma}^+$, as the example below shows.

The lower-right block of $\mathbf{\Omega}$ is $\mathbf{S}^+$. Therefore $\mathbf{\Omega}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$ if and only if $\mathbf{S}^+$ is block diagonal with respect to $(\mathcal{L}, \mathcal{R})$. The Moore–Penrose inverse acts blockwise on block-diagonal matrices and satisfies $(\mathbf{S}^+)^+ = \mathbf{S}$. It follows that this condition is equivalent to block diagonality of $\mathbf{S}$, which is equivalent to $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B}$ by Theorem 3.2.

Finally, the graph has the disjoint vertex blocks $\mathcal{B}, \mathcal{L}, \mathcal{R}$. If $\mathbf{\Omega}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$, it has no direct edge between the left and right blocks, and any path between them must pass through $\mathcal{B}$. Conversely, a direct left-right edge would be a path avoiding $\mathcal{B}$, so separation by $\mathcal{B}$ excludes every such edge. Thus the graph condition is equivalent to the zero cross block, completing the proof. □

The Moore–Penrose inverse need not preserve the zero block. Let B be a fair sign, $A = B$, and let C be independent of B with $\mu = \mathbf{E}[C] \in (-1, 1) \setminus \{0\}$. Then $A \perp\!\!\!\perp C \mid B$, because A is a function of B. Order the features as $(B; A, AB; C, BC)$, so that $\mathcal{B} = \{B\}$, $\mathcal{L} = \{A, AB\}$, and $\mathcal{R} = \{C, BC\}$. The feature AB is identically one, so its row and column of $\mathbf{\Sigma}$ vanish. Next, $\mathrm{Var}(B) = \mathrm{Var}(A) = \mathrm{Cov}(A, B) = \mathrm{Var}(BC) = 1$, $\mathrm{Cov}(B, BC) = \mathrm{Cov}(A, BC) = \mathbf{E}[C] = \mu$, and $\mathrm{Var}(C) = 1 - \mu^2$, while C is uncorrelated with B, A, and BC because $\mathrm{Cov}(C, BC) = \mathbf{E}[B] - \mu\,\mathbf{E}[B]\,\mathbf{E}[C] = 0$. Hence, on (B, A, BC) the covariance matrix is

$$\mathbf{M} = \boldsymbol{e}\boldsymbol{e}^\top + \mu(\boldsymbol{e}\boldsymbol{f}^\top + \boldsymbol{f}\boldsymbol{e}^\top) + \boldsymbol{f}\boldsymbol{f}^\top, \qquad \boldsymbol{e} = (1, 1, 0)^\top, \quad \boldsymbol{f} = (0, 0, 1)^\top.$$

Its range is spanned by the orthogonal vectors $\boldsymbol{e}$ and $\boldsymbol{f}$, and a direct check of the four Moore–Penrose conditions gives

$$\mathbf{M}^+ = \frac{1}{2(1-\mu^2)} \Big\{ \tfrac{1}{2}\boldsymbol{e}\boldsymbol{e}^\top - \mu(\boldsymbol{e}\boldsymbol{f}^\top + \boldsymbol{f}\boldsymbol{e}^\top) + 2\boldsymbol{f}\boldsymbol{f}^\top \Big\}.$$

Because $\mathbf{\Sigma}$ is block diagonal after permuting its features, $\mathbf{\Sigma}^+$ acts blockwise, and therefore

$$\mathbf{\Sigma}^+[A, BC] = \mathbf{M}^+[A, BC] = -\frac{\mu}{2(1-\mu^2)} \neq 0.$$

In contrast, $\mathbf{\Omega}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$ by the corollary.

D Dimension under a Markov condition

Under the assumptions of Remark 3.5, put $m_j(\boldsymbol{z}) = \mathbf{E}[X_j \mid \boldsymbol{X}_{M_j} = \boldsymbol{z}]$ on the supported parent configurations and choose arbitrary values in $[-1, 1]$ elsewhere. A binary conditional distribution is determined by its mean:

$$\mathbf{P}(X_j = x_j \mid \boldsymbol{X}_{M_j} = \boldsymbol{z}) = \frac{1 + x_j m_j(\boldsymbol{z})}{2}.$$

The chain rule and the Markov condition give

$$\mathbf{P}(\boldsymbol{X} = \boldsymbol{x}) = 2^{-p} \prod_{j=1}^{p} \{1 + x_j m_j(\boldsymbol{x}_{M_j})\}. \tag{D.1}$$

This identity includes structural zeros. Indeed, along any unsupported configuration, the first unsupported prefix has a supported predecessor and a zero conditional factor. The product therefore remains zero regardless of extensions at later unsupported parent configurations. Conversely, any functions $m_j : \{\pm 1\}^{|M_j|} \to [-1, 1]$ define a normalized pmf by this product: summing from x_p backward makes each last factor sum to two. Summing the later variables also verifies the stipulated conditional distributions on supported prefixes.

For a positive pmf, all parent configurations occur, and each has one free conditional probability in $(0, 1)$. These probabilities can be chosen independently and are recovered from the joint pmf by taking conditional probabilities. The map in (D.1) and its inverse are smooth, giving exactly $D = \sum_j 2^{|M_j|}$ free parameters. The Hadamard transform changes conditional-mean values into interaction coefficients invertibly. With structural zeros, unsupported configurations can instead give redundant parameters.

For an order-three sequence and $p \geq 3$, the count is $1 + 2 + 4 + 8(p-3) = 8p - 17$. The matrix size for two scalar endpoints given r_B bits is a separate calculation: $|\mathcal{B}| = 2^{r_B} - 1$ and $|\mathcal{L}| = |\mathcal{R}| = 2^{r_B}$, giving $3 \cdot 2^{r_B} - 1$ features. Neither calculation asserts low rank for the full interaction covariance of a positive law.

E CHARACTERIZATION OF THE GLOBAL MARKOV PROPERTY

We call a complete set a *maximal clique* if it is not contained in a larger complete set. We call a graph *chordal* if every cycle of length at least four has an edge joining two nonconsecutive vertices.

Relations between the Markov properties. The neighbors of a vertex separate it from all its nonneighbors, so the global property implies the local property. For nonadjacent u, v, let $R = V \setminus (\{u, v\} \cup N(u))$. The local property gives $\boldsymbol{X}_u \perp\!\!\!\perp (\boldsymbol{X}_v, \boldsymbol{X}_R) \mid \boldsymbol{X}_{N(u)}$. Moving $\boldsymbol{X}_R$ into the conditioning vector gives the pairwise property. Under positivity on the product of vertex supports, part (b) below proves that pairwise implies global. Thus they are equivalent under positivity.

For an example with structural zeros, let G consist of the two disconnected edges $\{1, 2\}$ and $\{3, 4\}$, and set $X_1 = X_2 = X_3 = X_4 = U$ for a fair sign U. Every vertex is determined by its neighbor, so the local property holds. The global property would require (X_1, X_2) and (X_3, X_4) to be independent, because the empty set separates the two components. The global property fails: all four signs equal one with probability $1/2$, whereas the component probabilities multiply to $1/4$.

Theorem E.1 (Global Markov property)**.** *In each of the following cases, the distribution of $\boldsymbol{X}$ is globally Markov on G if and only if $\mathbf{S}_{I,J|S} = \mathbf{0}$ for every listed triple (I, S, J).*

(a) For an arbitrary joint pmf and an arbitrary graph, use every triple for which S separates I from J. Equivalently, it suffices to use, for every $S \subseteq V$ and every connected component K of the graph $G - S$ obtained by deleting S, the triple $\big(K,\ S,\ V \setminus (S \cup K)\big)$ whenever $V \setminus (S \cup K) \neq \varnothing$.

(b) If $\mathbf{P}(\boldsymbol{X} = \boldsymbol{x}) > 0$ for every $\boldsymbol{x} \in \prod_{v \in V} \mathrm{Supp}(\boldsymbol{X}_v)$, in particular if the joint pmf is strictly positive, use

$$(I, S, J) = \big(\{u\},\ V \setminus \{u, v\},\ \{v\}\big)$$

for every pair of distinct nonadjacent vertices u, v.

(c) If G is chordal, order its maximal cliques $C_1, \ldots, C_m$ so that, for $j \geq 2$,

$$H_{j-1} = \bigcup_{i<j} C_i, \qquad S_j = C_j \cap H_{j-1}, \qquad R_j = C_j \setminus S_j,$$

and each S_j is contained in an earlier clique. Use the $m - 1$ triples

$$(I, S, J) = \big(R_j,\ S_j,\ H_{j-1} \setminus S_j\big), \qquad j = 2, \ldots, m.$$

Parts (a) and (c) allow arbitrary structural zeros, and part (b) allows unused binary patterns within individual vertex encodings. Empty conditioning sets give unconditional independence.

Part (a) translates graph separation into the covariance characterization of BEGIN; its reduced form uses one triple for each deleted set and each resulting connected component. The positivity condition in part (b) allows all graph separations to be deduced from the missing-edge conditions. It is imposed only on the product of the vertex supports, so it permits the structural zeros that every binary encoding of a multinomial variable with $K < 2^b$ categories creates within a vertex; it excludes only zeros produced by dependence between different vertices. Chordal structure gives a different reduction: part (c) requires only one condition for each added clique, with no positivity assumption. In that condition, the new variables are independent of all previously included variables outside the overlap. Parts (b) and (c) are the classical pairwise-to-global and decomposable-graph results (Lauritzen, 1996, Chapter 3) expressed through Theorem 3.2. Orderings as in (c) exist if and only if G is chordal. The proof follows below.

For four scalar signs on the cycle $1, 2, 3, 4, 1$, the only nontrivial separation conditions are

$$X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4), \qquad X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3).$$

By Theorem E.1(a), the two corresponding BEGIN cross blocks vanish if and only if the joint law is globally Markov on this cycle. This equivalence holds even when the pmf has structural zeros. Proposition F.1 in Appendix F gives the corresponding group intersections and the Ising representation under strict positivity, followed by a construction from compatible marginal laws.

PROOF OF THEOREM E.1

For every triple (I, S, J), Theorem 3.2 gives

$$\mathbf{S}[\mathcal{L}, \mathcal{R}] = \mathbf{0} \quad \Longleftrightarrow \quad \boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_S,$$

where the interaction index sets and the Schur complement are constructed separately for that triple. It therefore suffices to prove the corresponding conditional-independence characterizations. The assertions are immediate if $V = \varnothing$; hence assume $V \neq \varnothing$.

Part (a). The first list is the definition of the global Markov property combined with Theorem 3.2. Every triple in the reduced list is a separation, so the reduced conditions are necessary. Conversely, suppose that they hold, and let S separate I from J, with I and J nonempty. Let $K_1, \ldots, K_q$ be the connected components of $G - S$; no component meets both I and J, so $q \geq 2$. The reduced conditions give $\boldsymbol{X}_{K_i} \perp\!\!\!\perp \boldsymbol{X}_{\bigcup_{l \neq i} K_l} \mid \boldsymbol{X}_S$ for each i. Applying these for $i = 1, \ldots, q-1$ in turn, and discarding $K_1, \ldots, K_{i-1}$ from the other endpoint vector at step i by the decomposition property, shows that $\boldsymbol{X}_{K_1}, \ldots, \boldsymbol{X}_{K_q}$ are mutually conditionally independent given $\boldsymbol{X}_S$. No component meets both I and J, so I and J lie in the unions of two disjoint collections of components, and marginalizing gives $\boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_S$. No positivity assumption is used.

Part (b). If u and v are distinct nonadjacent vertices, then $V \setminus \{u, v\}$ separates them. Hence the global Markov property implies

$$\boldsymbol{X}_u \perp\!\!\!\perp \boldsymbol{X}_v \mid \boldsymbol{X}_{V \setminus \{u,v\}},$$

which proves necessity.

For sufficiency, we first record the intersection property in the form needed here. Let $\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{Z}$ be the random vectors at pairwise disjoint vertex sets. Under the positivity assumption in (b), the support of $\boldsymbol{X}$ is $\prod_{v \in V} \mathrm{Supp}(\boldsymbol{X}_v)$, so the support of $(\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{Z})$ is the product of the supports of $\boldsymbol{Y}$, $\boldsymbol{U}$, $\boldsymbol{W}$, and $\boldsymbol{Z}$. Then

$$\left.\begin{array}{l} \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{U} \mid (\boldsymbol{W}, \boldsymbol{Z}), \\ \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W} \mid (\boldsymbol{U}, \boldsymbol{Z}) \end{array}\right\} \quad \Longrightarrow \quad \boldsymbol{Y} \perp\!\!\!\perp (\boldsymbol{U}, \boldsymbol{W}) \mid \boldsymbol{Z}.$$

To verify this property directly, note that every conditional probability below is well defined when $(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z})$ lies in the product of the supports. The two assumed conditional independences imply, for every $\boldsymbol{y}$ and every such $(\boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z})$,

$$p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{z}).$$

Fixing one supported value $\boldsymbol{u}_0$ shows that $p(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{z}) = p(\boldsymbol{y} \mid \boldsymbol{u}_0, \boldsymbol{z})$ for every supported $\boldsymbol{w}$. Hence $p(\boldsymbol{y} \mid \boldsymbol{u}, \boldsymbol{w}, \boldsymbol{z})$ is constant as $(\boldsymbol{u}, \boldsymbol{w})$ varies over $\mathrm{Supp}(\boldsymbol{U}) \times \mathrm{Supp}(\boldsymbol{W})$. Averaging over $(\boldsymbol{U}, \boldsymbol{W})$ conditional on $\boldsymbol{Z} = \boldsymbol{z}$ identifies this constant as $p(\boldsymbol{y} \mid \boldsymbol{z})$. Thus $\boldsymbol{Y} \perp\!\!\!\perp (\boldsymbol{U}, \boldsymbol{W}) \mid \boldsymbol{Z}$, as required. Some positivity is needed: if all three one-dimensional vectors $\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{W}$ equal the same nondegenerate sign and $\boldsymbol{Z}$ is constant, both hypotheses hold but the conclusion fails.

Now suppose the stated BEGIN conditions hold. By Theorem 3.2,

$$\boldsymbol{X}_u \perp\!\!\!\perp \boldsymbol{X}_v \mid \boldsymbol{X}_{V \setminus \{u,v\}}$$

for every pair of distinct nonadjacent vertices. Consider any graphical separation $I - S - J$. Let A be the set of vertices reachable from I by paths that do not pass through S, and set

$$D = V \setminus (A \cup S).$$

Then $I \subseteq A$, $J \subseteq D$, and no edge joins a vertex in A to a vertex in D.

Consequently, for every $u \in A$ and $v \in D$,

$$\boldsymbol{X}_u \perp\!\!\!\perp \boldsymbol{X}_v \mid \big(\boldsymbol{X}_{A \setminus \{u\}}, \boldsymbol{X}_{D \setminus \{v\}}, \boldsymbol{X}_S\big).$$

For each fixed u, repeated application of the intersection property over the vertices of D gives

$$\boldsymbol{X}_u \perp\!\!\!\perp \boldsymbol{X}_D \mid \big(\boldsymbol{X}_{A \setminus \{u\}}, \boldsymbol{X}_S\big).$$

Applying the intersection property again, now over the vertices of A, yields

$$\boldsymbol{X}_D \perp\!\!\!\perp \boldsymbol{X}_A \mid \boldsymbol{X}_S.$$

Marginalizing to $I \subseteq A$ and $J \subseteq D$ therefore gives

$$\boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_S.$$

Since the graphical separation was arbitrary, $\boldsymbol{X}$ satisfies the global Markov property.

Part (c): necessity. Fix $j \geq 2$. We first show that S_j separates R_j from $H_{j-1} \setminus S_j$ in the full graph G. Build G by adding its maximal cliques in the given order. Let G_ℓ be the graph formed by the first ℓ cliques, with vertex set $H_\ell = \bigcup_{i \leq \ell} C_i$.

At step j, the vertices in R_j are new, and C_j meets the preceding graph exactly in S_j. Thus the new vertices can connect to the preceding vertices only through S_j. Deleting S_j therefore separates R_j from $H_{j-1} \setminus S_j$ in G_j.

We must check that adding later cliques cannot destroy this separation. Consider a later clique C_ℓ, where $\ell > j$. Its overlap with the graph already constructed is S_ℓ, and the running-intersection property places all of S_ℓ inside one earlier clique. Consequently, after deleting the fixed set S_j, the surviving overlap $S_\ell \setminus S_j$ is either empty or lies entirely in one connected component of $G_{\ell-1} - S_j$.

Adding C_ℓ therefore attaches its new vertices to at most one existing component. It cannot join two previously separate components. Repeating this argument for every later clique shows that the separation established at step j persists in the final graph $G_m = G$.

Hence, if $\boldsymbol{X}$ is globally Markov with respect to G, then

$$\boldsymbol{X}_{R_j} \perp\!\!\!\perp \boldsymbol{X}_{H_{j-1} \setminus S_j} \mid \boldsymbol{X}_{S_j}, \qquad j = 2, \ldots, m.$$

Theorem 3.2 gives the corresponding zero BEGIN cross blocks.

Part (c): sufficiency. Suppose the stated BEGIN cross blocks are zero. By Theorem 3.2,

$$\boldsymbol{X}_{R_j} \perp\!\!\!\perp \boldsymbol{X}_{H_{j-1} \setminus S_j} \mid \boldsymbol{X}_{S_j}, \qquad j = 2, \ldots, m.$$

In words, once the overlap S_j is known, the remaining earlier variables do not affect the conditional distribution of the new variables R_j.

Let p_U denote the marginal pmf of $\boldsymbol{X}_U$. For every preceding configuration with positive probability, the displayed conditional independence gives

$$p(\boldsymbol{x}_{R_j} \mid \boldsymbol{x}_{H_{j-1}}) = p(\boldsymbol{x}_{R_j} \mid \boldsymbol{x}_{S_j}).$$

Multiplying by the probability of that preceding configuration yields

$$p_{H_j}(\boldsymbol{x}_{H_j}) = p_{H_{j-1}}(\boldsymbol{x}_{H_{j-1}}) p(\boldsymbol{x}_{R_j} \mid \boldsymbol{x}_{S_j}).$$

This identity also holds when the preceding configuration has probability zero: the joint probability on the left and the product on the right are then both zero. Conditional distributions at separator values of probability zero may therefore be defined arbitrarily as probability distributions.

Starting with $H_1 = C_1$ and applying this identity successively gives

$$p(\boldsymbol{x}) = p_{C_1}(\boldsymbol{x}_{C_1}) \prod_{j=2}^{m} p(\boldsymbol{x}_{R_j} \mid \boldsymbol{x}_{S_j}).$$

Each factor involves only the variables in one clique, because $R_j \cup S_j = C_j$.

We now show why this factorization implies the global Markov property. Take any separation $I - S - J$. Let A be the set of vertices reachable from I by paths avoiding S, and let

$$D = V \setminus (A \cup S).$$

Then $I \subseteq A$, $J \subseteq D$, and no edge joins A to D.

A clique cannot contain a vertex from each of A and D, since those two vertices would have to be adjacent. Every clique factor therefore depends only on $(\boldsymbol{x}_A, \boldsymbol{x}_S)$ or only on $(\boldsymbol{x}_D, \boldsymbol{x}_S)$. Collecting the factors on the two sides gives

$$p(\boldsymbol{x}) = f(\boldsymbol{x}_A, \boldsymbol{x}_S)\, g(\boldsymbol{x}_D, \boldsymbol{x}_S)$$

for nonnegative functions f and g.

For any $\boldsymbol{x}_S$ with $p_S(\boldsymbol{x}_S) > 0$, summing over $\boldsymbol{x}_A, \boldsymbol{x}_D$ gives

$$p_S(\boldsymbol{x}_S) = \left(\sum_{\boldsymbol{y}_A} f(\boldsymbol{y}_A, \boldsymbol{x}_S) \right) \left(\sum_{\boldsymbol{y}_D} g(\boldsymbol{y}_D, \boldsymbol{x}_S) \right).$$

Dividing by this positive quantity yields

$$p(\boldsymbol{x}_A, \boldsymbol{x}_D \mid \boldsymbol{x}_S) = \frac{f(\boldsymbol{x}_A, \boldsymbol{x}_S)}{\sum_{\boldsymbol{y}_A} f(\boldsymbol{y}_A, \boldsymbol{x}_S)} \frac{g(\boldsymbol{x}_D, \boldsymbol{x}_S)}{\sum_{\boldsymbol{y}_D} g(\boldsymbol{y}_D, \boldsymbol{x}_S)}.$$

The first factor is a probability distribution on $\boldsymbol{x}_A$, and the second is a probability distribution on $\boldsymbol{x}_D$. Thus

$$\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_D \mid \boldsymbol{X}_S.$$

Marginalizing to $I \subseteq A$ and $J \subseteq D$ gives

$$\boldsymbol{X}_I \perp\!\!\!\perp \boldsymbol{X}_J \mid \boldsymbol{X}_S.$$

Since the separation was arbitrary, $\boldsymbol{X}$ is globally Markov with respect to G. □

F PROOFS FOR BEGIN CONCATENATION

A conditional component characterization. We first record an elementary fact valid for every finite pmf. A law is globally Markov on a graph if and only if, for every vertex set T and every $\boldsymbol{x}_T$ with positive probability, its conditional law given $\boldsymbol{X}_T = \boldsymbol{x}_T$ is a product of pmfs over the connected components of the graph after deleting T. To prove the forward direction, each component is separated from all remaining components by T. Apply these conditional independences successively to obtain their joint product. Conversely, two sets separated by T occupy disjoint collections of components; marginalizing the product gives their conditional independence. This argument uses neither positivity of all cells nor a clique factorization.

F.1 PROOF OF THEOREM 3.6

Proof. Because S is complete in both pieces, their edges on S agree. All other edges of either piece involve a vertex unique to that piece. Consequently $G[V_i] = G_i$, and there are no edges between A and C.

Part (a): necessity. Suppose p is globally Markov on G. We first show that its V_1 marginal is globally Markov on G_1. Let T separate sets I and J inside G_1. If a path in G joined I to J while avoiding T, every excursion of that path through C would have its endpoints in S. Replace each excursion with distinct endpoints by the edge joining them in the complete set S, and delete excursions with identical endpoints. The result would be a walk, and hence a path, in G_1 avoiding T, a contradiction. Thus this separation also holds in G. The conditional independence it gives under p is exactly the required one under its V_1 marginal. The argument for V_2 is the same.

The set S separates A from C in G. Hence $\boldsymbol{X}_A \perp\!\!\!\perp \boldsymbol{X}_C \mid \boldsymbol{X}_S$, which is equivalent to $\mathbf{S}_{A,C|S} = \mathbf{0}$ by Theorem 3.2. This proves the necessity of the global Markov properties on the pieces and the zero-block condition. Conditional independence is equivalent to (3.7) on every positive separator cell. If $p_S(\boldsymbol{x}_S) = 0$, all joint cells with this separator value have probability zero. This proves the equivalence with the displayed product formula, including zero cells.

Part (a): sufficiency. Suppose the global Markov properties on the pieces and the product formula hold. Fix $T \subseteq V_1 \cup V_2$ and $\boldsymbol{x}_T$ with $\mathbf{P}(\boldsymbol{X}_T = \boldsymbol{x}_T) > 0$. Put $T_i = T \cap V_i$. The two local conditioning events $\boldsymbol{X}_{T_i} = \boldsymbol{x}_{T_i}$ also have positive probability. Since each local law is globally Markov, its conditional pmf factors over the components of $G_i \setminus T_i$ by the component characterization above.

If $S \setminus T = \varnothing$, the denominator in (3.7) is the fixed positive number $p_S(\boldsymbol{x}_S)$. After removing T, the two pieces are disjoint, so their local components are exactly the components of $G \setminus T$. The conditional joint pmf is therefore proportional to a product of nonnegative functions, one for each component. Normalizing these factors gives a product of pmfs.

If $S \setminus T \neq \varnothing$, all its vertices lie in a single component D of $G \setminus T$, because they form a complete set. Every local component meeting $S \setminus T$ lies in D. Every other local component is a separate component of $G \setminus T$: a path to the other piece would have to reach $S \setminus T$, and no such local path exists. The denominator in (3.7) depends only on variables in D and fixed values in $\boldsymbol{x}_T$.

Combine all local factors belonging to D with the reciprocal denominator, and assign this combined factor the value zero whenever $p_S(\boldsymbol{x}_S) = 0$. Every other factor belongs to one component of $G \setminus T$.

This representation equals the conditional pmf up to a positive normalizing constant: on positive separator cells it follows from (3.7), and on zero separator cells both sides are zero. Variables in $S \backslash T$ that occur in both local factors are all inside D, so different component factors have disjoint variable sets. The conditional pmf has total mass one. Consequently the finite sum of each nonnegative component factor is positive, and normalizing the factors gives a product of component pmfs. The component characterization proves global Markov on G.

Part (b). Agreement of the marginals on S is necessary for any joint extension. If they agree, denote their common overlap marginal by q_S and define q by (3.7), with $p_{V_i} = q_i$ and $p_S = q_S$. For each positive separator cell, summing over $\boldsymbol{x}_C$ recovers q_1, because the corresponding sum of q_2 is $q_S(\boldsymbol{x}_S)$. When $q_S(\boldsymbol{x}_S) = 0$, nonnegativity gives $q_1(\boldsymbol{x}_A, \boldsymbol{x}_S) = q_2(\boldsymbol{x}_S, \boldsymbol{x}_C) = 0$ for all $\boldsymbol{x}_A, \boldsymbol{x}_C$, so the same marginal identity holds. Summing over $\boldsymbol{x}_A$ similarly recovers q_2, and summing either marginal proves that q has total mass one. Its conditional law given $\boldsymbol{X}_S$ is a product on every positive separator cell. Part (a) therefore proves that q is globally Markov on G.

Conversely, any globally Markov extension must make the two sets of variables outside the overlap conditionally independent given $\boldsymbol{X}_S$, since S separates them. Its specified marginals then force (3.7). This proves uniqueness. The same argument includes an empty separator, whose single marginal probability is one, and empty residual sets. □

F.2 Two BEGIN conditions and a four-cycle Ising model

Proposition F.1 (Two BEGIN conditions on a four-cycle)**.** *Let $\boldsymbol{X} = (X_1, X_2, X_3, X_4) \in \{\pm 1\}^4$ have joint pmf p. Consider the two BEGIN cross blocks associated with*

$$\langle X_1, X_2, X_4\rangle \cap \langle X_2, X_3, X_4\rangle = \langle X_2, X_4\rangle,$$
$$\langle X_1, X_2, X_3\rangle \cap \langle X_1, X_3, X_4\rangle = \langle X_1, X_3\rangle.$$

Both blocks are formed under p. The conditions

$$\mathbf{S}_{\{1\},\{3\}|\{2,4\}} = \mathbf{0}, \qquad \mathbf{S}_{\{2\},\{4\}|\{1,3\}} = \mathbf{0} \tag{F.1}$$

hold if and only if p is globally Markov on the four-cycle with edges $\{1,2\}$, $\{2,3\}$, $\{3,4\}$, and $\{1,4\}$. Structural zeros are allowed.

If $p(\boldsymbol{x}) > 0$ for every $\boldsymbol{x} \in \{\pm 1\}^4$, these conditions are also equivalent to an Ising representation. Specifically, there exist real coefficients $\theta_1, \ldots, \theta_4, \theta_{12}, \theta_{23}, \theta_{34}, \theta_{14}$ such that, with Z denoting the normalizing constant,

$$p(\boldsymbol{x}) = \frac{1}{Z} \exp\left\{\sum_{i=1}^{4} \theta_i x_i + \theta_{12} x_1 x_2 + \theta_{23} x_2 x_3 + \theta_{34} x_3 x_4 + \theta_{14} x_1 x_4\right\}. \tag{F.2}$$

Proof. By Theorem 3.2, (F.1) is equivalent to

$$X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4), \qquad X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3). \tag{F.3}$$

The only vertex sets whose deletion separates two remaining vertices of the four-cycle are $\{2,4\}$ and $\{1,3\}$. Thus (F.3) lists all its nontrivial separation conditions. Theorem E.1(a) proves the equivalence with the global Markov property, including structural zeros.

Now assume strict positivity. Write p_{124}, p_{234}, and p_{24} for the marginals on the indicated coordinates. The first conditional independence gives

$$\log p(\boldsymbol{x}) = \log p_{124}(x_1, x_2, x_4) + \log p_{234}(x_2, x_3, x_4) - \log p_{24}(x_2, x_4).$$

By uniqueness of the binary interaction expansion in (K.1), every coefficient of an interaction containing both x_1 and x_3 is zero. The second conditional independence similarly eliminates every interaction containing both x_2 and x_4. Every interaction of order three or four contains at least one of these two pairs. Hence only the constant, the four single-variable terms, and the four cycle-edge terms remain, giving (F.2).

Conversely, in (F.2), fixing (x_2, x_4) makes the exponent a sum of a function of x_1 and a function of x_3. Normalizing therefore gives $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4)$. Fixing (x_1, x_3) gives $X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3)$ in the same way. □

Construction from compatible marginals. Let q_{124} and q_{234} be pmfs on (X_1, X_2, X_4) and (X_2, X_3, X_4), respectively. Suppose their marginals on (X_2, X_4) agree, and denote this common marginal by q_{24}. The complete graphs on $\{1, 2, 4\}$ and $\{2, 3, 4\}$ have the complete overlap $\{2, 4\}$. Theorem 3.6(b) therefore gives the joint pmf

$$p(\boldsymbol{x}) = \begin{cases} \dfrac{q_{124}(x_1, x_2, x_4) q_{234}(x_2, x_3, x_4)}{q_{24}(x_2, x_4)}, & q_{24}(x_2, x_4) > 0, \\ 0, & q_{24}(x_2, x_4) = 0. \end{cases} \tag{F.4}$$

This is the unique joint law with the prescribed marginals satisfying $X_1 \perp\!\!\!\perp X_3 \mid (X_2, X_4)$. The union of the two complete graphs contains the additional edge $\{2, 4\}$. The same law is globally Markov after removing this edge if and only if it also satisfies $X_2 \perp\!\!\!\perp X_4 \mid (X_1, X_3)$.

For each fixed (x_1, x_3), the four probabilities obtained by varying (x_2, x_4) form a nonnegative 2×2 table. The second conditional independence holds exactly when every such table has determinant zero:

$$\begin{aligned} &p(x_1, 1, x_3, 1)\, p(x_1, -1, x_3, -1) \\ &\quad = p(x_1, 1, x_3, -1)\, p(x_1, -1, x_3, 1), \qquad x_1, x_3 \in \{\pm 1\}. \end{aligned} \tag{F.5}$$

For a table with positive total mass, this is equivalent to factorization of its normalized conditional pmf. A table with zero total mass imposes no conditional independence restriction and automatically satisfies the identity. Thus the criterion also allows structural zeros. If q_{124} and q_{234} are strictly positive on their binary cubes, (F.4) is strictly positive on all sixteen states and is an Ising model on the four-cycle if and only if (F.5) holds.

Connection with Figure 1(c). The transformation

$$(X_1, X_2, X_3, X_4) = (B, A_1 A_2, A_1 A_2 A_3, A_1 A_3)$$

is invertible, with

$$B = X_1, \qquad A_1 = X_2 X_3 X_4, \qquad A_2 = X_3 X_4, \qquad A_3 = X_2 X_3.$$

Consequently, the BEGIN condition in Figure 1(c) is the first condition in (F.3). Adding the second condition gives the global Markov property on the four-cycle. Under strict positivity, the Ising representation is in the transformed variables $\boldsymbol{X}$. It can involve higher-order interactions in the original variables; for example, $X_1 X_2 = B A_1 A_2$.

F.3 Concatenation along a tree

The two-piece result can be applied repeatedly. We call the arrangement a *running-intersection tree* when every variable that belongs to two pieces also belongs to every piece on the path between them. We refer to this requirement as the *running-intersection property*. This ensures that an added piece meets the previous pieces only through its intersection with its parent.

Corollary F.2 (Concatenation along a tree)**.** *Let $G = (V, E)$ be a finite undirected graph. Let $C_1, \dots, C_m$, with $m \geq 1$, cover V, and suppose every edge of G has both endpoints in at least one C_i. Suppose a tree on these sets has the running-intersection property and each intersection of adjacent sets is complete in G.*

Root the tree at C_1 and order the sets with each parent before its children. For $j = 2, \dots, m$, put

$$H_{j-1} = \bigcup_{i<j} C_i, \qquad S_j = C_j \cap H_{j-1}, \qquad R_j = C_j \setminus S_j.$$

Then S_j is the intersection of C_j with its parent.

(a) A joint pmf p is globally Markov on G if and only if each p_{C_i} is globally Markov on $G[C_i]$ and

$$\mathbf{S}_{R_j,\, H_{j-1} \setminus S_j \mid S_j} = \mathbf{0}, \qquad j = 2, \dots, m. \tag{F.6}$$

Each block is formed from the marginal on $H_{j-1} \cup C_j$.

(b) Suppose pmfs q_i on C_i are given and are globally Markov on $G[C_i]$. A globally Markov joint law on G with these marginals exists if and only if adjacent pieces have equal marginals on their intersections. Write q_{S_j} for these common marginals. The extension is unique among globally Markov laws and equals

$$q(\boldsymbol{x}) = \begin{cases} \dfrac{\prod_{i=1}^{m} q_i(\boldsymbol{x}_{C_i})}{\prod_{j=2}^{m} q_{S_j}(\boldsymbol{x}_{S_j})}, & q_{S_j}(\boldsymbol{x}_{S_j}) > 0 \text{ for every } j, \\ 0, & \text{otherwise}. \end{cases} \tag{F.7}$$

Structural zeros are allowed. An empty separator has marginal one, and for $m = 1$ the denominator is the empty product, equal to one.

The graph inside each piece may contain cycles and need not be complete. The tree describes how the pieces overlap. When the pieces are maximal cliques of a chordal graph, their global Markov properties impose no restrictions, and part (a) reduces to Theorem E.1(c). Thus the result extends the clique construction to larger pieces that already satisfy their own Markov properties.

For a first-order chain, take $C_j = \{j, j+1\}$. A globally Markov joint law with prescribed pair marginals exists if and only if neighboring pairs agree on their shared one-variable marginal, and it is then unique. For an existing joint law, the required conditions are

$$X_j \perp\!\!\!\perp (X_1, \ldots, X_{j-2}) \mid X_{j-1}, \qquad j = 3, \ldots, p.$$

These are the same Markov conditions used for the dimension reduction in Remark 3.5 with $r_{\max} = 1$. They involve all earlier variables outside the overlap.

PROOF OF COROLLARY F.2

Proof. We first justify the graph structure used in the induction. The case $m = 1$ is immediate. For $m > 1$, take the last set C_m in the stated ordering. It is a leaf of the tree, because a child would have a later index. Its intersection with the union of the remaining sets equals its intersection with its parent. Indeed, any vertex present in C_m and another set must, by running intersection along their connecting path, also be present in the parent. Thus this intersection is S_m and is complete.

No edge joins $R_m = C_m \setminus S_m$ to $H_{m-1} \setminus S_m$. If such an edge existed, edge coverage would put its endpoints together in some set. A vertex in R_m occurs in no remaining set, whereas the other endpoint is not in C_m, a contradiction. Furthermore, every edge with both endpoints in H_{m-1} is still covered after removing C_m. If its only previous covering set was C_m, both endpoints would lie in S_m. The parent contains this complete set and hence also contains that edge. Therefore the graph on the remaining vertices is precisely $G[H_{m-1}]$, its edges are covered by the remaining sets, and all the tree assumptions remain valid. Repeating the argument in reverse index order gives, for each j, a decomposition of $G[H_j]$ into $G[H_{j-1}]$ and $G[C_j]$ along S_j, where $H_j = H_{j-1} \cup C_j$. It also proves that S_j is the parent intersection as stated in the corollary.

Part (a): necessity. Starting with the full globally Markov law, apply Theorem 3.6(a) to the final leaf decomposition. It gives the global Markov property for the marginals on C_m and H_{m-1}, and the condition in (F.6) for $j = m$. Repeat on H_{m-1}, then on successive earlier prefixes. This yields the global Markov properties of all pieces and all $m - 1$ conditions, computed under the corresponding prefix marginals as required.

Part (a): sufficiency. Begin with p_{C_1}, which is globally Markov on $G[C_1]$ by assumption. Suppose the existing marginal on H_{j-1} has already been shown globally Markov. Its overlap with C_j is the complete set S_j; the marginal on C_j is globally Markov, and the joint marginal on H_j satisfies the jth condition. The graph decomposition above and Theorem 3.6(a) show that this particular marginal on H_j is globally Markov. Induction reaches the full law.

Part (b). A joint law necessarily gives equal overlap marginals to adjacent sets. Suppose these equalities hold. Start with q_1. At step j, the previously constructed law on H_{j-1} has the prescribed marginal on the parent of C_j and therefore has marginal q_{S_j} on S_j. Theorem 3.6(b) constructs a unique globally Markov extension to H_j with the prescribed new marginal q_j, retaining all earlier marginals. This proves existence by induction.

For uniqueness, a globally Markov extension on H_j has a globally Markov marginal on H_{j-1} by the leaf decomposition and Theorem 3.6(a). By induction that earlier marginal is already uniquely determined, and the two-piece uniqueness result then determines the extension on H_j. Thus the full extension is unique.

Multiplying the successive two-piece formulas gives (F.7) when every separator denominator is positive. If some $q_{S_j}(\boldsymbol{x}_{S_j})$ is zero, any joint extension with the prescribed marginals gives probability zero to every configuration with that separator value. This proves the second branch of the formula without assigning a value to a ratio with zero denominator. Empty separators contribute factors equal to one. □

G THE HADAMARD PRISM AND ITS PROPERTIES

The Hadamard prism. The interaction covariance also has a direct algebraic form. For an integer $p \geq 0$ and $\boldsymbol{y} \in \mathbb{R}^{2^p}$, write $\mathrm{diag}(\boldsymbol{y})$ for the diagonal matrix with diagonal $\boldsymbol{y}$. We call the matrix

$$\eta_p(\boldsymbol{y}) = 2^{-p}\mathbf{H}_p\mathrm{diag}(\mathbf{H}_p\boldsymbol{y})\mathbf{H}_p \tag{G.1}$$

the *Hadamard prism* of $\boldsymbol{y}$. Its (Λ, Γ) entry is $y_{\Lambda\oplus\Gamma}$. Let $\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{X}^{\otimes}]$, and let $\boldsymbol{p}$ collect the cell probabilities in the column order of $\mathbf{H}_p$. Then $\boldsymbol{\mu} = \mathbf{H}_p\boldsymbol{p}$ and

$$\mathrm{Cov}(\boldsymbol{X}^{\otimes}) = \eta_p(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^{\top} = \mathbf{H}_p\mathrm{diag}(\boldsymbol{p})\mathbf{H}_p - \boldsymbol{\mu}\boldsymbol{\mu}^{\top}. \tag{G.2}$$

Thus the interaction means determine every second moment of the power vector, including under structural zeros. Removing the identity row and column gives the covariance in Theorem 3.1; selecting the relevant interactions gives the covariance blocks used in Theorem 3.2. Related constructions appear in the literature on group-circulant matrices and Boolean Fourier analysis (Terras, 1999; O'Donnell, 2014).

Because $2^{-p/2}\mathbf{H}_p$ is symmetric and orthogonal, the eigenvalues of $\eta_p(\boldsymbol{y})$ are exactly the coordinates of $\mathbf{H}_p\boldsymbol{y}$. When $\boldsymbol{y} = \mathbf{E}[\boldsymbol{Y}^{\otimes}]$ for a random $\boldsymbol{Y} \in \{\pm 1\}^p$, these coordinates are 2^p times the cell probabilities (Lemma G.2(c)); see also Zhang (2019). Below, the positions $0, \ldots, 2^p - 1$ of vectors and matrices are identified with $\{0,1\}^p$ through their binary expansions, so that $(\mathbf{H}_p)_{\Lambda,\Theta} = (-1)^{\Lambda^{\top}\Theta}$ for $\Lambda, \Theta \in \{0,1\}^p$. The next two lemmas collect identities for this interaction algebra and establish (G.2).

Lemma G.1. *For $\boldsymbol{y}, \boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^{2^p}$, the Hadamard prism satisfies:*

(a) Linearity and symmetry. *For any constants $a_1, a_2 \in \mathbb{R}$,*

$$\eta_p(a_1\boldsymbol{y}_1 + a_2\boldsymbol{y}_2) = a_1\eta_p(\boldsymbol{y}_1) + a_2\eta_p(\boldsymbol{y}_2), \qquad \eta_p(\boldsymbol{y})^{\top} = \eta_p(\boldsymbol{y}). \tag{G.3}$$

(b) Commutativity.

$$\eta_p(\boldsymbol{y}_1)\eta_p(\boldsymbol{y}_2) = \eta_p(\boldsymbol{y}_2)\eta_p(\boldsymbol{y}_1). \tag{G.4}$$

(c) Recursion.

$$\eta_{p+1}\left(\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix}\right) = \begin{pmatrix}\eta_p(\boldsymbol{y}_1) & \eta_p(\boldsymbol{y}_2)\\ \eta_p(\boldsymbol{y}_2) & \eta_p(\boldsymbol{y}_1)\end{pmatrix}. \tag{G.5}$$

(d) Exchange identity.

$$\eta_p(\boldsymbol{y}_1)\boldsymbol{y}_2 = \eta_p(\boldsymbol{y}_2)\boldsymbol{y}_1. \tag{G.6}$$

(e) Entries. *For all $\Lambda, \Gamma \in \{0,1\}^p$,*

$$\eta_p(\boldsymbol{y})_{\Lambda,\Gamma} = y_{\Lambda\oplus\Gamma}. \tag{G.7}$$

Thus $\eta_p(\boldsymbol{y})$ is the group-circulant matrix of $\boldsymbol{y}$ on $(\mathbb{Z}_2)^p$, and $\eta_p(\boldsymbol{y}_1)\boldsymbol{y}_2$ is the dyadic convolution $\left(\sum_{\Gamma}(y_1)_{\Lambda\oplus\Gamma}(y_2)_{\Gamma}\right)_{\Lambda}$ of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$.

Lemma G.2. *Let $\boldsymbol{Y} \in \{\pm 1\}^p$ and let $\boldsymbol{Y}^{\otimes}$ be its power vector. Then:*

(a) Rank-one representation.

$$\eta_p(\boldsymbol{Y}^{\otimes}) = \boldsymbol{Y}^{\otimes}(\boldsymbol{Y}^{\otimes})^{\top}. \tag{G.8}$$

(b) Compatibility with multiplication. *For every deterministic vector* $\boldsymbol{y} \in \mathbb{R}^{2^p}$,

$$(\boldsymbol{y}^\top \boldsymbol{Y}^{\otimes})\, \eta_p(\boldsymbol{Y}^{\otimes}) = \eta_p(\boldsymbol{Y}^{\otimes})\, \eta_p(\boldsymbol{y}) = \eta_p(\boldsymbol{y})\, \eta_p(\boldsymbol{Y}^{\otimes}). \tag{G.9}$$

(c) Second moments. *Suppose* $\boldsymbol{Y}$ *is random, let* $\boldsymbol{p} \in \mathbb{R}^{2^p}$ *be its pmf ordered to match the columns of* $\mathbf{H}_p$*, and let* $\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{Y}^{\otimes}]$*. Then* $\boldsymbol{\mu} = \mathbf{H}_p \boldsymbol{p}$ *and*

$$\mathbf{E}\big[\boldsymbol{Y}^{\otimes}(\boldsymbol{Y}^{\otimes})^\top\big] = \eta_p(\boldsymbol{\mu}) = \mathbf{H}_p \mathrm{diag}(\boldsymbol{p}) \mathbf{H}_p, \tag{G.10}$$

whose eigenvalues are $2^p\, \mathbf{P}(\boldsymbol{Y} = \boldsymbol{y})$*,* $\boldsymbol{y} \in \{\pm 1\}^p$*. Consequently,* $\mathrm{Cov}(\boldsymbol{Y}^{\otimes}) = \eta_p(\boldsymbol{\mu}) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$.

G.1 Proof of Lemma G.1

Part (a) follows directly from the definition

$$\eta_p(\boldsymbol{y}) = \frac{1}{2^p} \mathbf{H}_p\, \mathrm{diag}(\mathbf{H}_p \boldsymbol{y})\, \mathbf{H}_p,$$

because $\boldsymbol{y} \mapsto \mathbf{H}_p \boldsymbol{y}$ and $\boldsymbol{y} \mapsto \mathrm{diag}(\boldsymbol{y})$ are linear, and the middle matrix is diagonal.

For part (b), use $\mathbf{H}_p^2 = 2^p \mathbf{I}$ to write

$$\eta_p(\boldsymbol{y}_1)\eta_p(\boldsymbol{y}_2) = \frac{1}{2^p} \mathbf{H}_p\, \mathrm{diag}(\mathbf{H}_p \boldsymbol{y}_1) \mathrm{diag}(\mathbf{H}_p \boldsymbol{y}_2)\, \mathbf{H}_p.$$

The same expression with $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ interchanged is identical because diagonal matrices commute.

For part (c), let

$$\mathbf{D}_+ := \mathrm{diag}\big(\mathbf{H}_p(\boldsymbol{y}_1 + \boldsymbol{y}_2)\big), \qquad \mathbf{D}_- := \mathrm{diag}\big(\mathbf{H}_p(\boldsymbol{y}_1 - \boldsymbol{y}_2)\big).$$

Using the Sylvester recursion for the Hadamard matrix,

$$\mathbf{H}_{p+1} = \begin{pmatrix} \mathbf{H}_p & \mathbf{H}_p \\ \mathbf{H}_p & -\mathbf{H}_p \end{pmatrix},$$

we have

$$\mathbf{H}_{p+1} \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{H}_p(\boldsymbol{y}_1 + \boldsymbol{y}_2) \\ \mathbf{H}_p(\boldsymbol{y}_1 - \boldsymbol{y}_2) \end{pmatrix}.$$

Therefore

$$\eta_{p+1}\left(\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \right) = \frac{1}{2^{p+1}} \begin{pmatrix} \mathbf{H}_p & \mathbf{H}_p \\ \mathbf{H}_p & -\mathbf{H}_p \end{pmatrix} \begin{pmatrix} \mathbf{D}_+ & 0 \\ 0 & \mathbf{D}_- \end{pmatrix} \begin{pmatrix} \mathbf{H}_p & \mathbf{H}_p \\ \mathbf{H}_p & -\mathbf{H}_p \end{pmatrix}.$$

Now $\mathbf{D}_+ + \mathbf{D}_- = 2\mathrm{diag}(\mathbf{H}_p \boldsymbol{y}_1)$ and $\mathbf{D}_+ - \mathbf{D}_- = 2\mathrm{diag}(\mathbf{H}_p \boldsymbol{y}_2)$, so the block multiplication yields

$$\eta_{p+1}\left(\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \right) = \begin{pmatrix} \eta_p(\boldsymbol{y}_1) & \eta_p(\boldsymbol{y}_2) \\ \eta_p(\boldsymbol{y}_2) & \eta_p(\boldsymbol{y}_1) \end{pmatrix}.$$

For part (d), note that

$$\eta_p(\boldsymbol{y}_1)\boldsymbol{y}_2 = \frac{1}{2^p} \mathbf{H}_p\big((\mathbf{H}_p \boldsymbol{y}_1) \odot (\mathbf{H}_p \boldsymbol{y}_2)\big),$$

where $\odot$ denotes entrywise Hadamard multiplication. The right-hand side is symmetric in $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, hence

$$\eta_p(\boldsymbol{y}_1)\boldsymbol{y}_2 = \eta_p(\boldsymbol{y}_2)\boldsymbol{y}_1.$$

For part (e), use $(\mathbf{H}_p)_{\Lambda,\Theta} = (-1)^{\Lambda^\top \Theta}$ and $(\mathbf{H}_p \boldsymbol{y})_\Theta = \sum_\Xi (-1)^{\Theta^\top \Xi} y_\Xi$. Since $(-1)^{\Lambda^\top \Theta + \Theta^\top \Xi + \Theta^\top \Gamma} = (-1)^{\Theta^\top (\Lambda \oplus \Xi \oplus \Gamma)}$ and $\sum_{\Theta \in \{0,1\}^p} (-1)^{\Theta^\top \Phi} = 2^p\, \mathbf{1}\{\Phi = \mathbf{0}\}$,

$$\eta_p(\boldsymbol{y})_{\Lambda,\Gamma} = \sum_\Xi y_\Xi \frac{1}{2^p} \sum_\Theta (-1)^{\Theta^\top (\Lambda \oplus \Xi \oplus \Gamma)} = y_{\Lambda \oplus \Gamma}.$$

The dyadic-convolution form of $\eta_p(\boldsymbol{y}_1)\boldsymbol{y}_2$ follows by multiplying (G.7) by $\boldsymbol{y}_2$. □

G.2 Proof of Lemma G.2

For part (a), fix a realization $\boldsymbol{y} \in \{\pm 1\}^p$. The columns of $\mathbf{H}_p$ are exactly the vectors $\boldsymbol{y}^{\otimes}$ as $\boldsymbol{y}$ ranges over $\{\pm 1\}^p$. Hence there exists a standard basis vector $\boldsymbol{e}(\boldsymbol{y})$ such that

$$\boldsymbol{y}^{\otimes} = \mathbf{H}_p \boldsymbol{e}(\boldsymbol{y}), \qquad \mathbf{H}_p \boldsymbol{y}^{\otimes} = 2^p \boldsymbol{e}(\boldsymbol{y}).$$

Using the definition of the Hadamard prism, we obtain

$$\eta_p(\boldsymbol{y}^{\otimes}) = \frac{1}{2^p} \mathbf{H}_p \operatorname{diag}\big(2^p \boldsymbol{e}(\boldsymbol{y})\big)\, \mathbf{H}_p = \mathbf{H}_p \boldsymbol{e}(\boldsymbol{y}) \boldsymbol{e}(\boldsymbol{y})^{\top} \mathbf{H}_p = \boldsymbol{y}^{\otimes} (\boldsymbol{y}^{\otimes})^{\top}.$$

Since this identity holds for every realization $\boldsymbol{y}$, it proves part (a).

For part (b), the exchange identity and part (a) give

$$\eta_p(\boldsymbol{y}) \boldsymbol{Y}^{\otimes} = \eta_p(\boldsymbol{Y}^{\otimes}) \boldsymbol{y} = (\boldsymbol{y}^{\top} \boldsymbol{Y}^{\otimes}) \boldsymbol{Y}^{\otimes}.$$

Since both matrices are symmetric, it follows that

$$\eta_p(\boldsymbol{Y}^{\otimes}) \eta_p(\boldsymbol{y}) = \boldsymbol{Y}^{\otimes} \big(\eta_p(\boldsymbol{y}) \boldsymbol{Y}^{\otimes}\big)^{\top} = (\boldsymbol{y}^{\top} \boldsymbol{Y}^{\otimes}) \eta_p(\boldsymbol{Y}^{\otimes}).$$

The two prism matrices commute by Lemma G.1(b), proving all equalities in part (b).

For part (c), the columns of $\mathbf{H}_p$ are the vectors $\boldsymbol{y}^{\otimes}$, so $\boldsymbol{\mu} = \sum_{\boldsymbol{y}} \mathbf{P}(\boldsymbol{Y} = \boldsymbol{y})\, \boldsymbol{y}^{\otimes} = \mathbf{H}_p \boldsymbol{p}$. By part (a) and the linearity in Lemma G.1(a),

$$\mathbf{E}\big[\boldsymbol{Y}^{\otimes} (\boldsymbol{Y}^{\otimes})^{\top}\big] = \mathbf{E}\big[\eta_p(\boldsymbol{Y}^{\otimes})\big] = \eta_p(\boldsymbol{\mu}).$$

Since $\mathbf{H}_p \boldsymbol{\mu} = \mathbf{H}_p^2 \boldsymbol{p} = 2^p \boldsymbol{p}$, the definition gives $\eta_p(\boldsymbol{\mu}) = \mathbf{H}_p \operatorname{diag}(\boldsymbol{p}) \mathbf{H}_p$, and its eigenvalues, the coordinates of $\mathbf{H}_p \boldsymbol{\mu}$, are $2^p \boldsymbol{p}$. Subtracting $\boldsymbol{\mu}\boldsymbol{\mu}^{\top}$ gives the covariance. □

H Binary approximation and its proof

We first specify the binary expansion at dyadic endpoints. We then give the approximation result for conditional event probabilities and its connection to the BEGIN block. These results prove the rate in Theorem 3.7(b); a separate argument establishes the stronger converse in part (a).

H.1 Dyadic partitions and conditional event probabilities

For $x \in [-1, 1]$ and an integer $d \geq 1$, define the cell index and midpoint by

$$k_d(x) = \min\big\{2^d - 1, \big\lfloor 2^{d-1}(x+1) \big\rfloor\big\}, \qquad T_d(x) = -1 + 2^{-d} + 2^{1-d} k_d(x).$$

The minimum assigns $x = 1$ to the last cell. Write $k_d(x) = \sum_{k=1}^{d} \epsilon_k(x) 2^{d-k}$ with $\epsilon_k(x) \in \{0, 1\}$, and set $A_k(x) = 2\epsilon_k(x) - 1$. These digits do not depend on $d \geq k$, and

$$T_d(x) = \sum_{k=1}^{d} A_k(x) 2^{-k}, \qquad |T_d(x) - x| \leq 2^{-d}.$$

In particular, $x = \sum_{k \geq 1} A_k(x) 2^{-k}$. We apply T_d coordinatewise and write

$$\mathcal{U}_d = \sigma\{T_d(\boldsymbol{U})\}, \qquad \mathcal{V}_d = \sigma\{T_d(\boldsymbol{V})\}, \qquad \mathcal{W}_d = \sigma\{T_d(\boldsymbol{W})\}.$$

These finite sigma-fields increase with d and satisfy

$$\sigma\Big(\bigcup_{d \geq 1} \mathcal{U}_d\Big) = \sigma(\boldsymbol{U}), \qquad \sigma\Big(\bigcup_{d \geq 1} \mathcal{V}_d\Big) = \sigma(\boldsymbol{V}), \qquad \sigma\Big(\bigcup_{d \geq 1} \mathcal{W}_d\Big) = \sigma(\boldsymbol{W}). \tag{H.1}$$

Indeed, the quantized vectors are measurable functions of the original vectors and converge to them pointwise.

To measure conditional dependence at resolution d, put

$$\Delta_d = \sup_{E \in \mathcal{U}_d,\, F \in \mathcal{W}_d} \mathbf{E}\big|\mathbf{P}(E \cap F \mid \mathcal{V}_d) - \mathbf{P}(E \mid \mathcal{V}_d)\mathbf{P}(F \mid \mathcal{V}_d)\big|.$$

Because the sigma-fields are finite, $\Delta_d = 0$ if and only if $\mathcal{U}_d \perp\!\!\!\perp \mathcal{W}_d \mid \mathcal{V}_d$.

Theorem H.1 (Dyadic conditional independence)**.** *Let* $(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}) \in [-1,1]^{r_U} \times [-1,1]^{r_V} \times [-1,1]^{r_W}$.

(a) If $\liminf_{d\to\infty} \Delta_d = 0$*, then* $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$*. In particular, this conclusion holds if* $\mathcal{U}_d \perp\!\!\!\perp \mathcal{W}_d \mid \mathcal{V}_d$ *for infinitely many* d*.*

(b) Under the continuity conditions of Theorem 3.7(b), if $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$*, then, for every* $d \geq 1$*,*

$$\Delta_d \leq L_{\boldsymbol{U}} L_{\boldsymbol{W}} r_V^{\kappa} 2^{2\kappa(1-d)-2}.$$

Under these continuity conditions, $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$ *if and only if* $\Delta_d \to 0$*, and in that case* $\Delta_d = O(4^{-\kappa d})$.

Let $\boldsymbol{A}^{(d)} \in \{\pm 1\}^{dr_U}$, $\boldsymbol{B}^{(d)} \in \{\pm 1\}^{dr_V}$, and $\boldsymbol{C}^{(d)} \in \{\pm 1\}^{dr_W}$ collect the first d sign digits of each coordinate of $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{W}$, respectively. Form $\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]$ as in Theorem 3.2. This is the cross block used in Theorem 3.7.

Corollary H.2 (BEGIN blocks of the leading bits)**.** *For every* $d \geq 1$*:*

(a) $\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$ *if and only if* $\mathcal{U}_d \perp\!\!\!\perp \mathcal{W}_d \mid \mathcal{V}_d$*. Hence, if this block is zero for infinitely many* d*, then* $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$*.*

(b) $\|\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]\|_{\max} \leq 4\Delta_d$*. Under conditional independence and the continuity conditions of Theorem 3.7(b), therefore,*

$$\|\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]\|_{\max} \leq L_{\boldsymbol{U}} L_{\boldsymbol{W}} r_V^{\kappa} 2^{2\kappa(1-d)}.$$

H.2 Proof of Theorem H.1

Part (a). We first show that, for every d and all $\mathcal{U}_d$-measurable f and $\mathcal{W}_d$-measurable g with values in $[0,1]$,

$$\mathbf{E}\big|\mathbf{E}[fg \mid \mathcal{V}_d] - \mathbf{E}[f \mid \mathcal{V}_d]\,\mathbf{E}[g \mid \mathcal{V}_d]\big| \leq \Delta_d. \tag{H.2}$$

Since $\mathcal{U}_d$ and $\mathcal{W}_d$ are finite sigma-fields, f and g take finitely many values, say $0 \leq a_1 < \cdots < a_m \leq 1$ and $0 \leq b_1 < \cdots < b_l \leq 1$. With $a_0 = b_0 = 0$,

$$f = \sum_{i=1}^{m} (a_i - a_{i-1}) \mathbf{1}\{f \geq a_i\}, \qquad g = \sum_{k=1}^{l} (b_k - b_{k-1}) \mathbf{1}\{g \geq b_k\},$$

where the events $\{f \geq a_i\}$ belong to $\mathcal{U}_d$, the events $\{g \geq b_k\}$ belong to $\mathcal{W}_d$, and the nonnegative weights in each sum total at most one. The conditional covariance given $\mathcal{V}_d$ is bilinear, so the left side of (H.2) is at most $\sum_{i,k} (a_i - a_{i-1})(b_k - b_{k-1})\Delta_d \leq \Delta_d$.

Now let $d_1 < d_2 < \cdots$ satisfy $\Delta_{d_n} \to 0$. This is possible because $\liminf_d \Delta_d = 0$, and in particular when $\Delta_d = 0$ for infinitely many d. Fix arbitrary events $E \in \sigma(\boldsymbol{U})$ and $F \in \sigma(\boldsymbol{W})$. Define their conditional-probability approximations by

$$f_d := \mathbf{E}[1_E \mid \mathcal{U}_d], \qquad g_d := \mathbf{E}[1_F \mid \mathcal{W}_d].$$

The martingale convergence theorem for increasing sigma-fields gives $f_d \to 1_E$ and $g_d \to 1_F$ in L^1. All these variables take values in $[0,1]$, so

$$|f_d g_d - 1_E 1_F| \leq |f_d - 1_E| + |g_d - 1_F|.$$

In particular, $f_d g_d \to 1_E 1_F$ in L^1. Conditional expectation contracts L^1 distances. Together with martingale convergence along $\mathcal{V}_d$, this gives

$$\mathbf{E}[f_d \mid \mathcal{V}_d] \longrightarrow \mathbf{E}[1_E \mid \sigma(\boldsymbol{V})], \qquad \mathbf{E}[g_d \mid \mathcal{V}_d] \longrightarrow \mathbf{E}[1_F \mid \sigma(\boldsymbol{V})] \quad \text{in } L^1,$$

and

$$\mathbf{E}[f_d g_d \mid \mathcal{V}_d] \longrightarrow \mathbf{E}[1_E 1_F \mid \sigma(\boldsymbol{V})] \quad \text{in } L^1.$$

For example, the first convergence follows from

$$\begin{aligned}\mathbf{E}\big|\mathbf{E}[f_d \mid \mathcal{V}_d] - \mathbf{E}[1_E \mid \sigma(\boldsymbol{V})]\big| &\leq \mathbf{E}|f_d - 1_E| \\ &\quad + \mathbf{E}\big|\mathbf{E}[1_E \mid \mathcal{V}_d] - \mathbf{E}[1_E \mid \sigma(\boldsymbol{V})]\big| \longrightarrow 0.\end{aligned}$$

Since the conditional means of f_d and g_d lie in $[0,1]$, their product also converges in L^1 to the product of the limits. By the triangle inequality and (H.2),

$$\begin{aligned}\mathbf{E}\big|\mathbf{E}[1_E 1_F \mid \sigma(\boldsymbol{V})] - \mathbf{E}[1_E \mid \sigma(\boldsymbol{V})]\,\mathbf{E}[1_F \mid \sigma(\boldsymbol{V})]\big| \\ \le \mathbf{E}\big|\mathbf{E}[1_E 1_F \mid \sigma(\boldsymbol{V})] - \mathbf{E}[f_d g_d \mid \mathcal{V}_d]\big| + \Delta_d \\ + \mathbf{E}\big|\mathbf{E}[f_d \mid \mathcal{V}_d]\,\mathbf{E}[g_d \mid \mathcal{V}_d] - \mathbf{E}[1_E \mid \sigma(\boldsymbol{V})]\,\mathbf{E}[1_F \mid \sigma(\boldsymbol{V})]\big|.\end{aligned}$$

Along $d = d_n$, all three terms on the right tend to zero. Hence

$$\mathbf{E}[1_E 1_F \mid \sigma(\boldsymbol{V})] = \mathbf{E}[1_E \mid \sigma(\boldsymbol{V})]\mathbf{E}[1_F \mid \sigma(\boldsymbol{V})] \quad \text{a.s.}$$

As E and F were arbitrary, this proves $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$.

Part (b). Fix a depth $d \ge 1$ and events $E \in \mathcal{U}_d$ and $F \in \mathcal{W}_d$. There are Borel sets $D_U \subseteq [-1,1]^{r_U}$ and $D_W \subseteq [-1,1]^{r_W}$ such that $E = \{\boldsymbol{U} \in D_U\}$ and $F = \{\boldsymbol{W} \in D_W\}$. Use the versions of the conditional laws satisfying the assumed total-variation bounds to define

$$\phi_E(\boldsymbol{v}) := \mathbf{P}(\boldsymbol{U} \in D_U \mid \boldsymbol{V} = \boldsymbol{v}), \qquad \psi_F(\boldsymbol{v}) := \mathbf{P}(\boldsymbol{W} \in D_W \mid \boldsymbol{V} = \boldsymbol{v}).$$

Thus, for all $\boldsymbol{v}, \boldsymbol{v}' \in [-1,1]^{r_V}$,

$$|\phi_E(\boldsymbol{v}) - \phi_E(\boldsymbol{v}')| \le L_{\boldsymbol{U}} \|\boldsymbol{v} - \boldsymbol{v}'\|_2^{\kappa}, \qquad |\psi_F(\boldsymbol{v}) - \psi_F(\boldsymbol{v}')| \le L_{\boldsymbol{W}} \|\boldsymbol{v} - \boldsymbol{v}'\|_2^{\kappa}.$$

Since $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$, the conditional joint probability of E and F is $\phi_E(\boldsymbol{V})\psi_F(\boldsymbol{V})$ almost surely. The tower property therefore gives

$$\begin{aligned}&\mathbf{P}(E \cap F \mid \mathcal{V}_d) - \mathbf{P}(E \mid \mathcal{V}_d)\mathbf{P}(F \mid \mathcal{V}_d) \\ &\quad = \mathbf{E}[\phi_E(\boldsymbol{V})\psi_F(\boldsymbol{V}) \mid \mathcal{V}_d] - \mathbf{E}[\phi_E(\boldsymbol{V}) \mid \mathcal{V}_d]\mathbf{E}[\psi_F(\boldsymbol{V}) \mid \mathcal{V}_d] \\ &\quad = \mathrm{Cov}\big(\phi_E(\boldsymbol{V}), \psi_F(\boldsymbol{V}) \mid \mathcal{V}_d\big). \qquad \text{(H.3)}\end{aligned}$$

Let $K \subseteq [-1,1]^{r_V}$ be a cell of the depth-d dyadic partition with $\mathbf{P}(\boldsymbol{V} \in K) > 0$. Its Euclidean diameter is at most $\sqrt{r_V}\,2^{1-d}$. Hence the range of ϕ_E on K has length at most $L_{\boldsymbol{U}} r_V^{\kappa/2} 2^{\kappa(1-d)}$, and the range of ψ_F on K has length at most $L_{\boldsymbol{W}} r_V^{\kappa/2} 2^{\kappa(1-d)}$. A real random variable whose range lies in an interval of length a has variance at most $a^2/4$: subtracting the midpoint bounds its squared deviation by $a^2/4$, and the mean minimizes expected squared deviation. It follows that, on $\{\boldsymbol{V} \in K\}$,

$$\mathrm{Var}\big(\phi_E(\boldsymbol{V}) \mid \mathcal{V}_d\big) \le \frac{L_{\boldsymbol{U}}^2}{4} r_V^{\kappa} 2^{2\kappa(1-d)},$$

$$\mathrm{Var}\big(\psi_F(\boldsymbol{V}) \mid \mathcal{V}_d\big) \le \frac{L_{\boldsymbol{W}}^2}{4} r_V^{\kappa} 2^{2\kappa(1-d)}.$$

Conditional Cauchy–Schwarz applied to (H.3) therefore yields, almost surely,

$$\big|\mathbf{P}(E \cap F \mid \mathcal{V}_d) - \mathbf{P}(E \mid \mathcal{V}_d)\mathbf{P}(F \mid \mathcal{V}_d)\big| \le L_{\boldsymbol{U}} L_{\boldsymbol{W}} r_V^{\kappa} 2^{2\kappa(1-d)-2}.$$

The cells of probability zero may be omitted, and there are only finitely many cells. Taking expectations and then the supremum over $E \in \mathcal{U}_d$ and $F \in \mathcal{W}_d$ proves

$$\Delta_d \le L_{\boldsymbol{U}} L_{\boldsymbol{W}} r_V^{\kappa} 2^{2\kappa(1-d)-2},$$

which tends to zero because $\kappa > 0$. The final statement of the theorem combines this bound with part (a), which needs no continuity. □

H.3 Proof of Corollary H.2

By the digit representation of T_d, $\sigma(\boldsymbol{A}^{(d)}) = \mathcal{U}_d$, $\sigma(\boldsymbol{B}^{(d)}) = \mathcal{V}_d$, and $\sigma(\boldsymbol{C}^{(d)}) = \mathcal{W}_d$. Hence $\mathcal{U}_d \perp\!\!\!\perp \mathcal{W}_d \mid \mathcal{V}_d$ if and only if $\boldsymbol{A}^{(d)} \perp\!\!\!\perp \boldsymbol{C}^{(d)} \mid \boldsymbol{B}^{(d)}$, which by Theorem 3.2 is equivalent to $\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$. The second statement in (a) then follows from Theorem H.1(a).

For (b), Proposition 3.3 shows that every entry of $\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]$ has the form $\mathbf{E}[B_\Theta\, \mathfrak{c}(\boldsymbol{B}^{(d)})]$, where B_Θ is an interaction of $\boldsymbol{B}^{(d)}$ and $\mathfrak{c}(\boldsymbol{B}^{(d)}) = \mathrm{Cov}(A_{\Lambda_1}, C_{\Lambda_2} \mid \boldsymbol{B}^{(d)})$ for interactions A_{Λ_1} of $\boldsymbol{A}^{(d)}$ and C_{Λ_2} of $\boldsymbol{C}^{(d)}$. Its absolute value is therefore at most $\mathbf{E}|\mathfrak{c}(\boldsymbol{B}^{(d)})|$. Write $A_{\Lambda_1} = 2 \cdot \mathbf{1}_E - 1$ and $C_{\Lambda_2} = 2 \cdot \mathbf{1}_F - 1$ with $E = \{A_{\Lambda_1} = 1\} \in \mathcal{U}_d$ and $F = \{C_{\Lambda_2} = 1\} \in \mathcal{W}_d$. By bilinearity, $\mathfrak{c}(\boldsymbol{B}^{(d)}) = 4\{\mathbf{P}(E \cap F \mid \mathcal{V}_d) - \mathbf{P}(E \mid \mathcal{V}_d)\mathbf{P}(F \mid \mathcal{V}_d)\}$, so $\mathbf{E}|\mathfrak{c}(\boldsymbol{B}^{(d)})| \le 4\Delta_d$. The final bound follows from Theorem H.1(b). □

H.4 Proof of Theorem 3.7

Part (a). Choose $d_n \to \infty$ along which $\|\mathbf{S}^{(d_n)}[\mathcal{L}, \mathcal{R}]\|_{\max} \to 0$. Fix an interaction f of finitely many leading bits of $\boldsymbol{U}$, an interaction g of finitely many leading bits of $\boldsymbol{W}$, and an interaction h of finitely many leading bits of $\boldsymbol{V}$, allowing $h = 1$. When either endpoint label is the identity, the conditional covariance of f and g is already zero, so assume both endpoint labels are nonidentity. For all sufficiently large d, the row labeled fh and the column labeled g belong to the cross block. Proposition 3.3 gives

$$\mathbf{S}^{(d)}[fh, g] = \mathbf{E}\left[h\left\{\mathbf{E}[fg \mid \mathcal{V}_d] - \mathbf{E}[f \mid \mathcal{V}_d]\mathbf{E}[g \mid \mathcal{V}_d]\right\}\right].$$

The three interactions are bounded. Martingale convergence and (H.1) imply

$$\mathbf{E}[fg \mid \mathcal{V}_d] - \mathbf{E}[f \mid \mathcal{V}_d]\mathbf{E}[g \mid \mathcal{V}_d] \longrightarrow Z := \mathbf{E}[fg \mid \boldsymbol{V}] - \mathbf{E}[f \mid \boldsymbol{V}]\mathbf{E}[g \mid \boldsymbol{V}] \quad \text{in } L^1.$$

For the product, use $|ab - a'b'| \le |a - a'| + |b - b'|$ when all four factors belong to $[-1, 1]$. Along $d = d_n$, the selected matrix entry tends to zero, while its expectation representation tends to $\mathbf{E}[hZ]$. Thus $\mathbf{E}[hZ] = 0$ for every finite interaction h of the conditioning bits.

For each fixed m, the interactions of $\boldsymbol{B}^{(m)}$, including the constant, span all functions measurable with respect to $\mathcal{V}_m$, even on a restricted support. It follows that $\mathbf{E}[Z \mid \mathcal{V}_m] = 0$. Since Z is bounded and $\sigma(\boldsymbol{V})$-measurable, another application of martingale convergence gives $Z = 0$ almost surely.

We have shown zero conditional covariance given $\boldsymbol{V}$ for every pair of finite endpoint interactions. By linearity, the same is true for every bounded $\mathcal{U}_m$-measurable function and every bounded $\mathcal{W}_m$-measurable function. Now fix arbitrary events $E \in \sigma(\boldsymbol{U})$ and $F \in \sigma(\boldsymbol{W})$, and set $f_m = \mathbf{E}[1_E \mid \mathcal{U}_m]$ and $g_m = \mathbf{E}[1_F \mid \mathcal{W}_m]$. These functions lie in $[0, 1]$ and converge in L^1 to 1_E and 1_F. Their zero conditional covariance says

$$\mathbf{E}[f_m g_m \mid \boldsymbol{V}] = \mathbf{E}[f_m \mid \boldsymbol{V}]\mathbf{E}[g_m \mid \boldsymbol{V}].$$

Boundedness and the L^1 contraction of conditional expectation allow passage to the limit, giving

$$\mathbf{P}(E \cap F \mid \boldsymbol{V}) = \mathbf{P}(E \mid \boldsymbol{V})\mathbf{P}(F \mid \boldsymbol{V}) \quad \text{almost surely.}$$

This proves $\boldsymbol{U} \perp\!\!\!\perp \boldsymbol{W} \mid \boldsymbol{V}$ without continuity or positivity assumptions.

Part (b). The bound follows from Theorem H.1(b) and Corollary H.2(b). It tends to zero because $\kappa > 0$. Conversely, convergence of the block norm to zero implies conditional independence by part (a). This proves the final equivalence. □

H.5 Why continuity is needed

Population conditional independence alone need not make the dyadic blocks small. Let V be uniform on $[-1, 1]$, define

$$U = \begin{cases} 2V + 1, & V < 0, \\ 2V - 1, & V \ge 0, \end{cases} \qquad W = U.$$

Both endpoints are functions of V, so $U \perp\!\!\!\perp W \mid V$. Apart from a null set of dyadic endpoints, the kth sign digit of U and W is the $(k + 1)$st sign digit of V. The sign digits of V are independent and fair. Thus the dth endpoint digit is independent of $\mathcal{V}_d$, and the two endpoint digits agree. The corresponding BEGIN entry is

$$\mathbf{S}^{(d)}[A_d(U), A_d(W)] = \mathbf{E}\{\mathrm{Cov}(A_d(U), A_d(W) \mid \mathcal{V}_d)\} = 1.$$

Every cross-block entry has absolute value at most one by Proposition 3.3. Hence $\|\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]\|_{\max} = 1$ for every d. Also $\Delta_d = 1/4$: the event that the dth endpoint digit is positive attains this value, and the conditional covariance of two indicators is at most $1/4$ in absolute value. The conditional laws in this example are point masses and fail the total-variation continuity assumption. Without a regularity assumption, population conditional independence therefore need not imply vanishing dyadic blocks. The converse in Theorem 3.7(a) requires no such assumption.

H.6 SHARPNESS FOR EXPONENT ONE

Let $r_V = 1$, let V be uniform on $[-1,1]$, and let $U, W \in \{\pm 1\} \subset [-1,1]$ be conditionally independent given V with $\mathbf{P}(U = 1 \mid V) = \mathbf{P}(W = 1 \mid V) = (1+V)/2$. Then $\mathrm{TV}\big(\mathcal{P}(U \mid V = v), \mathcal{P}(U \mid V = v')\big) = |v - v'|/2$, and similarly for W, so Theorem 3.7(b) applies with $\kappa = 1$ and $L_U = L_W = 1/2$. Here $\mathcal{U}_d = \sigma(U)$ and $\mathcal{W}_d = \sigma(W)$ for every d. Given $\mathcal{V}_d$, V is uniform on a cell of length 2^{1-d}, so by (H.3),

$$\mathbf{P}(U = 1, W = 1 \mid \mathcal{V}_d) - \mathbf{P}(U = 1 \mid \mathcal{V}_d)\mathbf{P}(W = 1 \mid \mathcal{V}_d) = \tfrac{1}{4}\,\mathrm{Var}(V \mid \mathcal{V}_d) = \tfrac{1}{48}\,4^{1-d}.$$

The other choices of events give the same absolute value or zero, so $\Delta_d = 4^{-d}/12$, one third of the bound $4^{-d}/4$ in Theorem H.1(b). Every bit of U equals U, every bit of W equals W, and the bits of V are independent fair signs. Thus every left feature is almost surely either an interaction of $\boldsymbol{B}^{(d)}$, whose residual vanishes, or U times such an interaction, and similarly on the right with W. Since $\mathrm{Cov}(U, W \mid \mathcal{V}_d) = 4 \cdot 4^{1-d}/48 = 4^{-d}/3$, Proposition 3.3 shows that the entries of $\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]$ are $\mathbf{E}[B_\Theta]\,4^{-d}/3$ or zero, and $\mathbf{E}[B_\Theta] = \mathbf{1}\{\Theta = \mathbf{0}\}$. Hence

$$\|\mathbf{S}^{(d)}[\mathcal{L}, \mathcal{R}]\|_{\max} = 4\Delta_d = \frac{4^{-d}}{3},$$

whereas Corollary H.2(b) gives the bound 4^{-d}.

Thus the factor 4 relating the block norm to Δ_d is attained, and the rate 4^{-d} cannot be improved in general when $\kappa = 1$. The example does not attain the constant in the Hölder upper bound.

I FINITE-SAMPLE BOUNDS AND SUBGROUP SEARCHES

For scalar endpoints and a fixed finite conditioning vector, we obtain two-sided finite-sample bounds for the empirical block. They depend on the number and minimum probability of supported conditioning states. We first define residual scores for subgroup searches, then prove the bounds using an exact within-cell covariance formula that avoids numerically inverting a singular matrix. A union bound gives simultaneous control over any finite search fixed in advance.

I.1 RESIDUALS OVER A FINITE CONDITIONING FAMILY

Theorem 3.2 suggests measuring residual conditional dependence by the cross block of its Schur complement. Given n observations of p binary signs $\boldsymbol{X} = (X_1, \ldots, X_p)$, for each pair (i, j) we search over sets $M \subseteq \{1, \ldots, p\} \setminus \{i, j\}$ with $|M| \le r_{\max}$ and subgroups $\mathcal{K} \le \langle \boldsymbol{X}_M \rangle$. For each choice, set

$$\mathcal{B} = \mathcal{K} \setminus \{1\}, \qquad \mathcal{L} = \{X_i g : g \in \mathcal{K}\}, \qquad \mathcal{R} = \{X_j g : g \in \mathcal{K}\},$$

and form the empirical cross block $\widehat{\mathbf{S}}_{i,j,M,\mathcal{K}}$ in (3.2), with covariances normalized by $1/n$. We define

$$Q_{\mathrm{bit}}(i, j) = \min_{M,\mathcal{K}} \|\widehat{\mathbf{S}}_{i,j,M,\mathcal{K}}\|_{\max}.$$

A small score indicates that some searched adjustment leaves little residual association; a large score indicates that the association persists across all searched choices. When $r_{\max} = 3$, each comparison uses at most 23 features. Equation (I.3) below computes the block from within-cell covariances.

The conditioning information is determined by the subgroup: the population block vanishes exactly when X_i and X_j are conditionally independent given the interactions in $\mathcal{K}$ (Corollary I.2). A proper subgroup can contain less information than $\boldsymbol{X}_M$. The score is also on a covariance scale (Proposition 3.3), so dependence in rare states receives less weight, and scores can be small for bits with very unbalanced marginal distributions. When several bits encode one variable, let a, b be the corresponding sets of bit indices and define

$$Q_{\mathrm{blk}}(a, b) = \max_{i \in a,\, j \in b} Q_{\mathrm{bit}}(i, j), \qquad a \neq b, \tag{I.1}$$

with zero diagonal. This maximum is one possible aggregation of scalar scores. The full vector characterization retains all endpoint interactions; aggregating scalar comparisons can omit dependence carried by these interactions. Corollary I.2 below gives simultaneous bounds for the prespecified residual search.

I.2 FINITE-SAMPLE BOUNDS

We first give bounds for two scalar endpoint bits and a conditioning vector fixed before observing the data. We then obtain simultaneous bounds for the finite search defined in Section I.1. All empirical covariances in this appendix use the normalization $1/n$.

Theorem I.1 (Finite-sample bounds for BEGIN entries)**.** *Let $(A_i, \boldsymbol{B}_i, C_i)_{i=1}^n$ be IID observations, where $A, C \in \{\pm 1\}$ and $\boldsymbol{B} \in \{\pm 1\}^{r_B}$ is a conditioning vector chosen in advance. Form*

$$\mathcal{B} = \langle \boldsymbol{B} \rangle \setminus \{1\}, \qquad \mathcal{L} = \{Ag : g \in \langle \boldsymbol{B} \rangle\}, \qquad \mathcal{R} = \{Cg : g \in \langle \boldsymbol{B} \rangle\},$$

and let $\widehat{\mathbf{S}}$ be the $(\mathcal{L}, \mathcal{R})$ cross block of the empirical Schur complement in (3.2)*. Suppose $\boldsymbol{B}$ has h_0 supported states, each with probability at least $\pi_{\min} > 0$, and*

$$\max_{\boldsymbol{b} \in \mathrm{Supp}(\boldsymbol{B})} |\operatorname{Cov}(A, C \mid \boldsymbol{B} = \boldsymbol{b})| \le \beta.$$

For $0 < \delta < 1$, set

$$\epsilon_n = \sqrt{\frac{\log(12h_0/\delta)}{n\pi_{\min}}}.$$

If $n\pi_{\min} \ge 8\log(2h_0/\delta)$, then, with probability at least $1 - \delta$, every supported conditioning state is observed and

$$\frac{1}{2} \max_{\boldsymbol{b} \in \mathrm{Supp}(\boldsymbol{B})} \mathbf{P}(\boldsymbol{B} = \boldsymbol{b}) |\operatorname{Cov}(A, C \mid \boldsymbol{B} = \boldsymbol{b})| - 3\epsilon_n \;\le\; \|\widehat{\mathbf{S}}\|_{\max} \;\le\; \beta + 6\epsilon_n. \tag{I.2}$$

Under $A \perp\!\!\!\perp C \mid \boldsymbol{B}$, one may take $\beta = 0$.

The theorem allows structural zeros and depends on the smallest probability among the supported conditioning states. Under conditional independence, $6\epsilon_n$ is a valid upper threshold with error probability at most δ. If some supported state satisfies

$$\mathbf{P}(\boldsymbol{B} = \boldsymbol{b}) |\operatorname{Cov}(A, C \mid \boldsymbol{B} = \boldsymbol{b})| \ge \eta > 0,$$

then the lower bound exceeds this null threshold whenever $\eta > 18\epsilon_n$, or equivalently

$$n\pi_{\min} > \frac{324\log(12h_0/\delta)}{\eta^2},$$

in addition to the sample-size condition of the theorem. The weaker condition $\eta > 6\epsilon_n$ only makes the lower bound positive; it does not separate it from the null threshold. The threshold depends on population support parameters; estimating these parameters requires additional justification. These guarantees concern a fixed comparison and a threshold justified by the bound, rather than a threshold selected from the same data.

I.3 PROOF OF THEOREM I.1

For each $\boldsymbol{b} \in \mathrm{Supp}(\boldsymbol{B})$, write

$$p_{\boldsymbol{b}} = \mathbf{P}(\boldsymbol{B} = \boldsymbol{b}), \qquad N_{\boldsymbol{b}} = \sum_{i=1}^n \mathbf{1}_{\{\boldsymbol{B}_i = \boldsymbol{b}\}}, \qquad \widehat{p}_{\boldsymbol{b}} = N_{\boldsymbol{b}}/n.$$

For $T \in \{A, C, AC\}$, let $\mu_T(\boldsymbol{b}) = \mathbf{E}[T \mid \boldsymbol{B} = \boldsymbol{b}]$ and let $\widehat{\mu}_T(\boldsymbol{b})$ be the average of T in that cell, with value zero when $N_{\boldsymbol{b}} = 0$. Define the population and empirical conditional covariances by

$$\mathfrak{c}(\boldsymbol{b}) = \mu_{AC}(\boldsymbol{b}) - \mu_A(\boldsymbol{b})\mu_C(\boldsymbol{b}), \qquad \widehat{\mathfrak{c}}(\boldsymbol{b}) = \widehat{\mu}_{AC}(\boldsymbol{b}) - \widehat{\mu}_A(\boldsymbol{b})\widehat{\mu}_C(\boldsymbol{b}).$$

The empirical conditional-covariance identity. Fix $g, h \in \langle \boldsymbol{B} \rangle$. Put $Y_i = A_i g(\boldsymbol{B}_i)$ and $Z_i = C_i h(\boldsymbol{B}_i)$, with cell means

$$m_i = g(\boldsymbol{B}_i)\widehat{\mu}_A(\boldsymbol{B}_i), \qquad \ell_i = h(\boldsymbol{B}_i)\widehat{\mu}_C(\boldsymbol{B}_i).$$

Let $\bar{Y}, \bar{Z}$ be the overall sample means, and let $\boldsymbol{z}_i$ collect the centered separator interactions $k(\boldsymbol{B}_i) - n^{-1}\sum_l k(\boldsymbol{B}_l)$, $k \in \mathcal{B}$. Then $\widehat{\boldsymbol{\Sigma}}_{\mathcal{B}} = n^{-1}\sum_i \boldsymbol{z}_i \boldsymbol{z}_i^\top$. The constant and the separator interactions span every function on the observed conditioning states. Hence there are vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$ such that

$$m_i - \bar{Y} = \boldsymbol{\alpha}^\top \boldsymbol{z}_i, \qquad \ell_i - \bar{Z} = \boldsymbol{\beta}^\top \boldsymbol{z}_i.$$

Within each observed cell, the residuals $Y_i - m_i$ and $Z_i - \ell_i$ sum to zero. Since $\boldsymbol{z}_i$ is constant within a cell, the two empirical covariance blocks involving the separator are $\boldsymbol{\alpha}^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{B}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathcal{B}}\boldsymbol{\beta}$. Their Schur correction is

$$\boldsymbol{\alpha}^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{B}} \widehat{\boldsymbol{\Sigma}}_{\mathcal{B}}^{+} \widehat{\boldsymbol{\Sigma}}_{\mathcal{B}}\boldsymbol{\beta} = \boldsymbol{\alpha}^\top \widehat{\boldsymbol{\Sigma}}_{\mathcal{B}}\boldsymbol{\beta} = \frac{1}{n}\sum_i (m_i - \bar{Y})(\ell_i - \bar{Z}).$$

This follows from the Moore–Penrose identity and remains valid when $\widehat{\boldsymbol{\Sigma}}_{\mathcal{B}}$ is singular. Subtracting the correction, expanding the products, and again using the zero sums of the residuals within each cell gives

$$\begin{aligned}\widehat{\mathbf{S}}[Ag, Ch] &= \frac{1}{n}\sum_i (Y_i - m_i)(Z_i - \ell_i) \\ &= \sum_{\boldsymbol{b}\in \mathrm{Supp}(\boldsymbol{B})} \widehat{p}_{\boldsymbol{b}}\, g(\boldsymbol{b}) h(\boldsymbol{b})\, \widehat{\mathfrak{c}}(\boldsymbol{b}). \end{aligned} \tag{I.3}$$

Empty cells contribute zero. Each entry combines the cell covariances with signed probability weights.

Cell counts and conditional means. Consider the event

$$\mathcal{E} = \{N_{\boldsymbol{b}} \geq np_{\boldsymbol{b}}/2 \text{ for every } \boldsymbol{b} \in \mathrm{Supp}(\boldsymbol{B})\}.$$

The binomial Chernoff bound gives

$$\mathbf{P}(\mathcal{E}^c) \leq \sum_{\boldsymbol{b}\in \mathrm{Supp}(\boldsymbol{B})} e^{-np_{\boldsymbol{b}}/8} \leq h_0 e^{-n\pi_{\min}/8} \leq \delta/2.$$

On $\mathcal{E}$, every supported cell is nonempty, $N_{\boldsymbol{b}} \geq n\pi_{\min}/2$, and $\widehat{p}_{\boldsymbol{b}} \geq p_{\boldsymbol{b}}/2$.

Conditional on $(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_n)$, the observations in each cell are independent with their corresponding conditional law. Since $A, C, AC \in [-1, 1]$, Hoeffding inequality gives

$$\mathbf{P}\big(|\widehat{\mu}_T(\boldsymbol{b}) - \mu_T(\boldsymbol{b})| > t \mid \boldsymbol{B}_1, \ldots, \boldsymbol{B}_n\big) \leq 2e^{-N_{\boldsymbol{b}} t^2/2}.$$

Set $t = 2\epsilon_n$. On $\mathcal{E}$, the right side is at most $\delta/(6h_0)$. A conditional union bound over the $3h_0$ choices of T and $\boldsymbol{b}$, followed by the bound on $\mathbf{P}(\mathcal{E}^c)$, shows that with probability at least $1 - \delta$, $\mathcal{E}$ holds and every conditional-mean error is at most t. On this event,

$$\begin{aligned}|\widehat{\mathfrak{c}}(\boldsymbol{b}) - \mathfrak{c}(\boldsymbol{b})| \leq |\widehat{\mu}_{AC}(\boldsymbol{b}) - \mu_{AC}(\boldsymbol{b})| + |\widehat{\mu}_A(\boldsymbol{b}) - \mu_A(\boldsymbol{b})|\,|\widehat{\mu}_C(\boldsymbol{b})| \\ + |\mu_A(\boldsymbol{b})|\,|\widehat{\mu}_C(\boldsymbol{b}) - \mu_C(\boldsymbol{b})| \leq 3t.\end{aligned}$$

The two bounds. By (I.3), $|g(\boldsymbol{b})h(\boldsymbol{b})| = 1$, and $\sum_{\boldsymbol{b}} \widehat{p}_{\boldsymbol{b}} = 1$,

$$|\widehat{\mathbf{S}}[Ag, Ch]| \leq \sum_{\boldsymbol{b}} \widehat{p}_{\boldsymbol{b}} |\widehat{\mathfrak{c}}(\boldsymbol{b})| \leq \beta + 3t.$$

This holds simultaneously for every entry. For the lower bound, let $\widehat{\boldsymbol{w}} \in \mathbb{R}^{2^{r_B}}$ have entries $\widehat{p}_{\boldsymbol{b}}\widehat{\mathfrak{c}}(\boldsymbol{b})$, with zero entries at unsupported states. Taking $g = B_\Gamma$ and $h = 1$ in (I.3) shows that the entries of $\mathbf{H}_{r_B}\widehat{\boldsymbol{w}}$ all occur in $\widehat{\mathbf{S}}$. Since $\mathbf{H}_{r_B}^\top \mathbf{H}_{r_B} = 2^{r_B}\mathbf{I}$,

$$\|\widehat{\mathbf{S}}\|_{\max} \geq \|\widehat{\boldsymbol{w}}\|_2 \geq \max_{\boldsymbol{b}} \widehat{p}_{\boldsymbol{b}} |\widehat{\mathfrak{c}}(\boldsymbol{b})|.$$

If $|\mathfrak{c}(\boldsymbol{b})| \geq 3t$, then

$$\widehat{p}_{\boldsymbol{b}} |\widehat{\mathfrak{c}}(\boldsymbol{b})| \geq \frac{p_{\boldsymbol{b}}}{2}\{|\mathfrak{c}(\boldsymbol{b})| - 3t\} \geq \frac{p_{\boldsymbol{b}}}{2}|\mathfrak{c}(\boldsymbol{b})| - \frac{3t}{2}.$$

If $|\mathfrak{c}(\boldsymbol{b})| < 3t$, the last expression is nonpositive, so the same final lower bound remains valid. Taking the maximum over cells and substituting $t = 2\epsilon_n$ proves (I.2). Under conditional independence, $\mathfrak{c}(\boldsymbol{b}) = 0$ on every supported state, proving the final assertion. □

I.4 UNIFORM BOUNDS FOR THE SUBGROUP SEARCH

Consider the search defined in Section I.1. For a subgroup $\mathcal{K} \leq \langle \boldsymbol{X}_M \rangle$ of order 2^k, choose $g_1, \ldots, g_k \in \mathcal{K}$ so that every element of $\mathcal{K}$ is a unique product of a subset of these interactions. Let

$$\boldsymbol{B}_{\mathcal{K}} = \big(g_1(\boldsymbol{X}_M), \ldots, g_k(\boldsymbol{X}_M)\big) \in \{\pm 1\}^k.$$

For $\mathcal{K} = \{1\}$ this is the empty conditioning vector. Let $\mathcal{F}$ be the exact, nonempty finite family of choices $\tau = (i, j, M, \mathcal{K})$ searched when computing the bit-pair scores. The family is fixed before the data are seen. For each $\tau \in \mathcal{F}$, let h_τ be the number of supported states of $\boldsymbol{B}_{\mathcal{K}}$ and π_τ their smallest probability. Define

$$c_\tau(\boldsymbol{b}) = \mathrm{Cov}(X_i, X_j \mid \boldsymbol{B}_{\mathcal{K}} = \boldsymbol{b}), \qquad \beta_\tau = \max_{\boldsymbol{b}} |c_\tau(\boldsymbol{b})|, \qquad \eta_\tau = \max_{\boldsymbol{b}} \mathbf{P}(\boldsymbol{B}_{\mathcal{K}} = \boldsymbol{b}) |c_\tau(\boldsymbol{b})|,$$

where both maxima are over $\mathrm{Supp}(\boldsymbol{B}_{\mathcal{K}})$. Write $\mathbf{S}_{i,j,M,\mathcal{K}}$ for the population counterpart of the empirical cross block used in the score.

Corollary I.2 (Guarantees for the searched residuals)**.** *For every searched choice $\tau = (i, j, M, \mathcal{K})$,*

$$\mathbf{S}_{i,j,M,\mathcal{K}} = \mathbf{0} \quad \Longleftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid \boldsymbol{B}_{\mathcal{K}} \quad \Longleftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid \sigma\{g(\boldsymbol{X}_M) : g \in \mathcal{K}\}.$$

Suppose the n observations of $\boldsymbol{X}$ are IID. Put $N = |\mathcal{F}|$, fix $0 < \delta < 1$, and let

$$\epsilon_\tau = \sqrt{\frac{\log(12 N h_\tau/\delta)}{n \pi_\tau}}.$$

If $n\pi_\tau \geq 8 \log(2Nh_\tau/\delta)$ for every $\tau \in \mathcal{F}$, then with probability at least $1 - \delta$, simultaneously for all $\tau = (i, j, M, \mathcal{K}) \in \mathcal{F}$,

$$\frac{1}{2}\eta_\tau - 3\epsilon_\tau \;\leq\; \|\widehat{\mathbf{S}}_{i,j,M,\mathcal{K}}\|_{\max} \;\leq\; \beta_\tau + 6\epsilon_\tau.$$

For any scored pair, let $\mathcal{F}_{ij}$ denote its nonempty subset of choices in $\mathcal{F}$. Because $\mathcal{F}$ is exactly the family used in the search, the event in the corollary also gives

$$\min_{\tau \in \mathcal{F}_{ij}} \left(\frac{1}{2}\eta_\tau - 3\epsilon_\tau \right) \leq Q_{\mathrm{bit}}(i, j) \leq \min_{\tau \in \mathcal{F}_{ij}} (\beta_\tau + 6\epsilon_\tau).$$

In particular, any searched choice satisfying conditional independence gives $Q_{\mathrm{bit}}(i, j) \leq 6\epsilon_\tau$. The bounds also pass to the block scores through the maximum in (I.1). These bounds concern residual scores; graph recovery requires additional conditions on the variable-level procedure. A proper subgroup changes the conditioning information, so its zero block is not a test of conditional independence given all of $\boldsymbol{X}_M$.

The need for all endpoint interactions can already be seen with two independent fair signs Y_1, Y_2. Let $A = Y_1 Y_2 \epsilon$, where $\epsilon \in \{\pm 1\}$ is independent of (Y_1, Y_2) and $\mathbf{E}[\epsilon] \neq 0$. Then A is independent of each Y_i separately, but $\mathbf{E}[A Y_1 Y_2] = \mathbf{E}[\epsilon] \neq 0$, so A is not independent of the vector (Y_1, Y_2). Scalar pair scores can therefore omit dependence detected by the full vector characterization.

Proof of Corollary I.2. The chosen interactions give the one-to-one correspondence $\Gamma \mapsto \prod_{l=1}^k g_l^{\Gamma_l}$ from $\{0, 1\}^k$ to $\mathcal{K}$. Addition modulo two of exponent vectors corresponds to multiplication in $\mathcal{K}$. Evaluation preserves multiplication, so the interactions of $\boldsymbol{B}_{\mathcal{K}}$ are exactly the features indexed by $\mathcal{K}$. Since $i, j \notin M$, adjoining either endpoint retains distinct formal labels. Thus the separator and endpoint blocks agree with those in Theorem 3.2 applied to $(X_i, \boldsymbol{B}_{\mathcal{K}}, X_j)$, including when their evaluations satisfy deterministic relations. The theorem proves the first equivalence. The second follows because the chosen interactions and all their products generate the same sigma-field.

Apply Theorem I.1 to each member of $\mathcal{F}$ with δ/N in place of δ. Its support parameters are h_τ and π_τ, and its error term becomes ϵ_τ. A union bound over the N choices proves simultaneous coverage. Taking minima over exactly the searched choices gives the stated score bounds. □

J WHY AN ARBITRARY REDUCED CONDITIONING BLOCK CAN FAIL

The full set $\mathcal{B} = \langle \boldsymbol{B} \rangle \setminus \{1\}$ is a universally valid conditioning block. More generally, a reduced block gives the same residuals whenever it and the constant span every function on $\mathrm{Supp}(\boldsymbol{B})$, while the full endpoint interaction blocks are retained. Indeed, both blocks then define the same orthogonal projection $Z \mapsto \mathbf{E}[Z \mid \boldsymbol{B}]$. Such a reduction is possible when structural zeros create redundant features. Without this spanning condition, neither implication is universally guaranteed. The following examples omit one separator interaction while retaining the full endpoint interaction blocks.

Let $\boldsymbol{B} = (B_1, B_2)$, where B_1 and B_2 are independent Rademacher variables, and let A and C be scalar binary variables. Then

$$\mathcal{B} = \{B_1, B_2, B_1B_2\}, \qquad \mathcal{L} = \{A, AB_1, AB_2, AB_1B_2\}, \qquad \mathcal{R} = \{C, B_1C, B_2C, B_1B_2C\}.$$

Let $\boldsymbol{\Sigma}$ be the covariance matrix of the interaction vector indexed by $\mathcal{B} \cup \mathcal{L} \cup \mathcal{R}$, ordered as $(\mathcal{B}, \mathcal{L}, \mathcal{R})$. We omit the interaction B_1B_2 and replace $\mathcal{B}$ by

$$\widetilde{\mathcal{B}} := \{B_1, B_2\} \subsetneq \mathcal{B}.$$

Let $\widetilde{\mathbf{S}}$ denote the generalized Schur complement obtained from Theorem 3.2(c) by using $\widetilde{\mathcal{B}}$ in place of $\mathcal{B}$:

$$\widetilde{\mathbf{S}} := \boldsymbol{\Sigma}[\mathcal{L} \cup \mathcal{R}, \mathcal{L} \cup \mathcal{R}] - \boldsymbol{\Sigma}[\mathcal{L} \cup \mathcal{R}, \widetilde{\mathcal{B}}]\, \boldsymbol{\Sigma}_{\widetilde{\mathcal{B}}}^{+}\, \boldsymbol{\Sigma}[\widetilde{\mathcal{B}}, \mathcal{L} \cup \mathcal{R}].$$

Because B_1 and B_2 are independent with mean zero and variance one,

$$\boldsymbol{\Sigma}_{\widetilde{\mathcal{B}}} = \mathbf{I}_2.$$

Hence, for every $L \in \mathcal{L}$ and $R \in \mathcal{R}$,

$$\widetilde{\mathbf{S}}[L, R] = \mathrm{Cov}(L, R) - \mathrm{Cov}(L, B_1)\,\mathrm{Cov}(B_1, R) - \mathrm{Cov}(L, B_2)\,\mathrm{Cov}(B_2, R). \tag{J.1}$$

Define

$$J := \frac{1 + B_1 + B_2 + B_1B_2}{4} = \mathbf{1}\{B_1 = 1,\ B_2 = 1\}.$$

If $D \in \langle B_1, B_2 \rangle = \{1, B_1, B_2, B_1B_2\}$, then $JD = J$. Therefore, for all $D, E \in \langle B_1, B_2 \rangle$,

$$\mathbf{E}[JD] = \mathbf{E}[JDE] = \mathbf{E}[J] = \frac{1}{4}. \tag{J.2}$$

Proposition J.1. *With the notation above, the following statements both hold.*

(a) There exists a model with $A \perp\!\!\!\perp C \mid \boldsymbol{B}$ but $\widetilde{\mathbf{S}}[\mathcal{L}, \mathcal{R}] \neq \mathbf{0}$.

(b) There exists a model with $\widetilde{\mathbf{S}}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$ but $A \not\perp\!\!\!\perp C \mid \boldsymbol{B}$.

Proof. (a) Fix $0 < \rho \leq \frac{1}{2}$. Let A and C be conditionally independent given $\boldsymbol{B}$, with

$$\mathbf{E}[A \mid \boldsymbol{B}] = \mathbf{E}[C \mid \boldsymbol{B}] = \rho J.$$

Fix $L \in \mathcal{L}$ and $R \in \mathcal{R}$. Then $L = AD$ and $R = CE$ for some $D, E \in \langle B_1, B_2 \rangle$. By (J.2),

$$\mathbf{E}[L] = \mathbf{E}[D\,\mathbf{E}[A \mid \boldsymbol{B}]] = \rho\,\mathbf{E}[JD] = \frac{\rho}{4}, \qquad \mathbf{E}[R] = \frac{\rho}{4}.$$

Also, since $\mathbf{E}[B_i] = 0$,

$$\mathrm{Cov}(L, B_i) = \mathbf{E}[ADB_i] = \mathbf{E}[DB_i\,\mathbf{E}[A \mid \boldsymbol{B}]] = \rho\,\mathbf{E}[JDB_i] = \frac{\rho}{4}, \qquad i = 1, 2,$$

and similarly $\mathrm{Cov}(B_i, R) = \rho/4$ for $i = 1, 2$. Because $A \perp\!\!\!\perp C \mid \boldsymbol{B}$,

$$\mathbf{E}[LR] = \mathbf{E}[DE\,\mathbf{E}[A \mid \boldsymbol{B}]\mathbf{E}[C \mid \boldsymbol{B}]] = \rho^2\,\mathbf{E}[JDE] = \frac{\rho^2}{4},$$

so

$$\mathrm{Cov}(L, R) = \frac{\rho^2}{4} - \frac{\rho^2}{16} = \frac{3\rho^2}{16}.$$

Substituting into (J.1) gives

$$\widetilde{\mathbf{S}}[L, R] = \frac{3\rho^2}{16} - 2\left(\frac{\rho}{4}\right)^2 = \frac{\rho^2}{16}.$$

Thus every entry of $\widetilde{\mathbf{S}}[\mathcal{L}, \mathcal{R}]$ equals $\rho^2/16$, and in particular $\widetilde{\mathbf{S}}[\mathcal{L}, \mathcal{R}] \neq \mathbf{0}$ although $A \perp\!\!\!\perp C \mid \boldsymbol{B}$.

(b) Keep the same conditional means,

$$\mathbf{E}[A \mid \boldsymbol{B}] = \mathbf{E}[C \mid \boldsymbol{B}] = \rho J,$$

but define the conditional joint distribution of (A, C) given $\boldsymbol{B}$ by

$$\mathbf{P}(A = a, C = c \mid \boldsymbol{B} = (1,1)) = \frac{1 + a\rho + c\rho + ac\,(3\rho^2/4)}{4}, \qquad a, c \in \{\pm 1\},$$

and let (A, C) be conditionally uniform on $\{\pm 1\}^2$ for the other three values of $\boldsymbol{B}$. For $0 < \rho \leq 1/2$ this is valid, since the four probabilities at $\boldsymbol{B} = (1,1)$ are nonnegative. The construction yields

$$\mathbf{E}[A \mid \boldsymbol{B}] = \mathbf{E}[C \mid \boldsymbol{B}] = \rho J, \qquad \mathbf{E}[AC \mid \boldsymbol{B}] = \frac{3\rho^2}{4}\, J.$$

Since

$$\mathbf{E}[AC \mid \boldsymbol{B} = (1,1)] = \frac{3\rho^2}{4} \neq \rho^2 = \mathbf{E}[A \mid \boldsymbol{B} = (1,1)]\mathbf{E}[C \mid \boldsymbol{B} = (1,1)],$$

we have $A \not\perp\!\!\!\perp C \mid \boldsymbol{B}$.

Now fix $L = AD \in \mathcal{L}$ and $R = CE \in \mathcal{R}$, with any $D, E \in \langle B_1, B_2 \rangle$. The calculations for $\mathbf{E}[L]$, $\mathbf{E}[R]$, $\mathrm{Cov}(L, B_i)$, and $\mathrm{Cov}(B_i, R)$ are the same as in part (a), so

$$\mathbf{E}[L] = \mathbf{E}[R] = \frac{\rho}{4}, \qquad \mathrm{Cov}(L, B_i) = \mathrm{Cov}(B_i, R) = \frac{\rho}{4}, \quad i = 1, 2.$$

Moreover,

$$\mathbf{E}[LR] = \mathbf{E}[DE\,\mathbf{E}[AC \mid \boldsymbol{B}]] = \frac{3\rho^2}{4}\,\mathbf{E}[JDE] = \frac{3\rho^2}{16},$$

hence

$$\mathrm{Cov}(L, R) = \frac{3\rho^2}{16} - \frac{\rho^2}{16} = \frac{\rho^2}{8}.$$

Substituting into (J.1), we obtain

$$\widetilde{\mathbf{S}}[L, R] = \frac{\rho^2}{8} - 2\left(\frac{\rho}{4}\right)^2 = 0.$$

Thus every entry of $\widetilde{\mathbf{S}}[\mathcal{L}, \mathcal{R}]$ is zero, so $\widetilde{\mathbf{S}}[\mathcal{L}, \mathcal{R}] = \mathbf{0}$ although $A \not\perp\!\!\!\perp C \mid \boldsymbol{B}$. □

This proposition shows that deleting $B_1 B_2$ does not give a universally valid criterion, even though (B_1, B_2) has full support. It does not preclude support-dependent reductions that preserve the entire space of conditioning functions. A search over proper subgroups that also changes the endpoint interaction blocks is a different operation: it tests conditional independence given the information generated by that subgroup, as described in Appendix I.1.

K Examples of interaction groups and conditioning

The notation in Section 3 records which binary signs enter an interaction. The following examples explain the multiplication rule, the shared conditioning group, and the role of deterministic relations.

Multiplication and addition modulo two. For binary signs, the relevant collection consists of formal interaction labels. Since $X_j^2 = 1$, multiplying two interactions removes every sign that appears twice. For example,

$$(X_1 X_2)(X_2 X_3) = X_1 X_3.$$

Addition modulo two records whether each factor occurs an even or odd number of times. Applying the rule from Section 3 separately to each coordinate gives

$$(1, 1, 0) \oplus (0, 1, 1) = (1, 0, 1).$$

Thus multiplication of interaction labels is addition modulo two of their exponent vectors. The zero exponent vector labels the identity 1, and every interaction is its own inverse. Associativity follows from associativity of addition modulo two.

A two-bit group and its subgroups. For two signs, the full interaction group is

$$\langle B_1, B_2\rangle = \{1, B_1, B_2, B_1B_2\}.$$

For example, $B_1(B_1B_2) = B_2$ and $(B_1B_2)^2 = 1$. The five subgroups are

$$\{1\}, \qquad \{1, B_1\}, \qquad \{1, B_2\}, \qquad \{1, B_1B_2\}, \qquad \langle B_1, B_2\rangle.$$

The subgroup $\{1, B_1B_2\}$ retains the interaction B_1B_2 without retaining either individual bit. If all four states of (B_1, B_2) have positive probability, this interaction distinguishes equal signs from opposite signs, but does not determine the two signs separately. Conditioning on this subgroup therefore means conditioning on B_1B_2, which gives less information than conditioning on (B_1, B_2). If the supported states make one retained interaction determine both signs, the two conditioning operations can instead agree. This distinction explains why a subgroup search must state which information it conditions on.

The group intersection in BEGIN. For three coordinate labels A, B, C, the two groups are

$$\langle A, B\rangle = \{1, A, B, AB\}, \qquad \langle B, C\rangle = \{1, B, C, BC\}.$$

Their intersection is $\langle B\rangle = \{1, B\}$. Removing the shared group gives the left and right sets $\mathcal{L} = \{A, AB\}$ and $\mathcal{R} = \{C, BC\}$; removing the constant from the shared group gives $\mathcal{B} = \{B\}$. The covariance matrix in Theorem 3.2 therefore uses the five interactions

$$(B, A, AB, C, BC).$$

Projecting the four endpoint interactions onto the constant and B leaves their conditional-mean residuals. The $(\mathcal{L}, \mathcal{R})$ block of their covariance is zero exactly when $A \perp\!\!\!\perp C \mid B$. This is the scalar example in Section 3.3 and Figure 1(a). No independence assumption on the original three signs is needed to define the groups or their intersection.

Distinct labels under a deterministic relation. Formal labels remain distinct when their random-variable values agree. To see why, let X_1 be a fair binary sign and set $X_2 = X_1$. The formal group still has four labels $1, X_1, X_2, X_1X_2$, but $X_2 = X_1$ and $X_1X_2 = 1$ almost surely. Accordingly,

$$\mathrm{Cov}\begin{pmatrix} X_1 \\ X_2 \\ X_1X_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Its rank is one, in agreement with Theorem 3.1, since the supported states are $(1, 1)$ and $(-1, -1)$. Keeping the original labels lets the covariance record these deterministic relations. Group intersections are taken using the labels before evaluating them on the support, and the generalized inverse accommodates the resulting singularity.

Recovering the original signs from the interactions. In the example in Section 3.3 and Figure 1(c), the three products A_1A_2, A_1A_3, and $A_1A_2A_3$ determine the original signs through

$$A_1 = (A_1A_2A_3)(A_1A_2)(A_1A_3),$$
$$A_2 = (A_1A_2A_3)(A_1A_3), \qquad A_3 = (A_1A_2A_3)(A_1A_2).$$

Consequently, conditional on (A_1A_2, A_1A_3), independence of B and $A_1A_2A_3$ is equivalent to independence of B and (A_1, A_2, A_3). This equivalence holds on arbitrary support.

Log-linear models and a higher-order interaction. For any strictly positive binary pmf π, the function $\log \pi(\boldsymbol{x})$ is defined on the whole binary cube. Let θ_Λ be the interaction coefficients of $\log \pi$. The Hadamard basis gives

$$\log \pi(\boldsymbol{x}) = \theta_{\mathbf{0}} + \sum_{\Lambda \neq \mathbf{0}} \theta_\Lambda x_\Lambda, \tag{K.1}$$

with

$$\theta_\Lambda = 2^{-p} \sum_{\boldsymbol{x} \in \{\pm 1\}^p} x_\Lambda \log \pi(\boldsymbol{x}), \qquad x_{\mathbf{0}} = 1.$$

Keeping only terms with one or two nonzero exponent entries gives the Ising family. To see a distribution with an interaction of higher order, fix $0 < |\rho| < 1$ and let X_1, X_2, X_3 have pmf

$$\pi(x_1, x_2, x_3) = \frac{1 + \rho x_1 x_2 x_3}{8}.$$

Summing over any coordinate makes each remaining pair uniform, so every pair is independent. Conditional on $X_3 = b$, both X_1 and X_2 have mean zero and

$$\mathrm{Cov}(X_1, X_2 \mid X_3 = b) = \rho b.$$

Thus the BEGIN block with endpoints X_1, X_2 and conditioning variable X_3 is

$$\mathbf{S}[\{X_1, X_1X_3\}, \{X_2, X_2X_3\}] = \begin{pmatrix} 0 & \rho \\ \rho & 0 \end{pmatrix}.$$

Its log pmf has a nonzero third-order interaction coefficient $\frac{1}{2}\log\{(1+\rho)/(1-\rho)\}$, so this distribution is outside the pairwise Ising family. The same pmf formula remains valid at $\rho = 1$ or $\rho = -1$, when four states have probability zero; the displayed BEGIN block also remains valid. Theorem 3.2 covers both the positive and restricted-support cases.

L Proof and calibration of the sparse interaction test

The observations $(A_i, \boldsymbol{B}_i, C_i)$ are independent and identically distributed. The scalar endpoints A_i, C_i take values in $\{-1, 1\}$. The collection $\mathcal{T}$ contains $M \geq 2$ distinct conditioning interactions and may include the constant. The conditional endpoint mean $m_A(\boldsymbol{B}) = \mathbf{E}[A \mid \boldsymbol{B}]$ is known. This is equivalent to knowing the binary assignment probabilities, since $\mathbf{P}(A = a \mid \boldsymbol{B}) = \{1 + am_A(\boldsymbol{B})\}/2$. No model for $\mathbf{E}[C \mid \boldsymbol{B}]$ is required.

L.1 Finite-sample level and power

Write

$$\mathfrak{c}(\boldsymbol{B}) = \mathrm{Cov}(A, C \mid \boldsymbol{B}), \qquad \beta_\Gamma = \mathbf{E}[B_\Gamma \mathfrak{c}(\boldsymbol{B})], \qquad u_\alpha = \sqrt{\frac{2\log(2M/\alpha)}{n}}.$$

Conditional centering gives

$$\mathbf{E}[\{A - m_A(\boldsymbol{B})\}C \mid \boldsymbol{B}] = \mathbf{E}[AC \mid \boldsymbol{B}] - m_A(\boldsymbol{B})\mathbf{E}[C \mid \boldsymbol{B}] = \mathfrak{c}(\boldsymbol{B}).$$

Thus $\widehat{\beta}_\Gamma$ is unbiased for the displayed BEGIN coordinate. Under $A \perp\!\!\!\perp C \mid \boldsymbol{B}$, condition on all observed $(\boldsymbol{B}_i, C_i)$. For each fixed Γ, the summands $\{A_i - m_A(\boldsymbol{B}_i)\}C_iB_{i,\Gamma}$ are then independent, have mean zero, and lie in intervals of length two. The exponential bound for bounded independent random variables yields

$$\mathbf{P}_0\left(|\widehat{\beta}_\Gamma| > t \mid (\boldsymbol{B}_i, C_i)_{i=1}^n\right) \leq 2\exp(-nt^2/2).$$

A union bound over M interactions proves the level statement, even conditional on the observed conditioning variables and responses. Neither fair conditioning bits nor sparsity is required.

Under any alternative, each summand lies in $[-2, 2]$. Choose a coordinate with largest absolute population mean and multiply by its population sign. The one-sided bounded-variable inequality gives

$$\mathbf{P}\left\{\max_{\Gamma\in\mathcal{T}}|\widehat{\beta}_\Gamma| \leq \max_{\Gamma\in\mathcal{T}}|\beta_\Gamma| - t\right\} \leq \exp(-nt^2/8).$$

Consequently a sufficient condition for power at least $1 - \delta$ is

$$\max_{\Gamma\in\mathcal{T}}|\beta_\Gamma| > \sqrt{\frac{2\log(2M/\alpha)}{n}} + \sqrt{\frac{8\log(1/\delta)}{n}}. \tag{L.1}$$

If $m_A \equiv 0$, all summands lie in $[-1, 1]$, so two can replace eight in the second term.

For independent fair conditioning bits, distinct interactions are orthonormal: $\mathbf{E}[B_\Gamma B_{\Gamma'}]$ equals one if the labels agree and zero otherwise. If $\mathfrak{c}(\boldsymbol{B}) = \sum_{\Gamma\in\mathcal{T}} \beta_\Gamma B_\Gamma$ has at most h nonzero coefficients, then

$$\mathbf{E}[\mathfrak{c}(\boldsymbol{B})^2] = \sum_{\Gamma\in\mathcal{T}} \beta_\Gamma^2 \le h \max_{\Gamma\in\mathcal{T}} |\beta_\Gamma|^2.$$

Combining this identity with (L.1) proves the explicit upper bound

$$\mathbf{E}[\mathfrak{c}(\boldsymbol{B})^2] > h \left\{ \sqrt{\frac{2\log(2M/\alpha)}{n}} + \sqrt{\frac{8\log(1/\delta)}{n}} \right\}^2. \tag{L.2}$$

For fixed error probabilities this is of order $h \log M/n$.

L.2 A matching lower bound with several active interactions

It is enough to consider the subclass with independent fair conditioning bits and balanced endpoint margins. Let P_0 make all variables independent fair signs. For a set $J \subseteq \mathcal{T}$ of exactly h labels and signs $\xi_\Gamma \in \{-1, 1\}$, define

$$f_{J,\xi}(\boldsymbol{B}) = a \sum_{\Gamma\in J} \xi_\Gamma B_\Gamma, \qquad \mathbf{P}_{J,\xi}(A = x, C = y \mid \boldsymbol{B}) = \frac{1 + xy f_{J,\xi}(\boldsymbol{B})}{4}.$$

Here $a > 0$ is a coefficient magnitude. If $ha \le 1$, every conditional probability is nonnegative, and it is positive when $ha < 1$. The conditional endpoint means are zero, the conditional covariance is $f_{J,\xi}$, and $\mathbf{E}[f_{J,\xi}(\boldsymbol{B})^2] = ha^2$. These are exactly h-interaction alternatives, all with the same known assignment law $\mathbf{P}(A = 1 \mid \boldsymbol{B}) = 1/2$.

Choose J uniformly among the h-element subsets of $\mathcal{T}$, choose its signs independently and fairly, and mix the corresponding n-observation laws. Denote this mixture by $\overline{P}$. For two independent draws (J, ξ) and (J', ξ') from the mixture, the likelihood ratios L, L' relative to $P_0^{\otimes n}$ satisfy

$$\mathbf{E}_0[LL'] = \left(1 + a^2 \sum_{\Gamma\in J\cap J'} \xi_\Gamma \xi'_\Gamma\right)^n \le \exp\left(na^2 \sum_{\Gamma\in J\cap J'} \xi_\Gamma \xi'_\Gamma\right).$$

The equality follows from orthonormality of the conditioning interactions; the inequality uses $1 + x \le e^x$. Let $K = |J \cap J'|$. Conditional on the two supports, averaging over their independent signs gives $\cosh(na^2)^K$. For $z \ge 1$, the hypergeometric overlap obeys

$$\begin{aligned}
\mathbf{E}[z^K] &= \sum_{\ell=0}^{h} \mathbf{E}\binom{K}{\ell}(z-1)^\ell \\
&= \sum_{\ell=0}^{h} \binom{h}{\ell} \frac{h(h-1)\cdots(h-\ell+1)}{M(M-1)\cdots(M-\ell+1)}(z-1)^\ell \\
&\le \{1 + (h/M)(z-1)\}^h \le \exp\{(h^2/M)(z-1)\}.
\end{aligned}$$

The product ratio for $\ell = 0$ equals one. Each factor in the ratio is no larger than h/M. Therefore the chi-square divergence satisfies

$$\chi^2(\overline{P}, P_0^{\otimes n}) \le \exp\left[\frac{h^2}{M}\{\cosh(na^2) - 1\}\right] - 1. \tag{L.3}$$

Fix desired error probabilities with $\Delta = 1 - \alpha - \delta > 0$, and choose $a = \rho/\sqrt{h}$. The condition $\rho\sqrt{h} \le 1$ ensures a valid probability mass function, allowing structural zeros at the boundary. Put

$$u = \log\left\{1 + \frac{M}{h^2}\log(1 + \Delta^2)\right\}.$$

If $n \le hu/\rho^2$, then $na^2 \le u$. Since $\cosh u - 1 \le e^u - 1$, inequality (L.3) gives $\chi^2(\overline{P}, P_0^{\otimes n}) \le \Delta^2$. Total variation distance is at most one half the square root of this divergence. For any randomized test $\phi \in [0, 1]$ of level at most α under P_0,

$$\mathbf{E}_{\overline{P}}\phi \le \alpha + \Delta/2 < 1 - \delta.$$

Some alternative in the mixture therefore has power less than $1-\delta$, although its squared signal equals ρ^2. Consequently uniform power $1-\delta$ requires

$$n > \frac{h}{\rho^2}\log\left\{1+\frac{M}{h^2}\log[1+(1-\alpha-\delta)^2]\right\}. \tag{L.4}$$

This is a lower bound for arbitrary tests, not only interaction tests. Knowing the assignment law does not affect the argument: that law is identical throughout the construction.

For fixed α, δ and $\varepsilon > 0$, if $h \le M^{1/2-\varepsilon}$, then $M/h^2 \ge M^{2\varepsilon}$. The logarithm in (L.4) is consequently bounded above and below by positive constant multiples of $\log M$ for $M \ge 2$. Rearranging (L.2) gives the matching sufficient sample order. The lower and upper bounds thus have the same order claimed in Theorem 4.1. □

L.3 Conditional randomization and its power

For each of $R \ge 1$ independent draws, hold $(\boldsymbol{B}_i, C_i)$ fixed and sample every A_i independently from its specified conditional law. Let $T_{\max}^{(\ell)}$ be the statistic from draw ℓ. Reject when

$$p_{\mathrm{CRT}} = \frac{1+\sum_{\ell=1}^{R}\mathbf{1}\{T_{\max}^{(\ell)} \ge T_{\max}\}}{R+1} \le \alpha. \tag{L.5}$$

Under conditional independence and conditional on the observed $(\boldsymbol{B}_i, C_i)$, the observed assignment vector and the R redraws are independent with the same law. Their statistics are exchangeable. Consequently the inclusive rank in (L.5) is a valid p value, including when ties occur, and the conditional rejection probability is at most α.

The finite-randomization test has the same sample order as Theorem 4.1 for fixed R with $1/(R+1) \le \alpha$. To see this, under any alternative define

$$u_R = \sqrt{\frac{2\log(4MR/\delta)}{n}}.$$

Given the observed $(\boldsymbol{B}_i, C_i)$, each randomized coordinate has mean zero and the same length-two bound used for null validity. A union bound therefore puts all R randomized maxima below or equal to u_R with probability at least $1-\delta/2$. The observed maximum exceeds u_R with probability at least $1-\delta/2$ whenever

$$\max_{\Gamma\in\mathcal{T}}|\beta_\Gamma| > u_R + \sqrt{\frac{8\log(2/\delta)}{n}}.$$

Both events then hold with probability at least $1-\delta$. On their intersection, $p_{\mathrm{CRT}} = 1/(R+1) \le \alpha$. The sparse bound $\max_\Gamma|\beta_\Gamma| \ge \rho/\sqrt{h}$ gives the asserted rate. Its constants may also depend on R.

L.4 Empirical BEGIN entries and unknown assignment laws

When the conditional law of A is unknown, let $\widehat{m}_A(\boldsymbol{b})$ and $\widehat{m}_C(\boldsymbol{b})$ be the corresponding sample means in each observed original conditioning cell $\boldsymbol{B} = \boldsymbol{b}$. Form the empirical interaction covariance using normalization $1/n$, and let $\widehat{\mathbf{S}}$ be its generalized Schur complement as in (3.2). The interactions and the constant span all functions on the empirical conditioning support. Hence their least-squares projections are exactly the cell means. It follows that

$$\begin{aligned}\widehat{\mathbf{S}}[AB_\Gamma, C] &= \frac{1}{n}\sum_i B_{i,\Gamma}\{A_i - \widehat{m}_A(\boldsymbol{B}_i)\}\{C_i - \widehat{m}_C(\boldsymbol{B}_i)\}\\ &= \frac{1}{n}\sum_i B_{i,\Gamma}\{A_i - \widehat{m}_A(\boldsymbol{B}_i)\}C_i.\end{aligned}$$

The second identity holds because the centered A_i sum to zero in each original conditioning cell. Replacing m_A by its cell mean therefore gives entries of the empirical BEGIN block.

For exact calibration, permute the observed A_i independently within each original conditioning cell, leaving $\boldsymbol{B}_i, C_i$ fixed, and recompute the maximum over the same fixed collection $\mathcal{T}$. Under

conditional independence, conditional on $(\boldsymbol{B}_i, C_i)$ and the number of positive assignments in each cell, the assignment vector is uniform over these within-cell arrangements. Independent uniform permutations therefore give exchangeable statistics, and the inclusive p value (L.5) is valid. The cell means are invariant under these permutations. If a cell contains one observation, its centered assignment is zero, which explains the loss of information when many cells are sparse. The power theorem assumes a known conditional law; its rate is not claimed for this procedure.

M EXACT BITWISE ALGORITHM

The following steps compute the statistic in (4.2). The candidate set is fixed before examining its maximum. We call an integer storing several observations a *packed word*. The *exclusive-or* operation, written XOR, adds its bits modulo two. The *population-count* operation counts the bits equal to one.

(1) Encode each sign x by the bit $(1 - x)/2$ and form the M conditioning products. Pack the observed bit columns into machine words.

(2) For each candidate Γ, form the bits of CB_Γ by exclusive-or. Compute once the offset $h_\Gamma = \sum_i m_A(\boldsymbol{B}_i) C_i B_{i,\Gamma}$.

(3) For the observed assignment and each independent conditional redraw, count d_Γ, the number of one bits in the exclusive-or of the assignment with CB_Γ. The coordinate sum is $n - 2d_\Gamma - h_\Gamma$. Mask unused bits in the final word before counting.

(4) Record the largest absolute coordinate sum, and compute the inclusive p value (L.5). The common divisor n does not change its ranking.

The bitwise sign count is an exact integer identity. Only the common centering offsets require floating-point arithmetic under nonuniform assignment. Let w be the number of bits in a machine word. With precomputed features, each assignment costs $O(M\lceil n/w \rceil)$ word operations. Forming order-at-most-k features costs at most $O(nkM)$ elementary operations; generating R assignments costs $O(Rn)$. The complete procedure therefore includes these costs in addition to the packed scans. The supplied code checks direct products, count-based and packed implementations on identical data and redraws before timing.

N SPARSE INTERACTION SIMULATIONS

Design. The eight coordinates of $\boldsymbol{B}$, and A, are independent fair signs. Given $(A, \boldsymbol{B})$, we generate the sign C with conditional mean $A\mathfrak{c}(\boldsymbol{B})$. For $h = 1, 2, 4$,

$$\mathfrak{c}(\boldsymbol{B}) = \frac{0.4}{\sqrt{h}} \sum_{j=1}^{h} B_{2j-1} B_{2j}.$$

The dense alternative is $\mathfrak{c}(\boldsymbol{B}) = 0.1(B_1, \ldots, B_4) H (B_5, \ldots, B_8)^\top$, where

$$H = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

It has 16 nonzero pair coefficients. Every alternative satisfies $\mathbf{E}\mathfrak{c}(\boldsymbol{B})^2 = 0.16$. Since $H^\top H = 4I$, Cauchy–Schwarz gives $|\mathfrak{c}(\boldsymbol{B})| \leq 0.8$ in the dense case; the same bound holds for the sparse cases. Thus all joint probabilities are positive. The null has $\mathfrak{c} = 0$. There are 200 replications of each distribution at each total sample size $n = 128, 256, 512, 1024$, giving 4000 datasets. We fixed the seed, supports, coefficients, method settings, and simulation sizes before comparing power.

Tests. All scans contain the same 37 products of at most two conditioning bits, including the constant. Max BEGIN takes the largest absolute empirical mean of AC times one of these products. The full G^2 statistic (Bishop et al., 1975; Agresti, 2013) sums the likelihood ratio statistics over

the original 256 conditioning cells. Gaussian KCI uses the empirical bandwidth and sample standardization as proposed in Zhang et al. (2011), as implemented in causal-learn (Zheng et al., 2024), with regularization 0.001. We use its Gaussian statistic and replace the asymptotic calibration by the shared randomizations, denoted KCI-CRT. The bandwidth multiplier is 1.2 for $n = 128$ and 0.7 for the other sample sizes. The forest-weighted CRT, motivated by the learned weights in Scheidegger et al. (2022), learns $\mathbf{E}(AC \mid \boldsymbol{B})$ using 100 regression trees on the first half sample, with minimum leaf size 10, all eight features eligible at a split, and random state 33. It tests the weighted residual numerator on the second half, using the sign of the fitted conditional mean as the weight and conditional randomization for calibration. The published WGCM uses studentized residual products. The conditional endpoint means are known to be zero in these designs.

All four statistics use the same 199 independent randomizations of A from its known fair law, holding (B, C) fixed. The p value counts statistics at least as large as the observed statistic and uses the plus-one correction. This gives finite-sample validity and can be conservative when statistics tie. The analytic-threshold version of Max BEGIN rejects above $\sqrt{2\log(2 \cdot 37/0.05)/n}$. With the same fixed product weights, the weighted-GCM residual numerators equal the coordinates of Max BEGIN in this design; the published studentized statistic uses a different normalization.

Results and computation. Table 1 includes every setting. Max BEGIN rejects $5, 7, 7, 5$ of 200 null replications at the four sample sizes. The simulation archive provides the individual outcomes, paired differences, and 95% Wilson intervals for all rates. Timing includes interaction construction, fitting where required, generation of the randomizations, and calibration, but excludes generation of the statistical dataset. We report sequential single-thread medians over ten fixed null datasets at each sample size. The machine uses an Intel Xeon Platinum 8272CL CPU at 2.60 GHz, Python 3.12, NumPy 2.3.5, and scikit-learn 1.8.0. The G^2 implementation removes conditioning cells in which C is constant, whose contribution is always zero, and uses integer counts with a lookup table for $j \log j$. The optimized and direct G^2 scores agree within 2×10^{-13} on 4000 calibration statistics. Packed BEGIN uses XOR and population count; its scores match matrix multiplication exactly.

Table 1: Rejection probabilities at level 0.05 over 200 fixed replications. All nonlearning tests use the same total sample size and 199 conditional randomizations; forest weighting uses a half sample for learning. The maximum binomial standard error is $\sqrt{0.25/200} \approx 0.0354$. The last column uses the stated conservative analytic threshold.

n	h	Max BEGIN	G^2	KCI-CRT	Forest	Threshold
128	0	0.025	0.045	0.025	0.040	0.005
128	1	0.880	0.150	0.065	0.070	0.780
128	2	0.725	0.130	0.075	0.075	0.425
128	4	0.490	0.155	0.075	0.065	0.210
128	16	0.180	0.130	0.070	0.035	0.050
256	0	0.035	0.040	0.040	0.030	0.000
256	1	1.000	0.400	0.085	0.490	1.000
256	2	0.995	0.380	0.060	0.375	0.950
256	4	0.930	0.355	0.080	0.310	0.730
256	16	0.605	0.410	0.080	0.260	0.225
512	0	0.035	0.055	0.055	0.025	0.000
512	1	1.000	0.840	0.050	0.930	1.000
512	2	1.000	0.820	0.105	0.895	1.000
512	4	1.000	0.890	0.080	0.810	0.995
512	16	0.950	0.835	0.070	0.820	0.640
1024	0	0.025	0.030	0.030	0.025	0.000
1024	1	1.000	1.000	0.075	1.000	1.000
1024	2	1.000	1.000	0.095	1.000	1.000
1024	4	1.000	0.995	0.075	1.000	1.000
1024	16	1.000	1.000	0.120	1.000	1.000

Table 2: Median full-test wall time in milliseconds, including 199 randomizations. Packed C++ is used for the main BEGIN timing. Matrix and NumPy packed versions compute the same statistic.

	Max BEGIN					
n	C++	Matrix	NumPy	G^2	KCI-CRT	Forest
128	1.12	0.93	1.13	0.68	2.39	109.77
256	1.51	1.31	1.29	2.44	14.80	119.66
512	2.40	2.06	1.98	7.13	69.46	119.53
1024	4.11	3.50	2.87	9.47	234.31	132.98

O CATEGORICAL MAX BEGIN IN THE MUSHROOM DATA

O.1 DATA AND ENCODING

The UCI Mushroom data contain 8,124 hypothetical field-guide records with 22 categorical attributes and an edibility class. We retain the stratum with `gill.attachment=f`, `veil.type=p`, `veil.color=w`, and `ring.number=o`, omit `stalk.root`, and drop the four attributes made constant by the selection. This leaves 7,288 records, of which 3,488 are labeled edible and 3,800 poisonous, and 18 variables. Their numbers of categories are 2 (class, bruises, gill spacing, gill size, stalk shape), 3 (stalk surface above the ring), 4 (cap surface, stalk surface below the ring, ring type), 5 (stalk color above the ring, spore print color), 6 (cap shape, stalk color below the ring, population, habitat), 8 (odor, gill color), and 10 (cap color). The recorded poisonous class also contains records of unknown edibility, and the records are not a random ecological sample. The calibration below assumes independent observations from a common distribution; the construction of the benchmark does not provide such a sampling design. The stratum and encoding were fixed before this analysis; the error control below does not cover their selection.

Each variable receives the primary encoding: its categories are sorted by their one-letter codes, numbered from zero, written with the minimal number of bits, 1, 2, 3, or 4, and each bit is converted to a sign. All 18 variables use 41 bits. Unused codewords, for example the six unused patterns of cap color, have zero counts; Theorems 3.1 and 3.2 allow these structural zeros, and the computation below never divides by an empty cell.

O.2 THE STATISTIC AND ITS EXACT COMPUTATION

Let $\boldsymbol{A} \in \{\pm 1\}^{r_A}$ and $\boldsymbol{C} \in \{\pm 1\}^{r_C}$ be the encoded endpoints and $\boldsymbol{B} \in \{\pm 1\}^{r_B}$ the concatenated encoded conditioning variables. For nonzero $\Lambda_1 \in \{0,1\}^{r_A}$ and $\Lambda_2 \in \{0,1\}^{r_C}$ and every $\Gamma \in \{0,1\}^{r_B}$, including $\Gamma = \mathbf{0}$, let $\widehat{m}_{\Lambda_1}(\mathbf{b})$ be the mean of A_{Λ_1} among the records with $\boldsymbol{B} = \mathbf{b}$. Write A_{i,Λ_1}, $B_{i,\Gamma}$, and C_{i,Λ_2} for the corresponding interactions evaluated on record i. Define

$$\widehat{\beta}_{\Lambda_1\Lambda_2\Gamma} = \frac{1}{n}\sum_{i=1}^{n} B_{i,\Gamma}\big\{A_{i,\Lambda_1} - \widehat{m}_{\Lambda_1}(\boldsymbol{B}_i)\big\}C_{i,\Lambda_2},$$
$$T = \max_{\Lambda_1\neq\mathbf{0},\ \Lambda_2\neq\mathbf{0},\ \Gamma} \big|\widehat{\beta}_{\Lambda_1\Lambda_2\Gamma}\big|. \tag{O.1}$$

This extends (4.2) with all endpoint interactions and the within-cell mean of Appendix L.4 in place of a known conditional mean. By the identity in Appendix L.4, $\widehat{\beta}_{\Lambda_1\Lambda_2\Gamma}$ is the entry $\widehat{\mathbf{S}}[A_{\Lambda_1}B_\Gamma, C_{\Lambda_2}]$ of the empirical generalized Schur complement, and the entry with a conditioning product on both sides depends only on the modulo-two sum of the two conditioning labels. Hence $T = \|\widehat{\mathbf{S}}[\mathcal{L},\mathcal{R}]\|_{\max}$ is the maximum absolute empirical BEGIN entry. The statistic is symmetric in the two endpoints, because the centered A_{Λ_1} sums to zero in every cell and C_{Λ_2} may therefore be centered as well. On any support, all population entries vanish exactly under conditional independence (Theorem 3.2).

The computation uses the count table $N(\mathbf{b}, a, c)$ indexed by the conditioning state and the two endpoint codes, and the within-cell residuals $R(\mathbf{b}, a, c) = N(\mathbf{b}, a, c) - N(\mathbf{b}, a, \cdot)N(\mathbf{b}, \cdot, c)/N(\mathbf{b})$ for nonempty cells. For the next identity, the table indices $\mathbf{b}$, a, and c are the corresponding $0/1$ codewords. A sign product is (-1) raised to the inner product of its label and its codeword. Thus

$$n\,\widehat{\beta}_{\Lambda_1\Lambda_2\Gamma} = \sum_{\mathbf{b},a,c} (-1)^{\Gamma^\top\mathbf{b}+\Lambda_1^\top a+\Lambda_2^\top c} R(\mathbf{b}, a, c), \tag{O.2}$$

where the sum runs over the codewords. One Walsh–Hadamard transform of R over all $r = r_A + r_B + r_C$ bits returns every coordinate at once, at cost $O(r2^r)$ after $O(n)$ counting. Coordinates whose endpoint label is constant on the support are zero and cannot attain a positive maximum. Ties among maximizing labels are common, because distinct conditioning products coincide up to sign on the supported cells; the reported label is the first in scan order. The C++ implementation supports statements with $r \le 23$. In the scan over subsets of two to six variables described below, the encoding uses at most $r = 19$ bits per statement. The C++ implementation evaluated all 385,101 statements in approximately 130 seconds, averaging about 0.34 milliseconds per statement. Independent dense Hadamard contractions match all ten results within 10^{-15}.

When a pair is conditioned on all sixteen remaining variables, r_B reaches 39 and the dense table is unavailable. The observed statistic is then computed exactly by affine compression: keep the conditioning codes whose residual table is nonzero, fix one of them as origin, and form a basis of the differences of the others over $\mathbb{Z}_2$. Every product of the original conditioning bits restricts, up to the sign at the origin, to a product of the resulting coordinates, and every such product is a restriction because the basis is linearly independent. The maximum of the transform on the compressed cube, computed in exact integer arithmetic after clearing the cell denominators, therefore equals the maximum on the original cube, and a maximizing original label is recovered by solving the binary linear equations. The compression serves only the observed statistic; the calibration retains the full family of 2^{r_B} conditioning products.

O.3 Exact calibration and finite-sample tail bounds

Under $\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{C} \mid \boldsymbol{B}$ and independent observations from a common distribution, conditional on all conditioning states, all endpoint values $\boldsymbol{C}_i$, and the multiset of endpoint values $\boldsymbol{A}_i$ in each cell, the arrangement of the $\boldsymbol{A}_i$ within each cell is uniform (Appendix L.4). The within-cell means in (O.1) are invariant under these arrangements, so T has an exact conditional permutation distribution without a model for either endpoint. Its exact permutation p value is the probability that a shuffled dataset gives a maximum at least as large as the observed T, conditional on the cells and their endpoint category counts. The reported values are upper bounds on this probability. The search below requires permutation p values of order 10^{-9}, which simulation cannot resolve. We therefore bound the exact permutation tail.

Fix nonzero endpoint labels and a cell of size N with k records having $A_{\Lambda_1} = 1$ and ℓ having $C_{\Lambda_2} = 1$. Under a uniform arrangement, the number H of records with both signs equal to one is hypergeometric with population N, ℓ successes, and k draws, and the contribution from the cell to $n\widehat{\beta}_{\Lambda_1\Lambda_2\Gamma}$ is $B_\Gamma(\mathbf{b})\,Y$ with $Y = 4(H - k\ell/N)$. Cells with a constant endpoint product, including singletons, contribute exactly zero. The contributions of different cells are independent. For $t > 0$ let $L_{\mathbf{b}}(t) = \max\{\log \mathbb{E}e^{tY_{\mathbf{b}}}, \log \mathbb{E}e^{-tY_{\mathbf{b}}}\}$, a finite hypergeometric sum. Markov inequality applied to $e^{\pm tn\widehat{\beta}}$, a union over both signs, the 2^{r_B} conditioning products, and all nonconstant endpoint pairs (Λ_1, Λ_2) gives, for every $t > 0$,

$$\mathbb{P}\big(nT_{\mathrm{perm}} \ge u\big) \;\le\; \min\Big\{1,\; 2 \cdot 2^{r_B} \sum_{\Lambda_1,\Lambda_2} \exp\Big[-tu + \sum_{\mathbf{b}} L_{\Lambda_1\Lambda_2\mathbf{b}}(t)\Big]\Big\}. \tag{O.3}$$

The bound holds for every $t > 0$, so its infimum over $t > 0$ is also a valid bound. The finite sums are evaluated in floating point with a small upward numerical adjustment. The independent check below uses the simpler bound that follows. Sampling without replacement has exponential moments no larger than sampling with replacement (Hoeffding, 1963, Theorem 4), and the bounded-variable exponential inequality then gives $\log \mathbb{E}e^{tY} \le 2t^2 \min(k, N-k, \ell, N-\ell)$. With $D_{\Lambda_1\Lambda_2}$ the sum of these minima over cells, the simpler envelope $2 \cdot 2^{r_B} \sum_{\Lambda_1,\Lambda_2} \exp\{-u^2/(8D_{\Lambda_1\Lambda_2})\}$ is also valid for $u > 0$; terms with $D_{\Lambda_1\Lambda_2} = 0$ are omitted because their coordinates are identically zero. The implementation reports the smaller bound at $u = nT - 10^{-8}$ when this is positive, and reports one otherwise. Exhaustive enumeration of all arrangements on small tables confirms that both envelopes dominate the exact tail at every attainable threshold, and 2,000-draw Monte Carlo permutations agree with the exact tails. The union bound is conservative because different conditioning products can coincide on the observed cells.

Table 3: The exhaustive Max BEGIN scan of the Mushroom stratum.

Variables in subset	Subsets	Statements per subset	Statements
2	153	1	153
3	816	3	2,448
4	3,060	6	18,360
5	8,568	10	85,680
6	18,564	15	278,460
Total	31,161		385,101

Table 4: Panel (a) of Figure 3: each pair given the other three variables. Upper stalk surface denotes the surface above the ring. Cells are the supported conditioning states.

Variables		Max	Cells	Informative Cells	Informative Records	Adjusted bound
Class	Odor	0.3463	22	6	4,460	$< 10^{-100}$
Class	Cap shape	0.0022	18	1	1,832	not rejected
Class	Upper stalk surface	0.0115	34	3	840	$< 10^{-28}$
Class	Ring type	0.0011	42	2	1,132	not rejected
Odor	Cap shape	0.1240	9	4	4,984	$< 10^{-33}$
Odor	Upper stalk surface	0.0525	20	5	2,490	$< 10^{-6}$
Odor	Ring type	0.2299	19	7	6,132	$< 10^{-100}$
Cap shape	Upper stalk surface	0.0000	14	5	2,832	not rejected
Cap shape	Ring type	0.0607	15	3	4,296	$< 10^{-5}$
Upper stalk surface	Ring type	0.2211	27	7	4,638	$< 10^{-100}$

O.4 Search, selection, and error control

The C++ engine evaluated T for every pair within every subset of two to six variables, conditioning on the remaining variables of the subset: 31,161 subsets and 385,101 statements (Table 3), in about 130 seconds. For each of the $\binom{18}{2} = 153$ pairs, the conditioning set can contain zero to four of the other sixteen variables, giving

$$\binom{18}{2} \sum_{j=0}^{4} \binom{16}{j} = 385{,}101.$$

This scan produced only statistics and support diagnostics. Seven candidate subsets were chosen for interpretation in this exploratory analysis and their 63 statements were bounded by (O.3). We report the five-variable subset containing class, odor, cap shape, stalk surface above the ring, and ring type. To account for the selection, each bound q is multiplied by

$$F = \binom{18}{2} 2^{16} = 10{,}027{,}008, \tag{O.4}$$

the number of endpoint pairs times the number of conditioning subsets of the other sixteen variables, and conditional independence is rejected when $Fq \leq .05$. By the union bound this controls the familywise error over every statement that the selection could have chosen, including subsets larger than six and the choice of displayed variables, under the fixed stratum and encoding. Nonrejection of a statement is not evidence of conditional independence.

O.5 Results

Table 4 lists the ten conditional independence tests for Figure 3(a). A conditioning cell is *informative* when both tested variables vary within it. The table reports upper bounds on permutation p values after the selection adjustment, rather than estimates of those p values. All seven rejected statements have adjusted bounds below 10^{-5}.

The class and upper stalk surface comparison illustrates how Max BEGIN identifies which category groups are conditionally associated. Exactly three cells vary in both endpoints, all odorless with

an evanescent ring; their cap shapes are flat, knobbed, and convex. Within them, silky stalks are poisonous in 8, 8, and 8 records and edible in none, while fibrous or smooth stalks are edible in 400, 16, and 400 records and poisonous in none. The maximizing stalk interaction assigns opposite signs to the silky category and the other two, and the maximizing conditioning product is the constant, so the three cells reinforce one another. As in Proposition 3.3, the statistic is the probability-weighted sum of the three within-cell covariances: two cells of 408 records with covariance .077, since 8 poisonous against 400 edible records leave little variance, and one cell of 24 records with covariance .889, giving .0115. Its adjusted bound is still below 10^{-28}. Exact factorization of the cap shape–upper stalk surface table in all 2,832 informative records is not evidence of population independence.

O.6 Conditioning on all sixteen remaining variables

For each of the 153 pairs conditioned on the other sixteen variables, 147 empirical blocks are exactly zero: every supported cell is a singleton, has a constant endpoint, or has a factorizing table. All comparisons involving the class have no informative cell. The six nonzero blocks pair stalk surface below the ring with cap shape, cap surface, cap color, gill color, population, and habitat; their residuals live in eight, eight, eight, four, eight, and eight cells of affine rank three, three, three, two, three, and three, holding 24 to 32 records, and their exact maximal numerators nT are 16, $32/3$, 16, $96/7$, 16, and 16, so T ranges from .0015 to .0022. Every bound (O.3) equals one. These nonrejections do not establish an empty conditional independence graph on all 18 variables. In particular, the absence of conditional variation prevents the class comparisons from distinguishing independence from dependence. This is the loss of information from sparse cells noted in Appendix L.4.

O.7 Coding invariance and robustness

The statistic depends on the encoding through its product dictionary. A recoding that is an affine bijection of the bit cube sends every product to a product, possibly with reversed sign, so it permutes the coordinates up to sign and leaves T, its permutation law, and the bound (O.3) unchanged. For a variable with at most four categories and minimal codes, every injective recoding extends to a permutation of the four two-bit patterns, and all 24 such permutations are affine. Variables with five or more categories admit recodings that are not affine. The graph in Figure 3(a) contains such variables and therefore has no general invariance guarantee. Fifty random injective minimal encodings of all 18 variables, drawn with a fixed seed, returned the same seven rejections shown in Figure 3(a) in every replication; the statistics of pairs involving variables with five or more categories varied, for instance from .278 to .346 for class and odor. The primary encoding was fixed before the search and was not changed.

Two further checks confirm the reported decisions. An independent implementation that uses only dense Hadamard contractions and the Hoeffding envelope, without the optimized bound, rejects exactly the same seven comparisons shown in Figure 3(a); its largest adjusted bound among them is .0397, for odor and the upper stalk surface. The affine compression agrees exactly with the dense transform on ten random tables, including the recovered maximizing labels.

O.8 Reproducibility

The analysis archive contains the retained records and category dictionary, the C++ and Python engines, the calibration module with its derivation, all 385,101 scan statistics with their maximizing labels, the 63 calibrated candidate statements, the full sixteen-variable results, the count tables behind the interpretations, the coding-sensitivity draws, the audit outputs, and the figure and table sources. Package versions and the input hash are recorded. The scan is deterministic; the only random elements are the coding-sensitivity seed and the Monte Carlo checks of the bounds.