Submitted to the Annals of Statistics

OPTIMAL NONPARAMETRIC TESTING OF MISSING COMPLETELY AT RANDOM, AND ITS CONNECTIONS TO COMPATIBILITY

BY THOMAS B. BERRETT[1,*] AND RICHARD J. SAMWORTH[2,†]

[1]*Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom*
tom.berrett@warwick.ac.uk

[2]*Statistical Laboratory, Centre for Mathematical Sciences, Wilberforce Road, Cambridge, CB3 0WB, United Kingdom*
r.samworth@statslab.cam.ac.uk

Given a set of incomplete observations, we study the nonparametric problem of testing whether data are Missing Completely At Random (MCAR). Our first contribution is to characterise precisely the set of alternatives that can be distinguished from the MCAR null hypothesis. This reveals interesting and novel links to the theory of Fréchet classes (in particular, compatible distributions) and linear programming, that allow us to propose MCAR tests that are consistent against all detectable alternatives. We define an incompatibility index as a natural measure of ease of detectability, establish its key properties, and show how it can be computed exactly in some cases and bounded in others. Moreover, we prove that our tests can attain the minimax separation rate according to this measure, up to logarithmic factors. Our methodology does not require any complete cases to be effective, and is available in the `R` package `MCARtest`.

1. Introduction. Over the last century, a plethora of algorithms have been proposed to address specific statistical challenges; in many cases these can be justified under modelling assumptions on the underlying data generating mechanism. When faced with a data set and a question of interest, the practitioner needs to assess the validity of the assumptions underpinning these statistical models, in order to determine whether or not they can trust the output of the method. Experienced practitioners recognise that mathematical assumptions can rarely be expected to hold exactly, and develop intuition (sometimes backed up by formal tests) about the seriousness of different violations.

One of the most commonly-encountered discrepancies between real data sets and models hypothesised in theoretical work is that of missing data. In fact, missingness may be even more serious than many other types of departure from a statistical model, in that it may be impossible even to run a particular algorithm without modification when data are missing. Once it is accepted that methods for dealing with missing data are essential, the primary concern is to understand the relationship between the data generating and missingness mechanisms. In the ideal situation, these two sources of randomness are independent, a setting known as Missing Completely At Random (MCAR). When this assumption holds, the analysis becomes much easier, because we can regard our observed data as a representative sample from the wider population. For instance, theoretical guarantees have recently been established in the MCAR setting for a variety of modern statistical problems, including high-dimensional regression (Loh and Wainwright, 2012), high-dimensional or sparse principal component analysis (Zhu, Wang and Samworth, 2022; Elsener and van de Geer, 2019), classification (Cai and Zhang, 2019), and precision matrix and changepoint estimation (Loh

*Research supported by Engineering and Physical Sciences Reseach Council (EPSRC) New Investigator Award EP/W016117/1.

†Research supported by EPSRC Programme grant EP/N031938/1, EPSRC Fellowship EP/P031447/1 and European Research Council Advanced grant 101019498.

and Tan, 2018; Follain, Wang and Samworth, 2022). The failure of this assumption, on the other hand, may introduce significant bias and necessitate further investigation of the nature of the dependence between the data and the missingness (Davison, 2003; Little and Rubin, 2019).

Our aim in this work is to study the fundamental problem of testing the null hypothesis that data are MCAR. It is important to recognise from the outset that in general there will exist alternatives (i.e. joint distributions of data and missingness that do not satisfy the MCAR hypothesis) for which no test could have power greater than its size. Indeed, to give a toy example, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0,1)$, but we only observe those X_i that are non-negative, then the joint distribution of our data is indistinguishable from the MCAR setting where $X_1, \ldots, X_n$ are a random sample from the folded normal distribution on $[0, \infty)$, and each X_i is observed independently with probability $1/2$.

The first main contribution of this work, then, is to determine precisely the set of alternatives that are distinguishable from our null hypothesis. Surprisingly, this question turns out to be relevant in several different subject areas, namely copula theory (Nelsen, 2007; Dall'Aglio, Kotz and Salinetti, 2012), portfolio risk management (Embrechts and Puccetti, 2010; Rüschendorf, 2013), coalition games (Vorobev, 1962), quantum contextuality (Bell, 1966; Clauser and Shimony, 1978) and relational databases (Maier, 1983). To describe our results briefly, we introduce the notation that when a random vector X takes values in a measurable space of the form $\mathcal{X} = \prod_{j=1}^{d} \mathcal{X}_j$ and when $S \subseteq \{1, \ldots, d\}$, we write $X_S := (X_j : j \in S)$ and $\mathcal{X}_S := \prod_{j \in S} \mathcal{X}_j$. Following, e.g., Joe (1997, Section 3), given a collection $\mathbb{S}$ of subsets of $\{1, \ldots, d\}$, and a collection of distributions $P_{\mathbb{S}} := (P_S : S \in \mathbb{S})$, we define their *Fréchet class* $\mathcal{F}(P_{\mathbb{S}})$ as the set of all distributions of X for which X_S has marginal distribution P_S for all $S \in \mathbb{S}$. In Section 2, we prove that it is only possible to detect that a joint distribution does not satisfy the MCAR hypothesis if the marginal distributions for which we have simultaneous observations are incompatible.

Our second contribution, in Section 3, is to introduce a new, universal test of the null hypothesis of compatibility, and consequently (by our result in Section 2) the MCAR hypothesis, in the discrete case. We prove that it has finite-sample Type I error control, and is consistent against all incompatible alternatives. These results therefore describe precisely what can be learnt about the plausibility of the MCAR hypothesis from data. Our methodology is based on a duality theorem due to Kellerer (1984) that gives a characterisation of compatibility, and allows us to define a notion of an *incompatibility index*, denoted $R(P_{\mathbb{S}})$. Although the result itself is rather abstract, we show how it motivates a test statistic that can be computed straightforwardly using linear programming. We further argue that a more specific and involved analysis can lead to improved tests in certain cases. For instance, when $d = 3$, with $\mathbb{S} = \big\{\{1,2\}, \{2,3\}, \{1,3\}\big\}$ and $|\mathcal{X}_1| = r$, $|\mathcal{X}_2| = s$ and $|\mathcal{X}_3| = 2$, we show by means of a minimax lower bound (Theorem 9) that our improved test achieves the optimal separation rate in $R(P_{\mathbb{S}})$ simultaneously in r, s and the sample sizes for each observation pattern, up to logarithmic factors.

The form of the incompatibility index is a supremum of a class of linear functionals, and exact expressions can become complicated as $|\mathbb{S}|$ and the alphabet sizes increase. In Section 3.4, we describe computational geometry algorithms that yield analytic expressions for $R(P_{\mathbb{S}})$; code is available in the `R` package `MCARtest` (Berrett and Samworth, 2022), and in principle, these can be applied for arbitrary $\mathbb{S}$. As illustrations, we provide examples with binary variables, where these expressions are more tractable. Moreover, as we show in Section 3.3, in some cases we can exploit the structure of $\mathbb{S}$ to reduce the computation of $R(P_{\mathbb{S}})$ to the computation of the analogous quantity for lower-dimensional settings, or at least to bound it in terms of these simpler quantities.

In Section 4, we explain how the methodology and theory described above extends to continuous data, or to variables having both continuous and discrete components. Here we have an additional approximation error in the minimax separation radius that depends on the smoothness of the densities of the continuous components. Section 5 is devoted to a theoretical and numerical study of a Monte Carlo version of our test, which uses bootstrap samples to generate the critical value, and which has similar guarantees to our universal test. Empirically, we find that this version also provides good Type I error control, and outperforms a test due to Fuchs (1982) even when this latter test is provided with additional complete cases (which are required for its application). Proofs of all of our results, as well as auxiliary results (which are prefaced with an 'S'), are deferred to the supplementary material (Berrett and Samworth, 2023).

Our theory is based on the study of marginal polytopes, which is a topical problem in convex geometry (Vlach, 1986; Wainwright and Jordan, 2008; Deza and Laurent, 2009). Indeed, these polytopes are known to be extremely complicated (De Loera and Kim, 2014), but are of considerable interest in hierarchical log-linear models (Eriksson et al., 2006), variational inference (Wainwright and Jordan, 2003), classical transportation (Kantorovich, 1942) (reprinted as Kantorovich (2006)) and max flow-min cut problems (Gale, 1957). In the special case where all variables are binary, marginal polytopes are equivalent to *correlation polytopes* or *cut polytopes*, which have been heavily studied in their own right (Deza and Laurent, 2009; Coons et al., 2020). In statistical contexts, recent work on hypothesis testing over polyhedral parameter spaces has sought to elucidate the link between the difficulty of the problem and the underlying geometry (Blanchard, Carpentier and Gutzeit, 2018; Wei, Wainwright and Guntuboyina, 2019).

Most prior work on testing the MCAR hypothesis has been developed within the context of parametric models such as multivariate normality (Little, 1988; Kim and Bentler, 2002; Jamshidian and Jalal, 2010), Poisson or multinomial contingency tables with at least some complete cases (Fuchs, 1982) or generalised estimating equations (Chen and Little, 1999; Qu and Song, 2002). Li and Yu (2015) study the nonparametric problem of testing whether or not a family of marginal distributions $P_{\mathbb{S}}$ is *consistent*, i.e. whether, for each $S, S' \in \mathbb{S}$ with $S \cap S' \neq \emptyset$, the marginal distributions of P_S and $P_{S'}$ on the coordinates in $S \cap S'$ agree with each other. Spohn et al. (2021) consider an equivalent problem, using random forest classification methods to test equalities of distributions. Consistency is a necessary, but not sufficient, condition for compatibility*. To the best of our knowledge, the current paper is the first both to characterise the set of detectable alternatives to the MCAR hypothesis, and to provide tests that have asymptotic power 1 against all such detectable alternatives while controlling the Type I error.

We conclude this introduction with some notation that is used throughout the paper. For $d \in \mathbb{N}$, we write $[d] := \{1, \ldots, d\}$, and also define $[\infty] := \mathbb{N}$. Given a countable set Ω, we write 2^{Ω} for its power set, and 1_{Ω} for the vector of ones indexed by the elements of Ω. If $S \subseteq [d]$, we denote $\mathbb{1}_S := (\mathbb{1}_{\{j \in S\}})_{j \in [d]} \in \{0,1\}^d$. For $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, write $x_S = (x_j : j \in S)$. For $x \in \mathbb{R}$, let $x_+ := \max(x, 0)$ and $x_- := \max(-x, 0)$. Given $a, b \geq 0$, we write $a \lesssim b$ to mean that there exists a universal constant $C > 0$ such that $a \leq Cb$, and, for a generic quantity x, write $a \lesssim_x b$ to mean that there exists C, depending only on x, such that $a \leq Cb$. We also write $a \asymp b$ to mean $a \lesssim b$ and $b \lesssim a$. For random elements X, Y, we write $X \perp\!\!\!\perp Y$

*However, in the special case where $[d] \in \mathbb{S}$, a necessary and sufficient condition for compatibility is that P_S is the marginal distribution on $\mathcal{X}_S$ of $P_{[d]}$, for each $S \in \mathbb{S} \setminus \{[d]\}$. In other words, in this case, consistency is sufficient for compatibility. A test of compatibility may therefore then be constructed by testing each of these hypotheses via $|\mathbb{S}| - 1$ two-sample tests and applying, e.g., a Bonferroni correction. More generally, this strategy may be applied whenever $\mathbb{S}$ is *decomposable* (Lauritzen, Speed and Vijayan, 1984; Lauritzen and Spiegelhalter, 1988).

Symbol	Type of mathematical object	Interpretation
$\mathbb{S}$	Subset of $2^{[d]}$	Set of possible observation patterns
$\mathcal{X}_1, \ldots, \mathcal{X}_d$	Measurable spaces	Component spaces for $X_1, \ldots, X_d$
$\mathcal{X}$	Product space $\prod_{j=1}^d \mathcal{X}_j$	Product space for $X = (X_1, \ldots, X_d)$
Ω	Random vector in $\{0,1\}^d$	Revelation vector
$X \circ \Omega$	Random vector in $\prod_{j=1}^d (\mathcal{X}_j \cup \{*\})$	Observed data
$\mathcal{X}_S$	$\prod_{j \in S} \mathcal{X}_j$	Partial product space for $X_S = (X_j : j \in S)$
$\mathcal{X}_{\mathbb{S}}$	$(\mathcal{X}_S : S \in \mathbb{S})$	Sequence of partial product spaces
$\mathcal{P}$	Set of distributions on $\mathcal{X}$	Possible distributions for X
$\mathcal{P}_S$	Set of distributions on $\mathcal{X}_S$	Possible distributions for X_S
$\mathcal{P}_{\mathbb{S}}$	$(\mathcal{P}_S : S \in \mathbb{S})$	Set of sequences of possible observed distributions
$\mathbb{A}$	Function from $\mathcal{P}$ to $\mathcal{P}_{\mathbb{S}}$	Marginalisation operator
$\mathcal{P}_{\mathbb{S}}^0$	Image of $\mathbb{A}$	Set of compatible sequences of distributions
$\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}})$	$\mathbb{A}^{-1}(P_{\mathbb{S}})$	Fréchet class
$R(P_{\mathbb{S}}, f_{\mathbb{S}})$	Element of $(-\infty, 1]$	Linear functional of $P_{\mathbb{S}}$
$\mathcal{G}_{\mathbb{S}}^+$	Set of functions from $\mathcal{X}_{\mathbb{S}}$ to $[-1,\infty)^{\lvert\mathbb{S}\rvert}$	Set of feasible $f_{\mathbb{S}}$
$R(P_{\mathbb{S}})$	$\sup_{f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^+} R(P_{\mathbb{S}}, f_{\mathbb{S}}) \in [0,1]$	Incompatibility index

TABLE 1
Key notation in the paper.

to mean that X and Y are independent. For probability measures P, Q on a measurable space $(\mathcal{Z}, \mathcal{C})$, we denote their total variation distance as $d_{\mathrm{TV}}(P,Q) := \sup_{C \in \mathcal{C}} |P(C) - Q(C)|$. For the reader's convenience, we include a table of key notation as Table 1.

2. Fréchet classes and non-detectable alternatives. We begin with a brief discussion of Fréchet classes, for which a good reference is Joe (1997, Section 3), as this will allow us to characterise the set of detectable alternatives of an MCAR test. Throughout the paper, for $d \in \mathbb{N}$ and measurable topological spaces $(\mathcal{X}_1, \mathcal{A}_1), \ldots, (\mathcal{X}_d, \mathcal{A}_d)$, we let $\mathcal{X} := \prod_{j=1}^d \mathcal{X}_j$. Given a collection $\mathbb{S}$ of subsets of $[d]$ and a set of distributions $P_{\mathbb{S}} = (P_S : S \in \mathbb{S})$, where P_S is defined on $\mathcal{X}_S$, we write $\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}})$ for the corresponding Fréchet class. As a simple example, if $\mathbb{S} = \big\{\{1\}, \ldots, \{d\}\big\}$, then $\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}})$ is the class of all joint distributions with specified marginals $P_{\{1\}}, \ldots P_{\{d\}}$. It is easy to see that this Fréchet class in non-empty, because it includes the product distribution $P_{\{1\}} \times \ldots \times P_{\{d\}}$. More generally, if $S_1, \ldots, S_m$ is a partition of $[d]$ and $\mathbb{S} = \{S_1, \ldots, S_m\}$, then $\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}})$ contains the corresponding product distribution. However, when $\mathbb{S}$ contains subsets that overlap, the Fréchet class $\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}})$ may be empty, or equivalently, $P_{\mathbb{S}}$ may be incompatible. One simple way in which this may occur is if $d = 2$ and $\mathbb{S} = \big\{\{1\}, \{1,2\}\big\}$, but $P_{\{1\}}$ and $P_{\{1,2\}}$ are not consistent. More interestingly, when $d \geq 3$ it may be the case that $P_{\mathbb{S}}$ is consistent but we still have $\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}}) = \emptyset$. For instance when $d = 3$ and $\mathbb{S} = \big\{\{1,2\}, \{1,3\}, \{2,3\}\big\}$, let $\rho_{23} = \rho_{13} = 2^{-1/2}$, let $\rho_{12} = -2^{-1/2}$ and, for $1 \leq i < j \leq 3$, let

$$P_{\{i,j\}} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{pmatrix}\right).$$

Then any joint distribution $P_{\{1,2,3\}}$ with these marginals would have 'covariance matrix'

$$\begin{pmatrix} 1 & -2^{-1/2} & 2^{-1/2} \\ -2^{-1/2} & 1 & 2^{-1/2} \\ 2^{-1/2} & 2^{-1/2} & 1 \end{pmatrix},$$

which has a negative eigenvalue.

We are now in a position to describe the main statistical question that motivates our work. Given $x = (x_1, \ldots, x_d) \in \mathcal{X}$ and $\omega = (\omega_1, \ldots, \omega_d) \in \{0,1\}^d$, we write $x \circ \omega$ for the element

of $\prod_{j=1}^d(\mathcal{X}_j \cup \{\star\})$ that has jth entry x_j if $\omega_j = 1$ and jth entry $\star$, denoting a missing value, if $\omega_j = 0$. Assume that we are given n independent copies of $X \circ \Omega$, where the pair (X, Ω) takes values in $\mathcal{X} \times \{0,1\}^d$, and wish to test the hypothesis $H_0 : X \perp\!\!\!\perp \Omega$, i.e. that entries of X are MCAR. This can be thought of as an independence test where we do not have complete observations, though we will see that the missingness leads to very different phenomena.

Let $\mathbb{S} := \big\{S \subseteq [d] : \mathbb{P}(\Omega = \mathbb{1}_S) > 0\big\}$ denote the set of all missingness patterns that could be observed. Writing P_S for the conditional distribution of X_S given that $\Omega = \mathbb{1}_S$, note that if our data are MCAR, then $P_{\mathbb{S}} := (P_S : S \in \mathbb{S})$ is compatible, because $X_S \overset{\mathrm{d}}{=} X_S|\{\Omega = \mathbb{1}_S\} \sim P_S$, so the Fréchet class $\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}})$ contains the distribution of X.

On the other hand, suppose now that our data are not MCAR, but that $P_{\mathbb{S}}$ is still compatible. If $\tilde{X}$ denotes a random vector, independent of Ω, whose distribution lies in the Fréchet class $\mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}})$, then $\tilde{X} \circ \Omega \overset{\mathrm{d}}{=} X \circ \Omega$, so no test of H_0 can have power at compatible alternatives that is greater than its size. The conclusion of this discussion is stated in Proposition 1 below, where we let Ψ denote the set of all (randomised) tests based on our observed data $X_1 \circ \Omega_1, \dots, X_n \circ \Omega_n$, i.e. the set of Borel measurable functions $\psi : \big(\prod_{j=1}^d(\mathcal{X}_j \cup \{\star\})\big)^n \to [0,1]$.

PROPOSITION 1. *Let $\mathcal{P}_0$ denote the set of distributions on $\mathcal{X} \times \{0,1\}^d$ that satisfy H_0, and let $\mathcal{P}_0'$ denote the set of distributions on $\mathcal{X} \times \{0,1\}^d$ for which the corresponding sequence of conditional distributions $P_{\mathbb{S}}$ is compatible. Then $\mathcal{P}_0 \subseteq \mathcal{P}_0'$, but for any $\psi \in \Psi$, we have*

$$\sup_{P \in \mathcal{P}_0'} \mathbb{E}_P \psi(X_1 \circ \Omega_1, \dots, X_n \circ \Omega_n) = \sup_{P \in \mathcal{P}_0} \mathbb{E}_P \psi(X_1 \circ \Omega_1, \dots, X_n \circ \Omega_n).$$

A consequence of Proposition 1 is that it is only possible to have non-trivial power against incompatible alternatives to H_0, and a search for optimal tests of the MCAR property may be reduced to looking for optimal tests of compatibility. In subsequent sections, we will construct tests of compatibility, noting that if such a test rejects the null hypothesis, then we can also reject the hypothesis of MCAR.

3. Testing compatibility. Let $\mathcal{P}_{\mathbb{S}}$ denote the set of sequences of the form $P_{\mathbb{S}} = (P_S : S \in \mathbb{S})$, where P_S is a distribution on $\mathcal{X}_S$, and let $\mathcal{P}_{\mathbb{S}}^0$ denote the subset of $\mathcal{P}_{\mathbb{S}}$ consisting of those $P_{\mathbb{S}}$ that are compatible. In testing compatibility, it is convenient to alter our model very slightly, so that we have deterministic sample sizes within each observation pattern. More precisely, given a collection $\mathbb{S} \subseteq 2^{[d]}$ and $P_{\mathbb{S}} = (P_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}$, we assume that we are given independent data $(X_{S,i})_{S \in \mathbb{S}, i \in [n_S]}$, where $X_{S,1}, \dots, X_{S,n_S} \overset{\mathrm{iid}}{\sim} P_S$ for each $S \in \mathbb{S}$. Our goal is to propose a test of $H_0' : P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0$, or equivalently, $H_0' : \mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}}) \neq \emptyset$. To this end, for $S \in \mathbb{S}$, let $\mathcal{G}_S^*$ denote the set of all bounded, upper semi-continuous functions on $\mathcal{X}_S$. We will exploit the characterisation of Kellerer (1984, Proposition 3.13), which states that $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0$ if and only if

(1)
$$\sum_{S \in \mathbb{S}} \int_{\mathcal{X}_S} f_S(x_S)\, dP_S(x_S) \geq 0 \quad \text{for all } (f_S : S \in \mathbb{S}) \in \prod_{S \in \mathbb{S}} \mathcal{G}_S^* \text{ with } \inf_{x \in \mathcal{X}} \sum_{S \in \mathbb{S}} f_S(x_S) \geq 0.$$

This duality theorem can be regarded as a potentially infinite-dimensional generalisation of Farkas's lemma (Farkas, 1902), which underpins the theory of linear programming.

We now show how (1) can be used to define a quantitative measure of incompatibility. For $S \in \mathbb{S}$, let $\mathcal{G}_S$ denote the subset of $\mathcal{G}_S^*$ consisting of functions taking values in $[-1, \infty)$, and let $\mathcal{G}_{\mathbb{S}} := \prod_{S \in \mathbb{S}} \mathcal{G}_S$. Given $f_S \in \mathcal{G}_S$ for each $S \in \mathbb{S}$, we write $f_{\mathbb{S}} := (f_S : S \in \mathbb{S}) \in \mathcal{G}_{\mathbb{S}}$. Now let

$$\mathcal{G}_{\mathbb{S}}^+ := \bigg\{ f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}} : \inf_{x \in \mathcal{X}} \sum_{S \in \mathbb{S}} f_S(x_S) \geq 0 \bigg\}.$$

Our key *incompatibility index*, then, is

$$R(P_{\mathbb{S}}) := \sup_{f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^{+}} R(P_{\mathbb{S}}, f_{\mathbb{S}}), \tag{2}$$

where

$$R(P_{\mathbb{S}}, f_{\mathbb{S}}) := -\frac{1}{|\mathbb{S}|} \sum_{S \in \mathbb{S}} \int_{\mathcal{X}_S} f_S(x_S)\, dP_S(x_S).$$

Since the choice $f_S \equiv 0$ for all $S \in \mathbb{S}$ means that the corresponding $f_{\mathbb{S}}$ belongs to $\mathcal{G}_{\mathbb{S}}^{+}$, we see that $R(P_{\mathbb{S}}) \geq 0$, and from (1), $R(P_{\mathbb{S}}) = 0$ if $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{0}$. Moreover, if (1) is violated by some $f_{\mathbb{S}} \in \mathcal{G}_{S}^{*}$ with $\inf_{x \in \mathcal{X}} \sum_{S \in \mathbb{S}} f_S(x_S) \geq 0$, then by scaling we may assume that $f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^{+}$, and hence $R(P_{\mathbb{S}}) > 0$ whenever $P_{\mathbb{S}} \notin \mathcal{P}_{\mathbb{S}}^{0}$. Finally, observe that we also have $R(P_{\mathbb{S}}) \leq 1$ for all $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$; the extreme case $R(P_{\mathbb{S}}) = 1$ corresponds to *strongly contextual* families of distributions, in the terminology of quantum contextuality (Abramsky and Brandenburger, 2011). When $|\mathcal{X}| < \infty$, we see from Theorem 2 below that $R(P_{\mathbb{S}}) < 1$ if and only if there exists $x \in \mathcal{X}$ with $P_S(\{x_S\}) > 0$ for all $S \in \mathbb{S}$.

THEOREM 2. *Suppose that $\mathcal{X}_j$ is a locally compact Hausdorff space*[†]*, for each $j \in [d]$, and that every open set in $\mathcal{X}$ is σ-compact. Then for any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$,*

(3)
$$R(P_{\mathbb{S}}) = \inf\big\{\epsilon \in [0,1] : P_{\mathbb{S}} \in (1-\epsilon)\mathcal{P}_{\mathbb{S}}^{0} + \epsilon \mathcal{P}_{\mathbb{S}}\big\} = 1 - \sup\big\{\epsilon \in [0,1] : P_{\mathbb{S}} \in \epsilon \mathcal{P}_{\mathbb{S}}^{0} + (1-\epsilon)\mathcal{P}_{\mathbb{S}}\big\}.$$

Remark: If $\mathcal{X}$ is second countable, then every open set in $\mathcal{X}$ is σ-compact.

Theorem 2 can be regarded as providing a dual representation for $R(P_{\mathbb{S}})$. In the quantum physics literature and for consistent families of distributions on discrete spaces, the second and third expressions in (3) are known as the *contextual fraction* (Abramsky, Barbosa and Mansfield, 2017). The first step of the proof of Theorem 2 is to apply the idea of *Alexandroff (one-point) compactification* (Alexandroff, 1924) to reduce the problem to compact Hausdorff spaces. Strong duality for linear programming (Isii, 1964, Theorem 2.3), combined with the Riesz representation theorem for positive linear functionals on the set of continuous functions on compact spaces, then allows us to deduce the result.

Another important property of our incompatibility index is the fact that it is 1-Lipschitz with respect to the total variation distance

$$d_{\mathrm{TV}}\big((P_S : S \in \mathbb{S}), (Q_S : S \in \mathbb{S})\big) := \sum_{S \in \mathbb{S}} d_{\mathrm{TV}}(P_S, Q_S)$$

on $\mathcal{P}_{\mathbb{S}}$.

PROPOSITION 3. *For any $P_{\mathbb{S}} = (P_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}$ and $Q_{\mathbb{S}} = (Q_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}$, we have*

$$|R(P_{\mathbb{S}}) - R(Q_{\mathbb{S}})| \leq d_{\mathrm{TV}}(P_{\mathbb{S}}, Q_{\mathbb{S}}).$$

With the basic properties of our incompatibility index now in place, we can now introduce the minimax framework for our hypothesis testing problem. Writing $n_{\mathbb{S}} := (n_S : S \in \mathbb{S}) \in \mathbb{N}^{\mathbb{S}}$, a test of H_0' is a measurable function $\psi'_{n_{\mathbb{S}}} : \prod_{S \in \mathbb{S}} \mathcal{X}_S^{n_S} \to [0,1]$, and we write $\Psi'_{n_{\mathbb{S}}}$ for the set of all such tests. Given $\rho \geq 0$, it is convenient to write

$$\mathcal{P}_{\mathbb{S}}(\rho) := \{P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}} : R(P_{\mathbb{S}}) \geq \rho\},$$

[†]A brief glossary of definitions of topological and measure-theoretic concepts used in this result and its proof is provided in Section S2 for the reader's convenience.

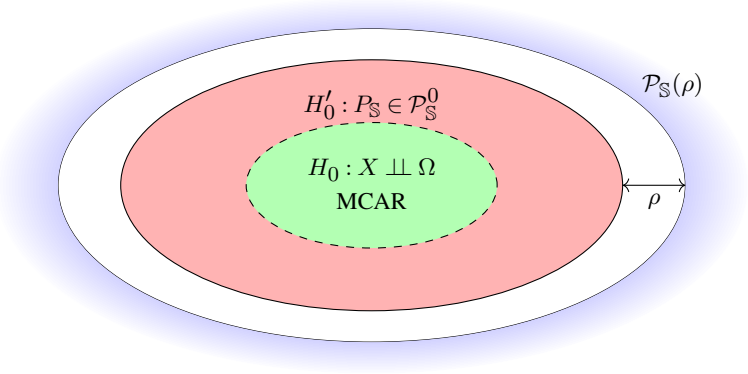

Fig 1: Illustration of our minimax testing framework.

so that $\mathcal{P}_{\mathbb{S}}(0) = \mathcal{P}_{\mathbb{S}}$, $\mathcal{P}_{\mathbb{S}}^0 = \mathcal{P}_{\mathbb{S}} \setminus \cup_{\epsilon>0}\mathcal{P}_{\mathbb{S}}(\epsilon)$ and $\mathcal{P}_{\mathbb{S}}(\epsilon) = \emptyset$ for $\epsilon > 1$. The minimax risk at separation ρ in this problem is defined as

$$\mathcal{R}(n_{\mathbb{S}}, \rho) := \inf_{\psi'_{n_{\mathbb{S}}} \in \Psi'_{n_{\mathbb{S}}}} \Big\{ \sup_{P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0} \mathbb{E}_{P_{\mathbb{S}}}(\psi'_{n_{\mathbb{S}}}) + \sup_{P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}(\rho)} \mathbb{E}_{P_{\mathbb{S}}}(1 - \psi'_{n_{\mathbb{S}}}) \Big\};$$

thus $\mathcal{R}(n_{\mathbb{S}}, \rho) = 0$ for $\rho > 1$. Finally, the minimax testing radius is defined as

$$\rho^*(n_{\mathbb{S}}) := \inf\big\{\rho \geq 0 : \mathcal{R}(n_{\mathbb{S}}, \rho) \leq 1/2\big\},$$

so that $\rho^*(n_{\mathbb{S}}) \leq 1$. This framework is illustrated in Figure 1.

3.1. *A universal test in the discrete case.* In this subsection, we will assume that $\mathcal{X}_j = [m_j]$ for every $j \in [d]$, where $m_j \in \mathbb{N}$. Given our data, for each $S \in \mathbb{S}$ and $A_S \in 2^{\mathcal{X}_S}$, define the empirical distribution of $(X_{S,i})_{i\in[n_S]}$ by

$$\widehat{P}_S(A_S) := \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbb{1}_{\{X_{S,i} \in A_S\}}$$

and write $\widehat{P}_{\mathbb{S}} := (\widehat{P}_S : S \in \mathbb{S})$. We propose to reject H_0' at the significance level $\alpha \in (0,1)$ if $R(\widehat{P}_{\mathbb{S}}) \geq C_\alpha$, where

$$C_\alpha := \frac{1}{2} \sum_{S\in\mathbb{S}} \Big(\frac{|\mathcal{X}_S| - 1}{n_S} \Big)^{1/2} + \Big\{ \frac{1}{2} \log(1/\alpha) \sum_{S\in\mathbb{S}} \frac{1}{n_S} \Big\}^{1/2}.$$

The following proposition provides size and power guarantees for this test.

PROPOSITION 4. *Fix $\alpha, \beta \in (0,1)$. Whenever $P_{\mathbb{S}} = (P_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}^0$, we have $\mathbb{P}_{P_{\mathbb{S}}}\big(R(\widehat{P}_{\mathbb{S}}) \geq C_\alpha\big) \leq \alpha$. Moreover, for any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$ satisfying*

$$R(P_{\mathbb{S}}) \geq C_\alpha + C_\beta,$$

we have $\mathbb{P}_{P_{\mathbb{S}}}\big(R(\widehat{P}_{\mathbb{S}}) \geq C_\alpha\big) \geq 1 - \beta$.

Proposition 4 reveals in particular that in addition to having guaranteed finite-sample size control, our test is consistent against any fixed, incompatible alternative; in other words, whenever $R(P_{\mathbb{S}}) > 0$, we have $\mathbb{P}_{P_{\mathbb{S}}}\big(R(\widehat{P}_{\mathbb{S}}) \geq C_\alpha\big) \to 1$ as $\min_{S\in\mathbb{S}} n_S \to \infty$. In combination with Proposition 1, then, we see that from a testing perspective, compatibility is the right

proxy for MCAR, in that distributions of (X, Ω) that do not satisfy the MCAR hypothesis are detectable if and only if their observed margins are incompatible. Moreover, we have the following upper bound on the minimax separation rate:

$$\rho^*(n_{\mathbb{S}}) \leq \sum_{S \in \mathbb{S}} \Big(\frac{|\mathcal{X}_S| - 1}{n_S} \Big)^{1/2} + 2 \Big(\log 2 \sum_{S \in \mathbb{S}} \frac{1}{n_S} \Big)^{1/2} \lesssim_{|\mathbb{S}|} \max_{S \in \mathbb{S}} \Big(\frac{|\mathcal{X}_S|}{n_S} \Big)^{1/2}.$$

As far as computation of the test statistic is concerned, observe that, writing $\mathcal{X}_{\mathbb{S}} := \{(S, x_S) : S \in \mathbb{S}, x_S \in \mathcal{X}_S\}$, we can identify $\mathcal{G}_{\mathbb{S}}$ with $[-1, \infty)^{\mathcal{X}_{\mathbb{S}}}$, and $\mathcal{G}_{\mathbb{S}}^+$ with a convex polyhedral subset of $[-1, \infty)^{\mathcal{X}_{\mathbb{S}}}$. Moreover, any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$ can be identified with an element of $[0, 1]^{\mathcal{X}_{\mathbb{S}}}$. We will show in Proposition 6 below that the supremum in (2) is attained. In fact, $R(P_{\mathbb{S}}, \cdot)$ is linear, so we can compute $R(\widehat{P}_{\mathbb{S}})$ using efficient linear programming algorithms.

3.2. *An improved test under additional information.* In this subsection, we show how in the discrete setting of Section 3.1, it may be possible to reduce the critical value of our test, while retaining finite-sample Type I error control, when certain information about the facet structure of relevant polytopes is available. This information does not depend on any quantities that are unknown to the practitioner, though exact computation may be a challenge when $|\mathbb{S}|$ or the alphabet sizes are large.

Before we can describe our improved test, it is helpful to study the geometric structure of the problem further. Regarding $\mathcal{G}_{\mathbb{S}}^+$ as a polyhedral convex subset of $[-1, \infty)^{\mathcal{X}_{\mathbb{S}}}$, it has a finite number of extreme points, so $\sup_{f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^+} R(P_{\mathbb{S}}, f_{\mathbb{S}}) = \max_{\ell \in [L]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)})$ for some $f_{\mathbb{S}}^{(1)}, \ldots, f_{\mathbb{S}}^{(L)} \in \mathcal{G}_{\mathbb{S}}^+$. Thus $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0$ if and only if $\max_{\ell \in [L]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)}) \leq 0$, and $\mathcal{P}_{\mathbb{S}}^0$ can be identified with a finite intersection of halfspaces, i.e. it can be identified with a convex polyhedron in $[0, 1]^{\mathcal{X}_{\mathbb{S}}}$. Now define the *marginal cone*[‡] $\mathcal{P}_{\mathbb{S}}^{0,*} := \{\lambda \cdot \mathcal{P}_{\mathbb{S}}^0 : \lambda \geq 0\}$. From the discussion above, $\mathcal{P}_{\mathbb{S}}^{0,*}$ can be identified with all non-negative multiples of a convex polyhedron, so can itself be identified with a convex polyhedral cone in $[0, \infty)^{\mathcal{X}_{\mathbb{S}}}$.

When $\emptyset \neq S_2 \subseteq S_1 \subseteq [d]$ and P_{S_1} is a measure on $\mathcal{X}_{S_1}$, we write $P_{S_1}^{S_2}$ for the marginal measure of P_{S_1} on $\mathcal{X}_{S_2}$. Recall that a family $P_{\mathbb{S}} = (P_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}$ is *consistent* if, whenever $S_1, S_2 \in \mathbb{S}$ have $S_1 \cap S_2 \neq \emptyset$, we have $P_{S_1}^{S_1 \cap S_2} = P_{S_2}^{S_1 \cap S_2}$. We let $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons}} \subseteq \mathcal{P}_{\mathbb{S}}$ denote the set of consistent families of distributions on $\mathcal{X}_{\mathbb{S}}$, with corresponding *consistent cone* $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons},*} := \{\lambda \cdot \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}} : \lambda \geq 0\}$ and *consistent ball* $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**} := \{\lambda \cdot \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}} : \lambda \in [0, 1]\}$. Thinking of $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$ as a convex polytope in $[0, \infty)^{\mathcal{X}_{\mathbb{S}}}$, the Minkowski sum $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ is also a convex polyhedral set, so has a finite number of facets (Rockafellar, 1997, Theorem 19.1). These facets fall into two categories: those that define the non-negativity conditions (i.e. $(P_{\mathbb{S}})_{(S,x_S)} = P_S(\{x_S\}) \geq 0$ for all $S \in \mathbb{S}$ and $x_S \in \mathcal{X}_S$), which are not of primary interest to us here, and the remainder, which we refer to as the set of *essential* facets. We remark that, in decomposable settings where $\mathcal{P}_{\mathbb{S}}^0 = \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$, there are no essential facets. More generally, regardless of whether $\mathbb{S}$ is decomposable, we still have the following:

PROPOSITION 5. *$\mathcal{P}_{\mathbb{S}}^0$ is a full-dimensional subset of $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$.*

In addition to the geometric insight of Proposition 5, it is also interesting from a statistical perspective when we consider testing compatibility against consistent alternatives (which captures the main essence of the problem in many examples; see the discussion at the end

[‡]Here, given $\lambda > 0$ and a distribution P on a measurable space $(\mathcal{Z}, \mathcal{C})$, the measure $\lambda \cdot P$ is defined in the obvious way by $(\lambda \cdot P)(C) := \lambda \cdot P(C)$ for $C \in \mathcal{C}$; likewise, for a family of distributions $\mathcal{P}$, we write $\lambda \cdot \mathcal{P} := \{\lambda \cdot P : P \in \mathcal{P}\}$.

of Section 3.2). It reveals a distinction with standard, fully-observed hypothesis testing problems (e.g. goodness-of-fit testing, two-sample testing, independence testing), where the null hypothesis parameter space is of lower dimension than that of the alternative hypothesis parameter space (e.g., Fienberg, 1968).

We are now in a position to present Proposition 6, whose main (second) part provides a decomposition of the incompatibility index.

PROPOSITION 6. *In the discrete setting above, the supremum in (2) and the infimum in (3) are attained. Moreover, writing F for the number of essential facets of $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$, there exist $f_{\mathbb{S}}^{(1)}, \dots, f_{\mathbb{S}}^{(F)} \in \mathcal{G}_{\mathbb{S}}^{+}$, depending only on $\mathbb{S}$ and $\mathcal{X}_{\mathbb{S}}$, such that for any $P_{\mathbb{S}} = (P_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}$, we have*

$$\max\Big\{\max_{\ell\in[F]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)})_+, \frac{1}{|\mathbb{S}|}\max_{S_1,S_2\in\mathbb{S}} d_{\mathrm{TV}}\big(P_{S_1}^{S_1\cap S_2}, P_{S_2}^{S_1\cap S_2}\big)\Big\} \le R(P_{\mathbb{S}})$$

$$\le \max_{\ell\in[F]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)})_+ + |\mathbb{S}|2^{|\mathbb{S}|+2}\cdot \max_{S_1,S_2\in\mathbb{S}} d_{\mathrm{TV}}\big(P_{S_1}^{S_1\cap S_2}, P_{S_2}^{S_1\cap S_2}\big), \tag{4}$$

where we interpret $\max_{\ell\in[0]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)})_+ = 0$.

Proposition 6 shows in particular that when $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$, the number of essential facets of $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ governs the complexity of the incompatibility index, and we can write $R(P_{\mathbb{S}})$ in irreducible form as

$$R(P_{\mathbb{S}}) = \max_{\ell\in[F]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)})_+.$$

For general $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$, Proposition 6 shows that $R(P_{\mathbb{S}})$ can be expressed as a maximum of this irreducible part and (up to a multiplicative factor depending only on $|\mathbb{S}|$) a total variation measure of inconsistency that quantifies the distance of $P_{\mathbb{S}}$ from $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$. As we will see below, the ideal situation is where we have knowledge of F, and we can then exploit this in the construction of powerful tests. For instance, when $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ and $\mathcal{X}_1 = [r]$, $\mathcal{X}_2 = [s]$ and $\mathcal{X}_3 = [2]$, we have $F = (2^r - 2)(2^s - 2)$; cf. Theorem 8 and the subsequent discussion. In more complicated examples, such knowledge may not be readily available, but we will also see, e.g. in Proposition 12 below, that it is nevertheless often possible to find bounds of the form

$$\max_{\ell\in[F']} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell),'})_+ \le R(P_{\mathbb{S}}) \le D_R \max_{\ell\in[F']} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell),'})_+ \tag{5}$$

for some known $D_R > 0$, $F' \in \mathbb{N}_0$, $f_{\mathbb{S}}^{(1),'}, \dots, f_{\mathbb{S}}^{(F'),'} \in \mathcal{G}_{\mathbb{S}}^{+} \cap [-1, |\mathbb{S}| - 1]^{\mathcal{X}_{\mathbb{S}}}$ and for all $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$. It then follows from (S11) in the proof of Proposition 6 that, in the upper bound in (4), we may replace $\max_{\ell\in[F]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)})_+$ by $D_R \max_{\ell\in[F']} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell),'})_+$.

Our alternative test rejects $H_0' : P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0$ at the significance level $\alpha \in (0,1)$ if and only if $R(\widehat{P}_{\mathbb{S}}) \ge C_\alpha' \equiv C_\alpha'\big(|\mathcal{X}_1|, \dots, |\mathcal{X}_d|, \mathbb{S}, (n_S : S \in \mathbb{S}), D_R, F'\big)$, where $C_\alpha' := \max(C_{\alpha,1}', C_{\alpha,2}')$, and

$$C_{\alpha,1}' := |\mathbb{S}|\Bigg\{\frac{2D_R^2 \log\big(\frac{2F'|\mathbb{S}|}{\alpha} \vee 1\big)}{\min_{S\in\mathbb{S}} n_S}\Bigg\}^{1/2},$$

$$C_{\alpha,2}' = |\mathbb{S}|\Bigg\{2^{2|\mathbb{S}|+7} \max_{\substack{S_1,S_2\in\mathbb{S}:\\ S_1\neq S_2, S_1\cap S_2\neq\emptyset}} \frac{|\mathcal{X}_{S_1\cap S_2}|\log 2 + \log\big(\frac{2|\mathbb{S}|(|\mathbb{S}|-1)}{\alpha}\big)}{n_{S_1} \wedge n_{S_2}}\Bigg\}^{1/2}.$$

Here, D_R, F' are such that (5) holds. If the number of essential facets F of $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ is known, then we may take $F' = F$ and $D_R = 1$. The following theorem provides size and power guarantees for this test.

THEOREM 7. *Fix $\alpha, \beta \in (0,1)$. If $P_{\mathbb{S}} = (P_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}^0$, then $\mathbb{P}_{P_{\mathbb{S}}}\big(R(\widehat{P}_{\mathbb{S}}) \geq C'_\alpha\big) \leq \alpha$. Moreover, there exists $M \equiv M(|\mathbb{S}|, D_R) > 0$ such that for any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$ satisfying*

$$R(P_{\mathbb{S}}) \geq M(C'_\alpha + C'_\beta), \tag{6}$$

we have $\mathbb{P}_{P_{\mathbb{S}}}\big(R(\widehat{P}_{\mathbb{S}}) \geq C'_\alpha\big) \geq 1 - \beta$.

Of course, by combining Proposition 4 and Theorem 7, we see that the test that rejects H_0' if $R(\widehat{P}_{\mathbb{S}}) \geq \min(C_\alpha, C'_\alpha) =: C_\alpha^{\min}$ remains of size α, so is an improved test that represents the best of both worlds. By taking $F' = F$ and $D_R = 1$, Proposition 4 and Theorem 7 now reveal that

$$\begin{aligned}\rho^*(n_{\mathbb{S}}) &\leq 2\min\big(MC'_{1/4}, C_{1/4}\big) \\ &\lesssim_{|\mathbb{S}|} \min\left\{\left(\frac{\log(F \vee 1)}{\min_{S\in\mathbb{S}} n_S} + \max_{\substack{S_1,S_2\in\mathbb{S}:\\ S_1\neq S_2, S_1\cap S_2\neq\emptyset}} \frac{|\mathcal{X}_{S_1\cap S_2}|}{n_{S_1} \wedge n_{S_2}}\right)^{1/2}, \max_{S\in\mathbb{S}}\left(\frac{|\mathcal{X}_S|}{n_S}\right)^{1/2}\right\}.\end{aligned}$$

By McMullen's Upper bound theorem (McMullen, 1970),

$$\log(F \vee 1) \lesssim_{|\mathbb{S}|} \log|\mathcal{X}| \cdot \max_{S\in\mathbb{S}} |\mathcal{X}_S|,$$

so that, when all sample sizes are of the same order of magnitude, we have $C'_\alpha + C'_\beta \lesssim_{|\mathbb{S}|} (C_\alpha + C_\beta) \cdot \log|\mathcal{X}|$. When tight bounds on $\log F$ are available, however, we may have that $C'_\alpha + C'_\beta$ is much smaller than $C_\alpha + C_\beta$; see the discussion following Theorem 8 below.

While these quantities are rather abstract, we can simplify them in certain cases. It is known from previous work (e.g. Vlach, 1986) that when $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ and $\mathcal{X} = [r] \times [s] \times [2]$ for some $r, s \in \mathbb{N}$, the marginal cone induced by the set of compatible measures is given by

$$\mathcal{P}_{\mathbb{S}}^{0,*} = \Big\{P_{\mathbb{S}} \equiv p_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},*} : \max_{A\subseteq[r], B\subseteq[s]} (-p_{AB\bullet} + p_{A\bullet 1} + p_{\bullet B1} - p_{\bullet\bullet 1}) \leq 0\Big\},$$

where, for example, $p_{AB\bullet} := P_{\{1,2\}}(A \times B)$ and $p_{\bullet\bullet 1} := P_{\{1,3\}}([r] \times \{1\}) = P_{\{2,3\}}([s] \times \{1\})$. However, the extension in the first part of Theorem 8 below, which provides an exact expression for the incompatibility index for an arbitrary family of consistent marginal distributions, is new. The second part provides a representation of $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ as an intersection of $F = (2^r - 2)(2^s - 2)$ closed halfspaces; thus, C'_α is known exactly, and can be used in our test of compatibility.

THEOREM 8. *Let $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ and $\mathcal{X} = [r] \times [s] \times [2]$ for some $r, s \in \mathbb{N}$. Then for any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$, we have*

$$R(P_{\mathbb{S}}) = 2 \max_{A\subseteq[r], B\subseteq[s]} (-p_{AB\bullet} + p_{A\bullet 1} + p_{\bullet B1} - p_{\bullet\bullet 1})_+. \tag{7}$$

Moreover,

$$\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**} = \Big\{P_{\mathbb{S}} \equiv p_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},*} : \max_{A\subseteq[r], B\subseteq[s]} (-p_{AB\bullet} + p_{A\bullet 1} + p_{\bullet B1} - p_{\bullet\bullet 1}) \leq 1/2\Big\}.$$

Remark: In the special case $s = 2$, the expression in (7) simplifies to

$$R(P_{\mathbb{S}}) = 2 \max_{j \in [2]} \Big\{ p_{\bullet j 1} - \sum_{i=1}^{r} \min(p_{ij\bullet}, p_{i\bullet 1}) \Big\}_+ . \tag{8}$$

This can be compared with corresponding expressions in the $d = 4$ cases that are given Example 14 and in Proposition S3.

From the expression for F in this case, we see that when $n_{\{1,2\}} = n_{\{2,3\}} = n_{\{1,3\}} = n/3$, we have

$$C'_\alpha + C'_\beta \asymp \Big\{ \frac{r + s + \log\big(1/(\alpha \wedge \beta)\big)}{n} \Big\}^{1/2}, \quad C_\alpha + C_\beta \asymp \Big\{ \frac{rs + \log\big(1/(\alpha \wedge \beta)\big)}{n} \Big\}^{1/2}.$$

More generally, as a consequence of Theorems 7 and 8,

$$\rho^*(n_{\mathbb{S}}) \lesssim \Big(\frac{r+s}{n_{\{1,2\}}} \Big)^{1/2} + \Big(\frac{r}{n_{\{1,3\}}} \Big)^{1/2} + \Big(\frac{s}{n_{\{2,3\}}} \Big)^{1/2}. \tag{9}$$

The main challenge in the proof of Theorem 8 is to establish (7), since the second part then follows using arguments from the proof of Proposition 6. Our strategy is to obtain matching lower and upper bounds on $R(P_{\mathbb{S}})$ via the primal and dual formulations (2) and (3) respectively. The lower bound requires, for each $A \subseteq [r]$ and $B \subseteq [s]$, a construction of $f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^+$ for which we can compute $R(P_{\mathbb{S}}, f_{\mathbb{S}})$. On the other hand, the upper bound relates $R(P_{\mathbb{S}})$ to the maximum two-commodity flow (Ahuja, Magnanti and Orlin, 1988, Chapter 17) through a specially-chosen network. Vlach (1986) gives a halfspace representation for $\mathcal{P}_{\mathbb{S}}^{0,*}$ using the max-flow min-cut theorem for a single-commodity flow through a simpler network; since there is no general max-flow min-cut theorem for two-commodity flows (Leighton and Rao, 1999), our proof is more involved.

Theorem 9 below provides a lower bound on the minimax testing radius in the setting of Theorem 8.

THEOREM 9. *Let* $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ *with* $|\mathcal{X}_1| = r$ *for some* $r \geq 2$, $|\mathcal{X}_2| = 2$ *and* $|\mathcal{X}_3| = 2$. *There exists a universal constant* $c > 0$ *such that*

$$\rho^*(n_{\mathbb{S}}) \geq c \max\Big\{ \frac{1}{\log r} \wedge \Big(\frac{r}{(n_{\{1,2\}} \wedge n_{\{1,3\}}) \log r} \Big)^{1/2}, \frac{1}{(\min_{S \in \mathbb{S}} n_S)^{1/2}} \Big\}.$$

Theorem 9 may be applied in $r \times s \times 2$ tables by noting that ρ^* cannot decrease when $|\mathcal{X}_S|$ increases, for any $S \in \mathbb{S}$. In particular, writing $\rho^*_{r,s,2}(n_{\mathbb{S}})$ here to emphasise the dependence on the alphabet sizes, in the main regime of interest where $n_{\{1,2\}} \geq (r+s)\log(r+s), n_{\{1,3\}} \geq r \log r$ and $n_{\{2,3\}} \geq s \log s$, we can conclude that

$$\begin{aligned} \rho^*_{r,s,2}(n_{\mathbb{S}}) &\geq \max\{\rho^*_{r,2,2}(n_{\mathbb{S}}), \rho^*_{2,s,2}(n_{\mathbb{S}})\} \\ &\gtrsim \Big(\frac{r+s}{n_{\{1,2\}} \log(r+s)} \Big)^{1/2} + \Big(\frac{r}{n_{\{1,3\}} \log r} \Big)^{1/2} + \Big(\frac{s}{n_{\{2,3\}} \log s} \Big)^{1/2}. \end{aligned}$$

When compared with our upper bound in (9), we see that our improved test is minimax rate-optimal, up to logarithmic factors.

The proof of Theorem 9 relies on Lemma S1 in Section S1, which provides a bound on the total variation distance between paired Poisson mixtures, and is an extension of both Wu and Yang (2016, Lemma 3) and Jiao, Han and Weissman (2018, Lemma 32). We remark that the sequences $P_{\mathbb{S}}$ constructed in our lower bound belong to $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$; in other words, the same lower bound on the minimax separation rate holds for testing against consistent alternatives.

3.3. *Reductions.* In this subsection, we show how, for certain $\mathbb{S} \subseteq 2^{[d]}$, the incompatibility index $R(P_{\mathbb{S}})$ can be expressed in terms of $R(P_{\mathbb{S}'})$ for some collection $\mathbb{S}' \subseteq 2^{\mathcal{J}}$, with $\mathcal{J}$ a proper subset of $[d]$. Conceptually, such formulae provide understanding of the facet structure of $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$, which in turn allows us to obtain tighter bounds on the critical values employed in our improved test (cf. Section 3.2). Computationally, these formulae extend the scope of results such as Theorem 8 by allowing us to provide explicit expressions for $R(P_{\mathbb{S}})$ in a wider range of examples. Finally, these reductions allow us to conclude that our improved test is minimax optimal (up to logarithmic factors) for wider classes of observation patterns $\mathbb{S}$.

Our first reduction considers a setting where there exists a subset of variables that are only observed as part of a single observation pattern within our class of possible patterns. Given $\mathbb{S} \subseteq 2^{[d]}$ and $J \subseteq [d]$, we write $\mathbb{S}^{-J} := \{S \cap J^c : S \in \mathbb{S}\}$.

PROPOSITION 10. *Let $\mathbb{S} \subseteq 2^{[d]}$, and suppose that $\emptyset \neq J \subseteq [d]$ and $S_0 \in \mathbb{S}$ are such that $J \subseteq S_0$ but $J \cap S = \emptyset$ for all $S \in \mathbb{S} \setminus \{S_0\}$. Writing $P_{\mathbb{S}}^{-J} := (P_S : S \in \mathbb{S} \setminus S_0, P_{S_0}^{S_0 \cap J^c})$, we have that if $P_{\mathbb{S}}^{-J} \in \mathcal{P}_{\mathbb{S}^{-J}}^{\mathrm{cons}}$, then $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$. Moreover, regardless of consistency,*

$$R(P_{\mathbb{S}}) = R(P_{\mathbb{S}}^{-J}).$$

As an illustration of Proposition 10 suppose that $\mathcal{X} = [r] \times [s] \times [2] \times [t] \times [u]$ and $\mathbb{S} = \big\{\{1,2,4\},\{2,3\},\{1,3,5\}\big\}$. Then $R(P_{\mathbb{S}}) = R(P_{\mathbb{S}^{-\{4,5\}}})$, and if $P_{\mathbb{S}}^{-\{4,5\}} \in \mathcal{P}_{\mathbb{S}^{-\{4,5\}}}^{\mathrm{cons}}$, then

$$R(P_{\mathbb{S}}) = 2 \max_{A \subseteq [r], B \subseteq [s]} \big(-p_{AB\bullet\bullet\bullet} + p_{A\bullet 1\bullet\bullet} + p_{\bullet B1\bullet\bullet} - p_{\bullet\bullet 1\bullet\bullet}\big)_+.$$

Moreover, when testing H_0' in this setting, we may take the same critical value as when $\mathbb{S} = \big\{\{1,2\},\{1,3\},\{2,3\}\big\}$ and $\mathcal{X} = [r] \times [s] \times [2]$, and analogously to (9), we obtain the following upper bound on the minimax testing radius:

$$\rho^*(n_{\mathbb{S}}) \lesssim \Big(\frac{r+s}{n_{\{1,2,4\}}}\Big)^{1/2} + \Big(\frac{r}{n_{\{1,3,5\}}}\Big)^{1/2} + \Big(\frac{s}{n_{\{2,3\}}}\Big)^{1/2}.$$

Our lower bound arguments from Theorem 9 can also be adapted to this five-dimensional setting. Briefly, mimicking the proof of Theorem 9, we should ensure in our choice of priors that the marginals over variables $\{1,2,4\}$, $\{2,3\}$ and $\{1,3,5\}$ agree with those for $\{1,2\}$, $\{2,3\}$ and $\{1,3\}$ respectively in that earlier proof. This can be achieved by taking, for example, $p_{ij\bullet k\bullet}$ to equal $p_{ij\bullet}/t$, where $p_{ij\bullet}$ was defined in the proof of Theorem 9. Arguing in this way allows us to conclude that the lower bound on $\rho^*(n_{\mathbb{S}})$ from Theorem 9 continues to hold, provided that we replace $n_{\{1,2\}}$ and $n_{\{1,3\}}$ with $n_{\{1,2,4\}}$ and $n_{\{1,3,5\}}$ respectively. In other words, our improved test is indeed minimax optimal up to logarithmic factors when $\mathbb{S} = \big\{\{1,2,4\},\{2,3\},\{1,3,5\}\big\}$ and $\mathcal{X} = [r] \times [s] \times [2] \times [t] \times [u]$.

Next, we consider a complementary situation where a subset of variables appears in all of our possible observation patterns. For the purposes of this result, we will assume that $(\mathcal{X}_j : j \in [d])$ are Polish spaces, so that regular conditional distributions and disintegrations are well-defined (e.g. Dudley (2018, Chapter 10) and Reeve, Cannings and Samworth (2021, Lemma 35)). Specifically, if $S \subseteq [d]$ and $J \subseteq S$, then there exists a family $(P_{S|x_J} : x_J \in \mathcal{X}_J)$ of probability measures on $\mathcal{X}_{S \cap J^c}$ with the properties that $x_J \mapsto P_{S|x_J}(B)$ is measurable for every measurable $B \subseteq \mathcal{X}_{S \cap J^c}$, and $\int_A P_{S|x_J}(B)\, dP_S^J(x_J) = P_S(A \times B)$ for all $A \in \mathcal{A}_J, B \in \mathcal{A}_{S \cap J^c}$. We then write $P_{\mathbb{S}|x_J} := (P_{S|x_J} : S \in \mathbb{S})$ for each $x_J \in \mathcal{X}_J$.

PROPOSITION 11. *Let $\mathbb{S} \subseteq 2^{[d]}$, and suppose that $J \subseteq [d]$ is such that $J \subseteq S$ for every $S \in \mathbb{S}$. Suppose further that there exists a distribution P^J on $\mathcal{X}_J$ such that $P^J_S = P^J$ for all $S \in \mathbb{S}$. Then*

$$R(P_{\mathbb{S}}) \leq \int_{\mathcal{X}_J} R(P_{\mathbb{S}|x_J})\, dP^J(x_J). \tag{10}$$

Moreover, in the discrete case where $\mathcal{X}_j = [m_j]$ for some $m_1, \ldots, m_d \in \mathbb{N} \cup \{\infty\}$, the inequality (10) is in fact an equality.

As an application of Proposition 11, suppose that $\mathbb{S} = \big\{\{1,2,3\},\{1,3,4\},\{1,2,4\}\big\}$, where $\mathcal{X}_1 = [r], \mathcal{X}_2 = [s], \mathcal{X}_3 = [t], \mathcal{X}_4 = [2]$, and where $P_{\mathbb{S}} \in \mathcal{P}^{\mathrm{cons}}_{\mathbb{S}}$. Then Proposition 11 combined with Theorem 8 yields that

$$R(P_{\mathbb{S}}) = 2\sum_{i=1}^{r} \max_{A\subseteq[s], B\subseteq[t]} (-p_{iAB\bullet} + p_{iA\bullet 1} + p_{i\bullet B1} - p_{i\bullet\bullet 1})_+.$$

This shows that in this setting we can find $f^{(1)}_{\mathbb{S}}, \ldots, f^{(F)}_{\mathbb{S}} \in \mathcal{G}^+_{\mathbb{S}}$ such that $R(P_{\mathbb{S}}) = \max_{\ell\in[F]} R(P_{\mathbb{S}}, f^{(\ell)}_{\mathbb{S}})_+$, with $F \leq \{(2^s-2)(2^t-2)\}^r \leq 2^{r(s+t)}$. Moreover, analogously to (9), we may argue that

$$\rho^*(n_{\mathbb{S}}) \lesssim \Big(\frac{r(s+t)}{n_{\{1,2,3\}}}\Big)^{1/2} + \Big(\frac{rs}{n_{\{1,2,4\}}}\Big)^{1/2} + \Big(\frac{rt}{n_{\{1,3,4\}}}\Big)^{1/2}.$$

It is a consequence of Proposition S2 that this rate is optimal, up to logarithmic factors.

Our final reduction result provides good upper and lower bounds on $R(P_{\mathbb{S}})$ in settings where there exists $J \in \mathbb{S}$ such that $[d]$ can be partitioned into (I, J, K), where every $S \in \mathbb{S}$ is a subset of either $I \cup J$ or $J \cup K$. As an alternative way of expressing this, if $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{S}$, we say $J \in \mathbb{S}$ is a *cut set* for $\mathbb{S}_1$ and $\mathbb{S}_2$ if $\mathbb{S}_1 \cap \mathbb{S}_2 = \{J\}$ and $(\cup_{S\in\mathbb{S}_1} S) \cap (\cup_{S\in\mathbb{S}_2} S) = J$.

PROPOSITION 12. *Let $\mathbb{S} \subseteq 2^{[d]}$, and suppose that $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{S}$ are such that J is a cut set for $\mathbb{S}_1$ and $\mathbb{S}_2$. Then for any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$, we have*

$$\max\{R(P_{\mathbb{S}_1}), R(P_{\mathbb{S}_2})\} \leq R(P_{\mathbb{S}}) \leq R(P_{\mathbb{S}_1}) + R(P_{\mathbb{S}_2}).$$

In Example 14(ii) below, we give an exact expression for $R(P_{\mathbb{S}})$ when $P_{\mathbb{S}} \in \mathcal{P}^{\mathrm{cons}}_{\mathbb{S}}$ in the special case where $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\},\{3,4\},\{1,4\}\big\}$ with $\mathcal{X}_j = [2]$ for all $j \in [4]$. Here, $\{1,3\}$ is a cut set for $\mathbb{S}_1 = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ and $\mathbb{S}_2 = \big\{\{1,3\},\{3,4\},\{1,4\}\big\}$ (see Figure 2(b)), and our calculations confirm that the conclusion of Proposition 12 holds with these choices of $\mathbb{S}_1, \mathbb{S}_2$ and J. More generally, when $\mathbb{S}$ is as above, $\mathcal{X} = [2]\times[r]\times[s]\times[t]$ and $P_{\mathbb{S}} \in \mathcal{P}^{\mathrm{cons}}_{\mathbb{S}}$, we can now see that $\tilde{R}(P_{\mathbb{S}}) \leq R(P_{\mathbb{S}}) \leq 2\tilde{R}(P_{\mathbb{S}})$, where

$$\tilde{R}(P_{\mathbb{S}}) := 2\max\Big\{ \max_{A\subseteq[r], B\subseteq[s]} (-p_{\bullet AB\bullet} + p_{1A\bullet\bullet} + p_{1\bullet B\bullet} - p_{1\bullet\bullet\bullet}), \\ \max_{A\subseteq[t], B\subseteq[s]} (-p_{\bullet\bullet BA} + p_{1\bullet\bullet A} + p_{1\bullet B\bullet} - p_{1\bullet\bullet\bullet}) \Big\}_+ .$$

Thus (5) holds with $D_R = 2$ and $F' = (2^s-2)\{(2^r-2)+(2^t-2)\} \leq 2^{s+\max(r,t)+1}$, so we can apply our test using the critical value C'_α with these choices. In particular, we can deduce from this that

$$\rho^*(n_{\mathbb{S}}) \lesssim \Big(\frac{r+s}{n_{\{2,3\}}}\Big)^{1/2} + \Big(\frac{r}{n_{\{1,2\}}}\Big)^{1/2} + \Big(\frac{s}{n_{\{1,3\}}}\Big)^{1/2} + \Big(\frac{s+t}{n_{\{3,4\}}}\Big)^{1/2} + \Big(\frac{t}{n_{\{1,4\}}}\Big)^{1/2}. \tag{11}$$

Proposition 12 also allows us to extend lower bounds on the minimax separation rate to this setting. Indeed, when $P_{\mathbb{S}_2}$ is compatible, we see that $R(P_{\mathbb{S}}) = R(P_{\mathbb{S}_1})$ so that the constructions in our lower bound in Theorem 9 can be extended to apply here. Indeed, if we define $p_{j\bullet i\bullet}$ as $p_{ij\bullet}$ was defined in proof of Theorem 9, define $p_{ji\bullet\bullet}$ as $p_{\bullet ij}$ was defined in proof of Theorem 9, and take $p_{\bullet ji\bullet} = p_{\bullet\bullet ij} = \mathbb{1}_{\{j\leq 2\}}/(2s)$ and $p_{i\bullet\bullet j} = \mathbb{1}_{\{j\leq 2\}}/4$ then $P_{\mathbb{S}_2}$ is compatible and the same calculations as in the proof of Theorem 9 show that

$$\rho^*(n_{\mathbb{S}}) \gtrsim \frac{1}{\log s} \wedge \left(\frac{s}{n_{\{1,3\}} \log s}\right)^{1/2}.$$

The constructions for the other terms in (11) are simple modifications of this and we see that our test is again minimax optimal up to logarithmic factors.

3.4. *Computation.* While C_α can be easily calculated for any test of compatibility and allows for a test with power against all incompatible alternatives, we have seen that C'_α can be smaller and lead to more powerful tests. Practical use of C'_α requires knowledge of the number F of essential facets of the polyhedral set $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$, or D_R and F' such that (5) holds. These are fully determined by $\mathbb{S}$ and $\mathcal{X}$, so in principle are known, but these polyhedral sets can be highly complex and explicit expressions for their numbers of essential facets are not generally available. Nevertheless, given particular $\mathbb{S}$ and $\mathcal{X}$, it is possible to compute explicit halfspace representations of $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ using well-developed packages for linear programming. In this section we describe some of the basic geometric concepts involved and how existing algorithms can be used in our setting. As our concern is to describe computational methods, we restrict attention to discrete settings where $|\mathcal{X}| < \infty$, so $\mathcal{P}_{\mathbb{S}}$ is finite-dimensional.

Existing work mentioned in the introduction has focused on the simpler problem of the computation of the facet structure of $\mathcal{P}_{\mathbb{S}}^0$, and we begin by describing the approach taken there. Given $P_{\mathbb{S}} := (P_S : S \in \mathbb{S}) \in \mathcal{P}_{\mathbb{S}}^0$, we write $p_S(x_S) := P_S(\{x_S\})$ for $(S, x_S) \in \mathcal{X}_{\mathbb{S}}$ and $p_{\mathbb{S}} := (p_S : S \in \mathbb{S}) \in [0,1]^{\mathcal{X}_{\mathbb{S}}}$. Thus

$$\begin{aligned}
\mathcal{P}_{\mathbb{S}}^0 &= \{P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}} : \mathcal{F}_{\mathbb{S}}(P_{\mathbb{S}}) \neq \emptyset\} \\
&= \left\{P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}} : \exists p \in [0,1]^{\mathcal{X}} \text{ s.t. } p_S(x_S) = \sum_{x_{S^c} \in \mathcal{X}_{S^c}} p(x_S, x_{S^c}) \, \forall S \in \mathbb{S}, x_S \in \mathcal{X}_S\right\} \\
&= \{P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}} : \exists p \in [0,1]^{\mathcal{X}} \text{ s.t. } \mathbb{A}p = p_{\mathbb{S}}\},
\end{aligned}$$

where the matrix $\mathbb{A} = (\mathbb{A}_{(S,y_S),x})_{(S,y_S)\in\mathcal{X}_{\mathbb{S}}, x\in\mathcal{X}} \in \{0,1\}^{\mathcal{X}_{\mathbb{S}}\times\mathcal{X}}$ has entries

$$\mathbb{A}_{(S,y_S),x} := \mathbb{1}_{\{x_S = y_S\}}. \tag{12}$$

Since each column of $\mathbb{A}$ has exactly $|\mathbb{S}|$ entries equal to 1 (one for each $S \in \mathbb{S}$), it follows that any $p \in [0,1]^{\mathcal{X}}$ with $\mathbb{A}p = p_{\mathbb{S}}$ satisfies $1_{\mathcal{X}}^T p = |\mathbb{S}|^{-1} 1_{\mathcal{X}_{\mathbb{S}}}^T \mathbb{A}p = |\mathbb{S}|^{-1} 1_{\mathcal{X}_{\mathbb{S}}}^T p_{\mathbb{S}} = 1$. We can therefore write $\mathcal{P}_{\mathbb{S}}^0$ as the convex hull of the columns of $\mathbb{A}$, with coefficients in the convex combination given by p. In the rest of this section, we adopt for compactness the convention that if $i \in [2]$, then $\bar{i} := 3 - i$, so that $\{\bar{i}\} = \{1,2\} \setminus \{i\}$.

EXAMPLE 13. *Consider the case $d = 3$, where $\mathbb{S} = \{\{1,2\},\{2,3\},\{1,3\}\}$ and $\mathcal{X} = [2]^3$. Here we have $|\mathcal{X}| = 8, |\mathcal{X}_{\mathbb{S}}| = 12$ and, if we order the 12 rows according to $(1,1,\bullet),(1,2,\bullet),(2,1,\bullet),(2,2,\bullet)$ for $S = \{1,2\}$, then $(\bullet,1,1),(\bullet,1,2),(\bullet,2,1),(\bullet,2,2)$ for*

$S = \{2,3\}$, *then* $(1,\bullet,1),(2,\bullet,1),(1,\bullet,2),(2,\bullet,2)$ *for* $S = \{1,3\}$, *we have*

$$\mathbb{A}^T = \begin{pmatrix} 1&0&0&0&1&0&0&0&1&0&0&0 \\ 1&0&0&0&0&1&0&0&0&0&1&0 \\ 0&1&0&0&0&0&1&0&1&0&0&0 \\ 0&1&0&0&0&0&0&1&0&0&1&0 \\ 0&0&1&0&1&0&0&0&0&1&0&0 \\ 0&0&1&0&0&1&0&0&0&0&0&1 \\ 0&0&0&1&0&0&1&0&0&1&0&0 \\ 0&0&0&1&0&0&0&1&0&0&0&0 \end{pmatrix}.$$

In this case, the polytope $\mathcal{P}^0_{\mathbb{S}}$ *has 16 facets; of these, 12 correspond to the simple non-negativity conditions* $p_{\mathbb{S}} \geq 0$*, while the remaining four essential facets are given by* $p_{i,\bar{j},\bullet} + p_{\bullet j 2} + p_{\bar{i},\bullet,1} \leq 1$ *for* $i,j \in [2]$ *(Vlach, 1986; Eriksson et al., 2006). More generally, when* $\mathcal{X} = [r] \times [s] \times [2]$ *for some* $r,s \in \mathbb{N}$*, the marginal polytope* $\mathcal{P}^0_{\mathbb{S}}$ *has* $(2^r-2)(2^s-2)+rs+2(r+s)$ *facets, with* $rs+2(r+s)$ *of these corresponding to simple non-negativity conditions.*

We now turn to the problem of computing the number of essential facets of the polyhedral set $\mathcal{P}^{0,*}_{\mathbb{S}} + \mathcal{P}^{\text{cons},**}_{\mathbb{S}}$, which is of more direct relevance in our context. As we see from Theorem 8 and the example above, in the special case $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ and $\mathcal{X} = [r] \times [s] \times [2]$, the structure of $\mathcal{P}^{0,*}_{\mathbb{S}} + \mathcal{P}^{\text{cons},**}_{\mathbb{S}}$ is similar to that of $\mathcal{P}^0_{\mathbb{S}}$; indeed, both polyhedral sets have the same numbers of essential and non-essential facets. However, the facet structure of $\mathcal{P}^{0,*}_{\mathbb{S}} + \mathcal{P}^{\text{cons},**}_{\mathbb{S}}$ is generally more complicated than that of $\mathcal{P}^0_{\mathbb{S}}$; Example 14 reveals that when $d = 4$ all irreducible choices of $\mathbb{S}$ except the simple chain pairs case exhibit this difference. The Minkowski sum $\mathcal{P}^{0,*}_{\mathbb{S}} + \mathcal{P}^{\text{cons},**}_{\mathbb{S}}$ is the convex hull of a set of directions (the columns of $\mathbb{A}$) and a set of points (the vertices of $\mathcal{P}^{\text{cons}}_{\mathbb{S}}$, together with the origin). Moreover, a halfspace representation of $\mathcal{P}^{\text{cons}}_{\mathbb{S}}$ is given by

$$\mathcal{P}^{\text{cons}}_{\mathbb{S}} = \Bigg\{ p_{\mathbb{S}} \in [0,\infty)^{\mathcal{X}_{\mathbb{S}}} : \sum_{x_S \in \mathcal{X}_S} p_S(x_S) = 1 \; \forall S \in \mathbb{S}, \\ \sum_{x_{S_1 \cap S_2^c} \in \mathcal{X}_{S_1 \cap S_2^c}} p_{S_1}(x_{S_1}) - \sum_{x_{S_1^c \cap S_2} \in \mathcal{X}_{S_1^c \cap S_2}} p_{S_2}(x_{S_2}) = 0 \, \forall x_{S_1 \cap S_2} \in \mathcal{X}_{S_1 \cap S_2}, S_1, S_2 \in \mathbb{S} \Bigg\},$$

and we can convert this to a vertex representation using software such as the `rcdd` package in `R` (Geyer and Meeden, 2021). In fact, as shown by Proposition 5, the equality constraints of $\mathcal{P}^{\text{cons}}_{\mathbb{S}}$ can be extracted from the equality constraints in the halfspace representation of $\mathcal{P}^0_{\mathbb{S}}$, a fact we use in our computations. The vertex representations of $\mathcal{P}^0_{\mathbb{S}}$ and $\mathcal{P}^{\text{cons}}_{\mathbb{S}}$ lead to a vertex representation of the sum $\mathcal{P}^{0,*}_{\mathbb{S}} + \mathcal{P}^{\text{cons},**}_{\mathbb{S}}$ that can then be converted back to a halfspace representation, again using software such as `rcdd`. The value of F is then given by the number of halfspaces in this representation, once we subtract the number of halfspaces defining $\mathcal{P}^{\text{cons}}_{\mathbb{S}}$.

To illustrate this computational approach, we find explicit expressions for $R(P_{\mathbb{S}})$ with $P_{\mathbb{S}} \in \mathcal{P}^{\text{cons}}_{\mathbb{S}}$, for all irreducible four-dimensional examples with binary variables. If $[d] \in \mathbb{S}$, then $\mathcal{P}^0_{\mathbb{S}} = \mathcal{P}^{\text{cons}}_{\mathbb{S}}$ so $F = 0$ and $R(P_{\mathbb{S}}) = 0$ for $P_{\mathbb{S}} \in \mathcal{P}^{\text{cons}}_{\mathbb{S}}$. We are therefore more interested in situations where $[d] \notin \mathbb{S}$, and where compatibility is not equivalent to consistency. By a combination of Propositions 10 and 11, the set of possible irreducible observation patterns $\mathbb{S}$ in the case $d = 4$ with $[d] \notin \mathbb{S}$ is the following:

- Chain pairs: $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\big\}$;
- All pairs except one: $\mathbb{S} = \big\{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{3,4\}\big\}$;

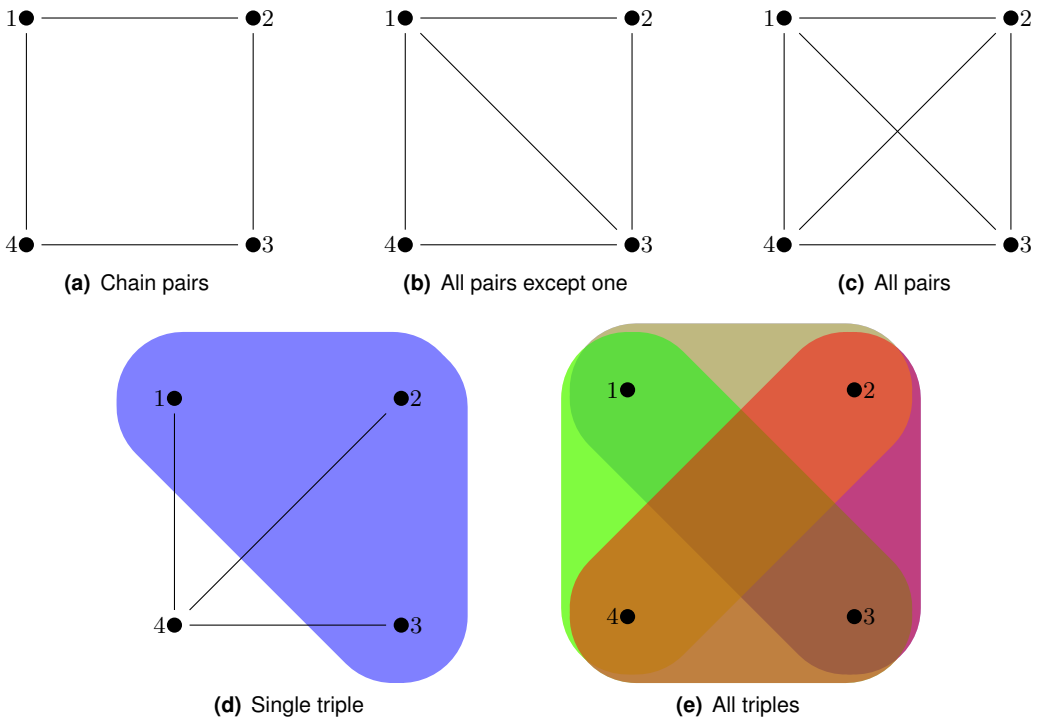

Fig 2: Irreducible observation patterns $\mathbb{S}$ with $d = 4$.

- All pairs: $\mathbb{S} = \big\{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}\big\}$;
- Single triple: $\mathbb{S} = \big\{\{1,2,3\},\{1,4\},\{2,4\},\{3,4\}\big\}$;
- All triples: $\mathbb{S} = \big\{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}\big\}$.

These patterns are illustrated in Figure 2.

EXAMPLE 14. *Let $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}_3 = \mathcal{X}_4 = [2]$. For $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$, the following statements hold:*

(i) *When* $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{3,4\},\{1,4\}\big\}$,

$$R(P_{\mathbb{S}}) = 2 \max_{k,\ell \in [2]} \Big\{ p_{\bullet\bullet k\ell} - p_{\bullet 2k\bullet} - \sum_{i=1}^{2} \min(p_{i1\bullet\bullet}, p_{i\bullet\bullet\ell}) \Big\}_+ \tag{13}$$
$$= 2 \max_{i,j,k \in [2]} (p_{ij\bullet\bullet} - p_{\bullet jk\bullet} - p_{\bullet\bullet\bar{k}1} - p_{i\bullet\bullet 2})_+.$$

From the second representation, we see that we may take $F = 8$. In fact, in this example the facet structure of $\mathcal{P}_{\mathbb{S}}^0$ is again closely related to the facet structure of $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$; indeed, by Hoşten and Sullivant (2002, Theorem 3.5),

$$\mathcal{P}_{\mathbb{S}}^0 = \Big\{ P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}} : \max_{i,j,k \in [2]} (p_{ij\bullet\bullet} - p_{\bullet jk\bullet} - p_{\bullet\bullet\bar{k},1} - p_{i\bullet\bullet 2}) \le 0 \Big\}$$

is a non-redundant halfspace representation. We give an analytic extension of (13) to $\mathcal{X}_1 = [r]$ for general $r \in \mathbb{N}$ in Proposition S3 of Section S1.

(ii) *When* $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\},\{3,4\},\{1,4\}\big\}$,

$$R(P_{\mathbb{S}}) = 2 \max \Big[0, \max_{j \in [2]} \Big\{ p_{\bullet j1\bullet} - \sum_{i=1}^{2} \min(p_{ij\bullet\bullet}, p_{i\bullet 1\bullet}) \Big\} \Big],$$

$$\max_{\ell\in[2]}\Big\{p_{\bullet\bullet 1\ell}-\sum_{i=1}^{2}\min(p_{i\bullet 1\bullet},p_{i\bullet\bullet\ell})\Big\},\ \max_{i,j,\ell\in[2]}(p_{\bullet\bullet 1\ell}-p_{ij\bullet\bullet}-p_{\bar{i}\bullet\bullet\ell}-p_{\bullet\bar{j}1\bullet})\Big]$$
$$=\max\big\{R(P_{\mathbb{S}}^{123}),R(P_{\mathbb{S}}^{134}),R(P_{\mathbb{S}\setminus\{\{1,3\}\}})\big\}.$$

Here, we write, e.g., $R(P_{\mathbb{S}}^{123})$ instead of $R(P_{\mathbb{S}}^{\{1,2,3\}})$ for notational simplicity. This is a simple example where $\mathcal{P}_{\mathbb{S}}^{0,*}+\mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ has a more complex facet structure than that of $\mathcal{P}_{\mathbb{S}}^{0}$. Indeed, Proposition 12 shows that $\max\{R(P_{\mathbb{S}}^{123}),R(P_{\mathbb{S}}^{134})\}\le R(P_{\mathbb{S}})\le R(P_{\mathbb{S}}^{123})+R(P_{\mathbb{S}}^{134})$, and hence that $P_{\mathbb{S}}$ is compatible if and only if $P_{\mathbb{S}}^{123}$ and $P_{\mathbb{S}}^{134}$ are compatible. On the other hand, writing

$$p_{\bullet\bullet 1\ell}-p_{ij\bullet\bullet}-p_{\bar{i}\bullet\bullet\ell}-p_{\bullet\bar{j}1\bullet}=(p_{\bullet\bullet 1\ell}-p_{i\bullet 1\bullet}-p_{\bar{i}\bullet\bullet\ell})+(p_{\bullet j1\bullet}-p_{ij\bullet\bullet}-p_{\bar{i}\bullet 1\bullet}),\tag{14}$$

shows that our expressions for $R(P_{\mathbb{S}})$ are non-redundant: $\mathcal{P}_{\mathbb{S}}^{0,*}+\mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ has $F=16$ essential facets, while $\mathcal{P}_{\mathbb{S}}^{0}$ only has 8.

(iii) *When $\mathbb{S}=\big\{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}\big\}$,*

$$R(P_{\mathbb{S}})=\max\Big[R(P_{\mathbb{S}}^{123}),R(P_{\mathbb{S}}^{124}),R(P_{\mathbb{S}}^{134}),R(P_{\mathbb{S}}^{234}),$$
$$2\max_{i,j,k,\ell\in[2]}(-p_{ij\bullet\bullet}-p_{\bullet jk\bullet}-p_{i\bullet k\bullet}+p_{i\bullet\bullet\ell}+p_{\bullet j\bullet\ell}+p_{\bullet\bullet k\ell}-p_{\bullet\bullet\bullet\ell}),$$
$$R(P_{\mathbb{S}\setminus\{\{1,3\},\{2,4\}\}}),R(P_{\mathbb{S}\setminus\{\{1,4\},\{2,3\}\}}),R(P_{\mathbb{S}\setminus\{\{1,2\},\{3,4\}\}})\Big].\tag{15}$$

Here, we see from the `R` code output that $P_{\mathbb{S}}$ is compatible if and only if the first two lines of (15) are non-positive. The final line can be bounded above by twice the first line: as in (14), we have, for example, that

$$R(P_{\mathbb{S}\setminus\{\{1,3\},\{2,4\}\}})=\max_{i,j,\ell\in[2]}(p_{\bullet\bullet 1\ell}-p_{ij\bullet\bullet}-p_{\bar{i}\bullet\bullet\ell}-p_{\bullet\bar{j}1\bullet})_+\le R(P_{\mathbb{S}}^{134})+R(P_{\mathbb{S}}^{123}).$$

We see from (15) that we may take $F=4\times4+16+3\times8=56$. In the `R` output, the half-space representation of $\mathcal{P}_{\mathbb{S}}^{0,*}+\mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ has 93 rows, 13 of which are equality constraints coming from the consistency conditions, 24 of which are non-negativity constraints, and the remaining 56 correspond to essential facets reflected in our expression $R(P_{\mathbb{S}})$ above. Here $\mathcal{P}_{\mathbb{S}}^{0}$ has 32 essential facets.

(iv) *When $\mathbb{S}=\big\{\{1,2,3\},\{1,4\},\{2,4\},\{3,4\}\big\}$, we have compatibility if and only if $P_{\mathbb{S}}^{124},P_{\mathbb{S}}^{134},P_{\mathbb{S}}^{234}$ are compatible and*

$$\tilde{p}_{ijk\ell}:=p_{ijk\bullet}+p_{\bar{i}\bullet\bullet\ell}+p_{\bullet\bar{j}\bullet\ell}+p_{\bullet\bullet\bar{k}\ell}-p_{\bullet\bullet\bullet\ell}\ge0$$

for all $i,j,k,\ell\in[2]$. These conditions are non-redundant, so that $\mathcal{P}_{\mathbb{S}}^{0}$ has $3\times4+16=28$ essential facets. Further,

$$R(P_{\mathbb{S}})=\max\Big[R(P_{\mathbb{S}}^{124}),R(P_{\mathbb{S}}^{134}),R(P_{\mathbb{S}}^{234}),-\frac{3}{2}\min_{i,j,k,\ell\in[2]}\tilde{p}_{ijk\ell},-\min_{i,j,k,\ell\in[2]}(\tilde{p}_{ijk\ell}+\tilde{p}_{\bar{i}\bar{j}\bar{k}\bar{\ell}})$$
$$-\min_{i,j,k,\ell\in[2]}\big\{\tilde{p}_{\bar{i}\bar{j}\bar{k}\bar{\ell}}+\min(p_{ij\bullet\bullet}-p_{i\bullet\bullet\ell}+p_{\bullet\bar{j}\bullet\ell},p_{\bullet jk\bullet}-p_{\bullet\bullet k\ell}+p_{\bullet\bar{j}\bullet\ell},p_{i\bullet k\bullet}-p_{\bullet\bullet k\ell}+p_{\bar{i}\bullet\bullet\ell})\big\}\Big],$$

so that we may take $F=3\times4+2\times16+3\times16=92$. From the above expression it also follows that

$$R(P_{\mathbb{S}})\le\max\big\{R(P_{\mathbb{S}}^{124}),R(P_{\mathbb{S}}^{134}),R(P_{\mathbb{S}}^{234})\big\}+2\max_{i,j,k,\ell\in[2]}(-\tilde{p}_{ijk\ell})_+.$$

(v) *When* $\mathbb{S} = \big\{\{1,2,3\},\{1,2,4\},\{1,3,4\},\{2,3,4\}\big\}$ *the marginal polytope* $\mathcal{P}^0_{\mathbb{S}}$ *has 32 essential facets. Indeed, writing* $\mathbb{S}_J := \mathbb{S} \setminus ([4] \setminus \{J\})$*, we have that* $P_{\mathbb{S}}$ *is compatible if and only if* $P_{\mathbb{S}_J|x_J}$ *is compatible for all* $J \in [4]$ *and* $x_J \in [2]$*. Moreover,*

$$R(P_{\mathbb{S}}) = \frac{1}{2} \max_{J\in[4]} \max_{x_J\in[2]} \big\{3p^J(x_J)R(P_{\mathbb{S}_J|x_J}) + p^J(\bar{x}_J)R(P_{\mathbb{S}_J|\bar{x}_J})\big\}$$
$$\leq 2 \max_{J\in[4]} \max_{x_J\in[2]} p^J(x_J)R(P_{\mathbb{S}_J|x_J}),$$

and $F = 4 \times 2 \times 4 \times 4 = 128$*. To see where these numbers come from, consider* $J = 1$ *and* $x_J = 1$*, and note that*

$$\frac{1}{2}\{3p^1(1)R(P_{\mathbb{S}_1|1}) + p^1(2)R(P_{\mathbb{S}_1|2})\}$$
$$= -\min\Big\{0, 3 \min_{j,k\in[2]} (p_{1jk\bullet} + p_{1,\bar{j}\bullet 1} - p_{1\bullet k1})\Big\} - \min\Big\{0, \min_{j',k'\in[2]} (p_{2j'k'\bullet} + p_{2,\bar{j}'\bullet 2} - p_{2\bullet k'2})\Big\}.$$

This is the maximum of 5×5 *linear functionals of* $P_{\mathbb{S}_J}$*, but all those where* 0 *is chosen in the first term are redundant in the final expression for* $R(P_{\mathbb{S}})$*, as are all those where* $(j',k') = (\bar{j},\bar{k})$*. Thus, for each value of* (J, x_J)*, there are* 4×4 *non-redundant essential facets.*

4. Mixed discrete and continuous variables. In this section, we consider a setting of mixed discrete and continuous variables, where there exist positive integers $d_0 \leq d$ such that $\mathcal{X} = [0,1)^{d_0} \times \prod_{j=d_0+1}^{d} [m_j]$, with $m_1, \ldots, m_d \in \mathbb{N} \cup \{\infty\}$. The case where the continuous components take values in other spaces, e.g. $\mathbb{R}$, can be handled using similar techniques. We assume that we observe independent random variables $(X_{S,i} : S \in \mathbb{S}, i \in [n_S])$, with $X_{S,i} \sim P_S$ taking values in $\mathcal{X}_S := [0,1)^{S\cap[d_0]} \times \prod_{j\in S\cap([d]\setminus[d_0])} [m_j]$. Given a vector of bandwidths $h = (h_1, \ldots, h_{d_0}) \in (0,\infty)^{d_0}$, we partition $[0,1)^{d_0} \times \prod_{j\in[d]\setminus[d_0]} [m_j]$ as

$$[0,1)^{d_0} \times \prod_{j\in[d]\setminus[d_0]} [m_j] = \bigcup_{(k_1,\ldots,k_d)\in\mathcal{K}_h} \Big(\prod_{j=1}^{d_0} I_{h_j,k_j} \times \prod_{j\in[d]\setminus[d_0]} \{k_j\}\Big),$$

where $\mathcal{K}_h := [\lceil 1/h_1 \rceil] \times \cdots \times [\lceil 1/h_{d_0} \rceil] \times \prod_{j\in[d]\setminus[d_0]} [m_j]$ and

$$I_{h_j,k_j} := \big[(k_j - 1)h_j, (k_j h_j) \wedge 1\big).$$

Let $\mathcal{G}^+_{\mathbb{S},h}$ denote the set of sequences of functions $(f_S : S \in \mathbb{S})$ where each $f_S : \mathcal{X}_S \to [-1,\infty)$ is piecewise constant on all sets of the form $\prod_{j\in S\cap[d_0]} I_{h_j,k_j} \times \prod_{j\in S\cap([d]\setminus[d_0])} \{k_j\}$ and where the sequence satisfies

$$\inf_{x\in\mathcal{X}} \sum_{S\in\mathbb{S}} f_S(x_S) \geq 0.$$

We further define

$$R_h(P_{\mathbb{S}}) := \sup_{f_{\mathbb{S}}\in\mathcal{G}^+_{\mathbb{S},h}} R(P_{\mathbb{S}}, f_{\mathbb{S}}).$$

Recalling our definition of $C^{\min}_\alpha$ from Section 3.2, in this mixed continuous and discrete setting, we reject the null hypothesis that $P_{\mathbb{S}} = (P_S : S \in \mathbb{S}) \in \mathcal{P}^0_{\mathbb{S}}$ at the level $\alpha \in (0,1)$ if

$$R_h(\widehat{P}_{\mathbb{S}}) \geq C^{\min}_\alpha\big(\lceil 1/h_1 \rceil, \ldots, \lceil 1/h_{d_0} \rceil, m_{d_0+1}, \ldots, m_d, \mathbb{S}, (n_S : S \in \mathbb{S})\big) =: C^*_\alpha,$$

where $\widehat{P}_S$ is the empirical distribution of $X_{S,1}, \ldots, X_{S,n_S}$ for $S \in \mathbb{S}$, and $\widehat{P}_{\mathbb{S}} = (\widehat{P}_S : S \in \mathbb{S})$.

For $d' \in \mathbb{N}$, $r = (r_1, \ldots, r_{d'}) \in (0,1]^{d'}$ and $L > 0$, let $\mathcal{H}_{d'}(r, L)$ denote the class of functions that are (r, L)-Hölder on $[0,1)^{d'}$, i.e. the set of functions $p : [0,1)^{d'} \to \mathbb{R}$ satisfying

$$|p(z_1, \ldots, z_{d'}) - p(z'_1, \ldots, z'_{d'})| \leq L \sum_{j=1}^{d'} |z_j - z'_j|^{r_j}$$

for all $(z_1, \ldots, z_{d'}), (z'_1, \ldots, z'_{d'}) \in [0,1)^{d'}$. Now let $\mathcal{P}_{\mathbb{S},r,L}$ denote the set of sequences of distributions $(P_S : S \in \mathbb{S})$ where P_S is a distribution on $\mathcal{X}_S$ having density p_S with respect to the Cartesian product of Lebesgue measure on $[0,1)^{S \cap [d_0]}$ and counting measure on $\prod_{j \in S \cap ([d] \setminus [d_0])} [m_j]$ satisfying the condition that the conditional density $x_{S \cap [d_0]} \mapsto p_S(x_{S \cap [d_0]} | x_{S \cap ([d] \setminus [d_0])})$ belongs to $\mathcal{H}_{|S \cap [d_0]|}(r, L)$ for all $x_{S \cap ([d] \setminus [d_0])} \in \prod_{j \in S \cap ([d] \setminus [d_0])} [m_j]$.

THEOREM 15. *In the above setting, let $\alpha, \beta \in (0,1)$ and suppose that $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S},r,L}$. Then the probability of a Type I error for our test is at most α. Moreover, if*

$$R(P_{\mathbb{S}}) \geq L(|\mathbb{S}| - 1) \sum_{j=1}^{d_0} h_j^{r_j} + C^*_\alpha + C^*_\beta,$$

then the probability of a Type II error is at most β.

We now specialise the upper bound of Theorem 15 to our main three-dimensional example. When $\mathbb{S} = \big\{\{1,2\}, \{2,3\}, \{1,3\}\big\}$ and $\mathcal{X} = [0,1)^2 \times \{1,2\}$ we have for $h_1, h_2 \in (0,1)$ that

$$L(|\mathbb{S}| - 1)(h_1^{r_1} + h_2^{r_2}) + C^*_\alpha + C^*_\beta \lesssim_{L,|\mathbb{S}|,\alpha,\beta} h_1^{r_1} + h_2^{r_2} + \Big(\frac{1/h_1 + 1/h_2}{\min_{S \in \mathbb{S}} n_S}\Big)^{1/2},$$

and we can choose h_1, h_2 to minimise this right-hand side. We can take $h_1 = h_2 = (\min_{S \in \mathbb{S}} n_S)^{-\frac{1}{1+2(r_1 \wedge r_2)}}$ and $\alpha = \beta = 1/4$ to deduce the minimax upper bound

$$\rho^*(n_{\mathbb{S}}) \lesssim_{L,|\mathbb{S}|} \Big(\min_{S \in \mathbb{S}} n_S\Big)^{-\frac{r_1 \wedge r_2}{1+2(r_1 \wedge r_2)}}.$$

5. Numerical studies. The tests introduced in Section 3 provide finite-sample Type I error control over the entire null hypothesis parameter space $\mathcal{P}^0_{\mathbb{S}}$. However, this may lead to conservative tests in particular examples, so we first present an alternative, Monte Carlo-based approach to constructing the critical value for our test. The first part of Proposition 6 and the dual formulation (3) mean that we can write

$$\widehat{P}_{\mathbb{S}} = \{1 - R(\widehat{P}_{\mathbb{S}})\}\widehat{Q}_{\mathbb{S}} + R(\widehat{P}_{\mathbb{S}})\widehat{T}_{\mathbb{S}},$$

where $\widehat{Q}_{\mathbb{S}} \in \mathcal{P}^0_{\mathbb{S}}$ and $\widehat{T}_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$. Here $\widehat{Q}_{\mathbb{S}}$ can be thought of as a closest compatible sequence of marginal distributions to $\widehat{P}_{\mathbb{S}}$ (in particular, if $\widehat{P}_{\mathbb{S}} \in \mathcal{P}^0_{\mathbb{S}}$, then $\widehat{Q}_{\mathbb{S}} = \widehat{P}_{\mathbb{S}}$). Moreover, $\widehat{Q}_{\mathbb{S}}$ can be computed straightforwardly at the same time as our test statistic $R(\widehat{P}_{\mathbb{S}})$. In particular, recall from the proof of Theorem 8 that

$$R(\widehat{P}_{\mathbb{S}}) = 1 - \max\big\{1_{\mathcal{X}}^T p : p \in [0,\infty)^{\mathcal{X}}, \mathbb{A}p \leq \widehat{p}_{\mathbb{S}}\big\},$$

where $\widehat{p}_{\mathbb{S}} := (\widehat{p}_S : S \in \mathbb{S})$ is the sequence of mass functions associated with $\widehat{P}_{\mathbb{S}}$. Writing $\widehat{p}$ for an optimal solution to this linear program, and assuming initially that $1_{\mathcal{X}}^T \widehat{p} \neq 0$, the closest compatible sequence $\widehat{Q}_{\mathbb{S}}$ is the sequence of distributions associated with the sequence of probability mass functions $\mathbb{A}\widehat{p}/(1_{\mathcal{X}}^T \widehat{p})$. If $1_{\mathcal{X}}^T \widehat{p} = 0$, so $R(\widehat{P}_{\mathbb{S}}) = 1$, then we simply take $\widehat{Q}_{\mathbb{S}}$ to be the sequence of discrete uniform distributions on $\mathcal{X}_{\mathbb{S}}$.

It is therefore natural to generate a critical value by drawing B bootstrap samples from $\widehat{Q}_{\mathbb{S}}$, computing the corresponding empirical distributions $\widehat{Q}_{\mathbb{S}}^{(1)}, \ldots, \widehat{Q}_{\mathbb{S}}^{(B)}$ and test statistics $R(\widehat{Q}_{\mathbb{S}}^{(1)}), \ldots, R(\widehat{Q}_{\mathbb{S}}^{(B)})$, and rejecting H_0' at significance level $\alpha \in (0,1)$ if and only if

$$1 + \sum_{b=1}^{B} \mathbb{1}_{\{R(\widehat{Q}_{\mathbb{S}}^{(b)}) \geq R(\widehat{P}_{\mathbb{S}})\}} \leq \alpha(B+1).$$

We are to give both theoretical and numerical backing for this test. Given $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$ and $\epsilon > 0$, define

$$B_\epsilon(P_{\mathbb{S}}) := \left\{ P'_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}} : \sum_{S \in \mathbb{S}} d_{\mathrm{TV}}(P_S, P'_S) \leq \epsilon \right\} \text{ and } (\mathcal{P}_{\mathbb{S}}^0)^{-\epsilon} := \left\{ P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0 : B_\epsilon(P_{\mathbb{S}}) \subseteq \mathcal{P}_{\mathbb{S}}^0 \right\}.$$

The following result shows that our Monte–Carlo test is uniformly valid over expanding subsets of $\mathcal{P}_{\mathbb{S}}^0$, and uniformly powerful over alternatives separated from the null.

PROPOSITION 16. *Recall the definition of C_α from Section 3.1. We have*

$$\sup_{P_{\mathbb{S}} \in (\mathcal{P}_{\mathbb{S}}^0)^{-C_\alpha}} \mathbb{P}_{P_{\mathbb{S}}} \left(1 + \sum_{b=1}^{B} \mathbb{1}_{\{R(\widehat{Q}_{\mathbb{S}}^{(b)}) \geq R(\widehat{P}_{\mathbb{S}})\}} \leq \alpha(B+1) \right) \leq \alpha.$$

Moreover, if $B \geq 2(1-\alpha)/\alpha$ and $R(P_{\mathbb{S}}) \geq 2\sqrt{2}(C_\alpha + C_\beta)$, then

$$\mathbb{P}_{P_{\mathbb{S}}} \left(1 + \sum_{b=1}^{B} \mathbb{1}_{\{R(\widehat{Q}_{\mathbb{S}}^{(b)}) \geq R(\widehat{P}_{\mathbb{S}})\}} > \alpha(B+1) \right) \leq \beta.$$

In our first experiments, we took $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ with $\mathcal{X} = [r] \times [2]^2$ and $r \in \{2,4,6\}$; we fix $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$ by setting, for each $i \in [r]$,

$$p_{i\bullet\bullet} = \frac{1}{r}, \quad p_{\bullet 1\bullet} = p_{\bullet\bullet 1} = \frac{1}{2}, \quad p_{i\bullet 1} = \frac{1}{2r}, \quad p_{i1\bullet} = \frac{1+(-1)^i}{2r} \tag{16}$$

and varying $p_{\bullet 21}$ to adjust the incompatibility index. Indeed, with these choices, we have $R(P_{\mathbb{S}}) = 2(p_{\bullet 21} - 1/4)_+$ by Theorem 8. Our Monte Carlo test was applied with $n_{\mathbb{S}} = (200, 200, 200)$, $B = 99$ and $\alpha = 0.05$, and we repeated our experiments 5000 times in each setting.

In this setting where we do not have complete cases available and $P_{\mathbb{S}}$ is consistent, we are not aware of alternative methods that would have non-trivial power. Nevertheless, in order to provide some comparison, we can furnish the tests of Fuchs (1982) and Spohn et al. (2021) with an additional $n_{\{1,2,3\}} = 200$ observations from the distribution on $\mathcal{X}$ having mass function $p_{ijk} = \{1 + (-1)^{i+j}\}/(4r)$ for $i \in [r]$ and $j,k \in [2]$, which ensures that $\mathbb{A}p$ is a closest compatible sequence to $P_{\mathbb{S}}$, in our terminology above. In particular, $\mathbb{A}p$ satisfies all equalities in (16), as well as $(\mathbb{A}p)_{\bullet 21} = 1/4$. We emphasise that these complete cases were not accessed by our method. The Fuchs (1982) test was of similar speed to our approach, so it was again possible to repeat each experiment 5000 times and we compare the power curves with those of our test in Figure 3. However, the PKLM test of Spohn et al. (2021) (applied with the default choices of tuning parameters) was considerably slower, so we only studied its performance under the extreme values of $R(P_{\mathbb{S}})$ (i.e. $R(P_{\mathbb{S}}) = 0$ and $R(P_{\mathbb{S}}) = 0.25$), and we only conducted 200 repetitions for each setting; see Table 2.

From Figure 3 and Table 2, we see that all three tests have good control of the size of the test, and in fact the Fuchs and PKLM tests are slightly conservative. Despite the extra

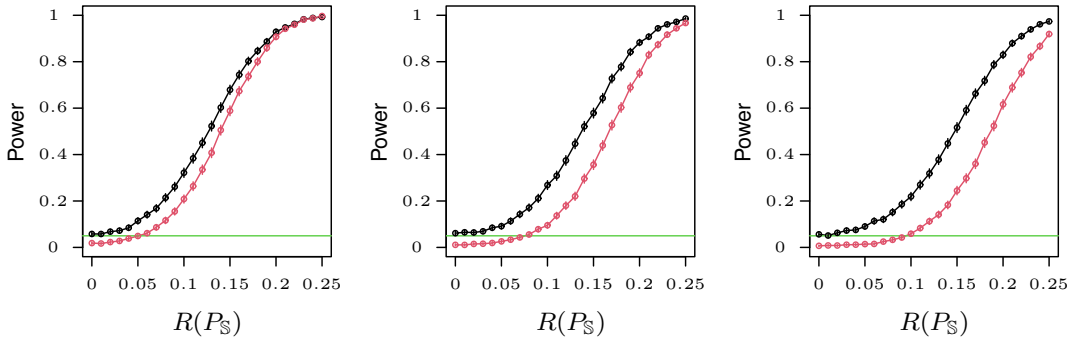

Fig 3: Power curves for our Monte Carlo test (black) and Fuchs's test (red). Error bars show three standard errors. Here, $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ with $\mathcal{X} = [r] \times [2]^2$ and $r = 2$ (left), $r = 4$ (middle) and $r = 6$ (right).

	$p_{\bullet 21} = 0.25$	$p_{\bullet 21} = 0.375$
$r = 2$	0.01	0.50
$r = 4$	0.02	0.485
$r = 6$	0.035	0.395

TABLE 2

Rejection rates for PKLM over 200 repetitions. Here $p_{\bullet 21} = 0.25$ corresponds to $R(P_{\mathbb{S}}) = 0$ and $p_{\bullet 21} = 0.375$ corresponds to $R(P_{\mathbb{S}}) = 0.25$.

complete cases that are available to the alternative methods, though, our test is significantly more powerful, with the difference in power increasing as r increases.

In our second set of experiments, we took $d = 5$, $n_{\mathbb{S}} = (500, 500, 500, 500, 500)$, $\mathcal{X} = [2]^5$ and $\mathbb{S} = \big\{\{1,2,3,4\},\{1,2,3,5\},\{1,2,4,5\},\{1,3,4,5\},\{2,3,4,5\}\big\}$. For $\epsilon \in [0.2, 0.35]$ and $i, j, k, \ell, m \in [2]$, we set

$$p_{ijk\ell\bullet} = \frac{1 + \epsilon(-1)^{i+j+k+\ell}}{16}, \quad p_{ijk\bullet m} = \frac{1 + \epsilon(-1)^{i+j+k+m}}{16}, \quad p_{ij\bullet\ell m} = \frac{1 + \epsilon(-1)^{i+j+\ell+m}}{16}$$

$$p_{i\bullet k\ell m} = \frac{1 + \epsilon(-1)^{i+k+\ell+m}}{16}, \quad p_{\bullet jk\ell m} = \frac{1 - \epsilon(-1)^{j+k+\ell+m}}{16},$$

for which $R(P_{\mathbb{S}}) = (5\epsilon - 1)_+/4$. In this case, we applied the Fuchs test for several different choices of the number of complete cases, namely $n_{\{1,2,3,4,5\}} \in \{25, 50, 100, 200\}$. The complete case distribution p was chosen so that $\mathbb{A}p$ was a closest compatible sequence to $P_{\mathbb{S}}$. Figure 4 shows the corresponding power curves, along with that of our test. In this example, our test is the only one that controls the Type I error at the nominal level, so none of the Fuchs tests are reliable here. We also see that the additional complete cases are crucial for the power of the Fuchs test, and that the power of our test remains competitive even without these observations.

Finally, we present in Figure 5 the results of investigations into the computational run time of our methodology. The computation of $R(\cdot)$ on the bootstrap samples is parallelisable, so we simply report the time taken for a single computation of $R(\cdot)$. The linear programming was carried out using the `gurobi` software in `R` (Gurobi Optimization, LLC, 2021); this takes advantage of sparse matrix representations of $\mathbb{A}$.

For $d \geq 3, m \in [d-1]$ and $r \geq 2$ we consider the collection of observation patterns $\mathbb{S} = \big\{\{1,\ldots,m\},\{2,\ldots,m+1\},\ldots,\{d,1,\ldots,m-1\}\big\}$ and set $\mathcal{X} = [r]^d$. We generate inputs $P_{\mathbb{S}}$ randomly as empirical distributions associated with discrete uniform distributions, with each sample size being 10000. The number of constraints in the linear program here is $r^d + dr^m$

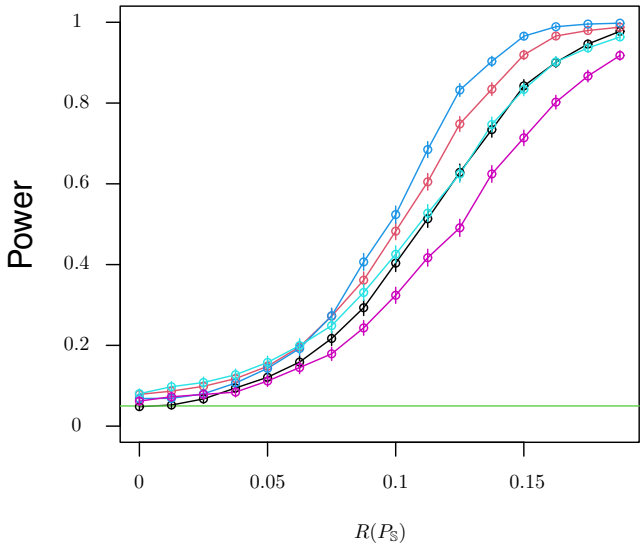

Fig 4: Power curves for our Monte Carlo test (black) and Fuchs's test with the latter test being applied with an additional $n_{\{1,2,3,4,5\}} = 25$ (magenta), 50 (cyan), 100 (red) and 200 (blue) complete cases. Error bars show three standard errors. Here, $\mathbb{S} = \big\{\{1,2,3,4\},\{1,2,3,5\},\{1,2,4,5\},\{1,3,4,5\},\{2,3,4,5\}\big\}$, with $\mathcal{X} = [2]^5$.

and the number of variables is dr^m. In our first set of examples, shown in the left panel of Figure 5 we consider $d = 3, 4, 5$, take $m = 2$ and vary r. When $d = 3$ and $r = 50$, the time taken was approximately 52 seconds; when $d = 5$ and $r = 15$ and the time taken was approximately 46 seconds. In our second set of examples, given in the right panel of Figure 5 we consider $m = 2, 3$ and $r = 2, 3$ and vary d. When $r = m = 2$ and $d = 20$, it took around 25 seconds to compute $R(P_{\mathbb{S}})$; when $r = 3$, $m = 3$ and $d = 13$, the time taken was approximately 66 seconds.

Acknowledgements: We thank Danat Duisenbekov and Sean Jaffe for their assistance in speeding up the computational algorithms, as well as the anonymous reviewers for their constructive comments, which helped to improve the paper.

SUPPLEMENTARY MATERIAL

Supplementary material: Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility
(doi: 10.1214/00-AOSXXXXSUPP; .pdf). The supplement contains proofs of our main results, as well as some auxiliary results.

REFERENCES

ABRAMSKY, S., BARBOSA, R. S. and MANSFIELD, S. (2017). Contextual fraction as a measure of contextuality. *Phys. Rev. Lett.* **119** 050504.

ABRAMSKY, S. and BRANDENBURGER, A. (2011). The sheaf-theoretic structure of non-locality and contextuality. *New J. Phys.* **13** 113036.

AHUJA, R. K., MAGNANTI, T. L. and ORLIN, J. B. (1988). *Network Flows*. Cambridge, Massachusetts.

ALEXANDROFF, P. (1924). Über die Metrisation der im Kleinen kompakten topologischen Räume. *Mathematische Annalen* **92** 294–301.

BELL, J. S. (1966). On the problem of hidden variables in quantum mechanics. *Rev. Mod. Phys.* **38** 447.

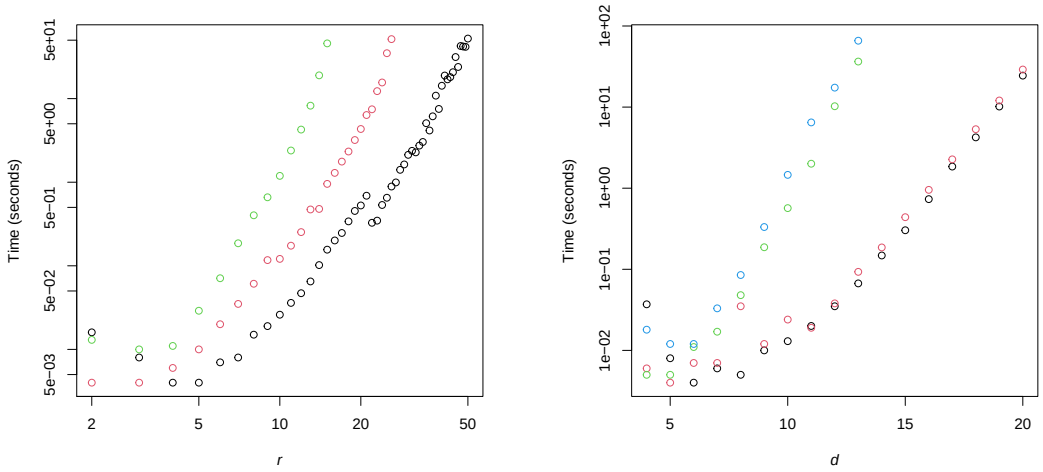

Fig 5: Time taken to compute $R(\cdot)$ in various settings. In the left-hand plot the $d = 3, 4, 5$ examples are represented by the black, red and green points, respectively. In the right-hand plot the $r = 2$ results are shown in black and red for $m = 2, 3$, respectively, and the $r = 3$ results are shown in green and blue for $m = 2, 3$, respectively.

BERRETT, T. B. and SAMWORTH, R. J. (2022). MCARtest: Optimal nonparametric testing of Missing Completely At Random R package version 1.1, available at https://cran.r-project.org/web/packages/MCARtest/index.html.

BERRETT, T. B. and SAMWORTH, R. J. (2023). Supplementary material to 'Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility'. *Submitted*.

BLANCHARD, G., CARPENTIER, A. and GUTZEIT, M. (2018). Minimax Euclidean separation rates for testing convex hypotheses in $\mathbb{R}^d$. *Electr. J. Statist.* **12** 3713–3735.

CAI, T. T. and ZHANG, L. (2019). High dimensional linear discriminant analysis: Optimality, adaptive algorithm and missing data. *J. Roy. Statist. Soc., Ser. B* **81** 675–705.

CHEN, H. Y. and LITTLE, R. (1999). A test of missing completely at random for generalised estimating equations with missing data. *Biometrika* **86** 1–13.

CLAUSER, J. F. and SHIMONY, A. (1978). Bell's theorem. Experimental tests and implications. *Rep. Prog. Phys.* **41** 1881.

COONS, J. I., CUMMINGS, J., HOLLERING, B. and MARAJ, A. (2020). Generalized cut polytopes for binary hierarchical models. *Algebraic Statistics, to appear*.

DALL'AGLIO, G., KOTZ, S. and SALINETTI, G. (2012). *Advances in Probability Distributions with Given Marginals: Beyond the Copulas*. Springer Science & Business Media.

DAVISON, A. C. (2003). *Statistical Models*. Cambridge University Press.

DE LOERA, J. A. and KIM, E. D. (2014). Combinatorics and geometry of transportation polytopes: an update. In *Discrete Geometry and Algebraic Combinatorics* 37–76. Amer. Math. Soc. Providence, RI.

DEZA, M. M. and LAURENT, M. (2009). *Geometry of Cuts and Metrics*. Springer.

DUDLEY, R. M. (2018). *Real Analysis and Probability*. CRC Press.

ELSENER, A. and VAN DE GEER, S. (2019). Sparse spectral estimation with missing and corrupted measurements. *Stat* **8** e229.

EMBRECHTS, P. and PUCCETTI, G. (2010). Bounds for the sum of dependent risks having overlapping marginals. *J. Multivar. Anal.* **101** 177–190.

ERIKSSON, N., FIENBERG, S. E., RINALDO, A. and SULLIVANT, S. (2006). Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models. *J. Symb. Comput.* **41** 222–233.

FARKAS, J. (1902). Theorie der einfachen Ungleichungen. *Journal für die Reine und Angewandte Mathematik* **1902** 1–27.

FIENBERG, S. E. (1968). The geometry of an $r \times c$ contingency table. *Ann. Math. Statist.* **39** 1186–1190.

FOLLAIN, B., WANG, T. and SAMWORTH, R. J. (2022). High-dimensional changepoint estimation with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B* **84** 1023–1055.

FUCHS, C. (1982). Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Amer. Statist. Assoc.* **77** 270–278.
GALE, D. (1957). A theorem on flows in networks. *Pacific J. Math* **7** 1073–1082.
GEYER, C. J. and MEEDEN, G. D. (2021). rcdd: Computational Geometry R package version 1.5, available at https://cran.r-project.org/web/packages/rcdd/index.html.
GUROBI OPTIMIZATION, LLC (2021). Gurobi Optimizer Reference Manual.
HOŞTEN, S. and SULLIVANT, S. (2002). Gröbner bases and polyhedral geometry of reducible and cyclic models. *J. Comb. Theory Ser. A.* **100** 277–301.
ISII, K. (1964). Inequalities of the types of Chebyshev and Cramér-Rao and mathematical programming. *Ann. Inst. Statist. Math.* **16** 277–293.
JAMSHIDIAN, M. and JALAL, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika* **75** 649–674.
JIAO, J., HAN, Y. and WEISSMAN, T. (2018). Minimax estimation of the L_1 distance. *IEEE Trans. Inf. Theory* **64** 6672–6706.
JOE, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. CRC Press.
KANTOROVICH, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)* **37** 199–201.
KANTOROVICH, L. V. (2006). On the translocation of masses. *J. Math. Sci.* **133** 1381–1382.
KELLERER, H. G. (1984). Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **67** 399–432.
KIM, K. H. and BENTLER, P. M. (2002). Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika* **67** 609–623.
LAURITZEN, S. L., SPEED, T. and VIJAYAN, K. (1984). Decomposable graphs and hypergraphs. *J. Aust. Math. Soc.* **36** 12–29.
LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Statist. Soc., Ser. B* **50** 157–194.
LEIGHTON, T. and RAO, S. (1999). Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM* **46** 787–832.
LI, J. and YU, Y. (2015). A nonparametric test of missing completely at random for incomplete multivariate data. *Psychometrika* **80** 707–726.
LITTLE, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *J. Amer. Statist. Assoc.* **83** 1198–1202.
LITTLE, R. J. and RUBIN, D. B. (2019). *Statistical Analysis with Missing Data*. John Wiley & Sons.
LOH, P.-L. and TAN, X. L. (2018). High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electr. J. Statist.* **12** 1429–1467.
LOH, P.-L. and WAINWRIGHT, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.* **40** 1637–1664.
MAIER, D. (1983). *The Theory of Relational Databases*. Computer Science Press, Rockville.
MCMULLEN, P. (1970). The maximum numbers of faces of a convex polytope. *Mathematika* **17** 179–184.
NELSEN, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.
QU, A. and SONG, P. X.-K. (2002). Testing ignorable missingness in estimating equation approaches for longitudinal data. *Biometrika* **89** 841–850.
REEVE, H. W., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Optimal subgroup selection. *arXiv preprint arXiv:2109.01077*.
ROCKAFELLAR, R. T. (1997). *Convex Analysis*. Princeton University Press.
RÜSCHENDORF, L. (2013). *Mathematical Risk Analysis*. Springer.
SPOHN, M.-L., NÄF, J., MICHEL, L. and MEINSHAUSEN, N. (2021). PKLM: A flexible MCAR test using Classification. *arXiv preprint arXiv:2109.10150*.
VLACH, M. (1986). Conditions for the existence of solutions of the three-dimensional planar transportation problem. *Discret. Appl. Math.* **13** 61–78.
VOROBEV, N. N. (1962). Consistent families of measures and their extensions. *Theory Probab. Appl.* **7** 147–163.
WAINWRIGHT, M. J. and JORDAN, M. I. (2003). Variational inference in graphical models: The view from the marginal polytope. In *Proceedings of the Annual Allerton Conference on Communication Control and Computing* **41** 961–971.
WAINWRIGHT, M. J. and JORDAN, M. I. (2008). *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc.
WEI, Y., WAINWRIGHT, M. J. and GUNTUBOYINA, A. (2019). The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *Ann. Statist.* **47** 994–1024.
WU, Y. and YANG, P. (2016). Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory* **62** 3702–3720.
ZHU, Z., WANG, T. and SAMWORTH, R. J. (2022). High-dimensional principal component analysis with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B* **84** 2000–2031.