Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility

Richard J. Samworth

University of Cambridge

'Statistics in the Big Data era' workshop
in honour of Peter J. Bickel

3 June 2022

Collaborator

Tom Berrett

Why care about missing data?

Why care about missing data?

The best solution to handle missing data is to have none.

– R.A. Fisher

Why care about missing data?

The best solution to handle missing data is to have none.

– R.A. Fisher

Consider a complete-case analysis with an $n \times d$ matrix, where each entry is observed independently with probability $p = 0.99$.

Why care about missing data?

The best solution to handle missing data is to have none.

– R.A. Fisher

Consider a complete-case analysis with an $n \times d$ matrix, where each entry is observed independently with probability $p = 0.99$.

- When $d = 5$, around 95% of observations are retained

Why care about missing data?

The best solution to handle missing data is to have none.

– R.A. Fisher

Consider a complete-case analysis with an $n \times d$ matrix, where each entry is observed independently with probability $p = 0.99$.

- When $d = 5$, around 95% of observations are retained
- When $d = 300$, only around 5% of observations are retained.

Why care about missing data?

The best solution to handle missing data is to have none.

– R.A. Fisher

Consider a complete-case analysis with an $n \times d$ matrix, where each entry is observed independently with probability $p = 0.99$.

- When $d = 5$, around 95% of observations are retained
- When $d = 300$, only around 5% of observations are retained.

Missingness represents one of the most common gaps between theory and practice; it can render methodology unreliable or inapplicable.

Missingness mechanisms

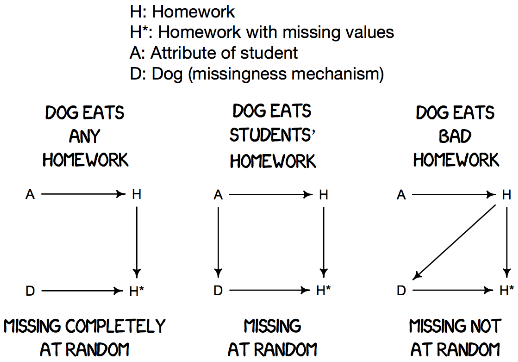


Image credit: Richard McElreath

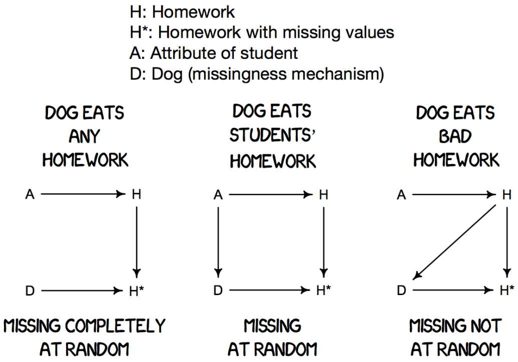


Image credit: Richard McElreath

The simplest setting is where data are MCAR; it makes the analysis is much easier and more interpretable (Loh & Wainwright, 2012; Belloni, Rosenbaum & Tsybakov, 2017; Loh & Tan, 2018; Zhu, Wang & S., 2019; Elsener & van de Geer, 2019; Cai & Zhang, 2019; Follain, Wang & S., 2022).

Formal setting

Given $x = (x_1, \ldots, x_d) \in \prod_{j=1}^d \mathcal{X}_j =: \mathcal{X}$ and $\omega = (\omega_1, \ldots, \omega_d) \in \{0,1\}^d$, define the jth component of $x \circ \omega \in \prod_{j=1}^d (\mathcal{X}_j \cup \{\star\})$ by

$$(x \circ \omega)_j := \begin{cases} x_j & \text{if } \omega_j = 1 \\ \star & \text{if } \omega_j = 0. \end{cases}$$

We observe independent copies of the random vector $X \circ \Omega$, where (X, Ω) takes values in $\mathcal{X} \times \{0,1\}^d$. Our aim is to test the MCAR null hypothesis

$$H_0 : X \perp\!\!\!\perp \Omega.$$

Formal setting

Given $x = (x_1, \ldots, x_d) \in \prod_{j=1}^d \mathcal{X}_j =: \mathcal{X}$ and $\omega = (\omega_1, \ldots, \omega_d) \in \{0,1\}^d$, define the jth component of $x \circ \omega \in \prod_{j=1}^d (\mathcal{X}_j \cup \{\star\})$ by

$$(x \circ \omega)_j := \begin{cases} x_j & \text{if } \omega_j = 1 \\ \star & \text{if } \omega_j = 0. \end{cases}$$

We observe independent copies of the random vector $X \circ \Omega$, where (X, Ω) takes values in $\mathcal{X} \times \{0,1\}^d$. Our aim is to test the MCAR null hypothesis

$$H_0 : X \perp\!\!\!\perp \Omega.$$

Write $\mathbb{S} := \{S \subseteq [d] : \mathbb{P}(\Omega = \mathbb{1}_S) > 0\}$ for the set of possible observation patterns. Let P_S denote the distribution of $X_S := (X_j)_{j \in S}$ conditional on $\Omega = \mathbb{1}_S$, and let $P_{\mathbb{S}} := (P_S : S \in \mathbb{S})$.

Little's test

For *Gaussian* data where *all pairs* of variables are observed together, the EM algorithm can be used to find MLEs for the population mean and covariance matrix.

Little (1988) estimates means within each observation pattern and compares to null MLEs with LR test:

$$d^2 = \sum_{j=1}^{J} m_j (\bar{\mathbf{y}}_{\text{obs},j} - \hat{\boldsymbol{\mu}}_{\text{obs},j}) \tilde{\boldsymbol{\Sigma}}_{\text{obs},j}^{-1} (\bar{\mathbf{y}}_{\text{obs},j} - \hat{\boldsymbol{\mu}}_{\text{obs},j})^T.$$

Fuchs' test

When $\mathcal{X}$ is discrete and complete cases are available ($[d] \in \mathbb{S}$), the EM algorithm can be used to find the MLE for the population distribution.

Fuchs (1982) derived the LR test statistic that compares this to observed counts. With a large number of complete cases its null distribution is approximately χ^2.

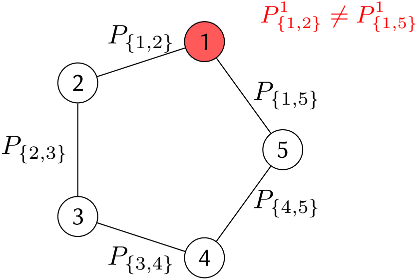
$$G^2 = \sum_i \sum_j \sum_k y_{ijk}{}^{ABC} \ln \frac{y_{ijk}{}^{ABC}}{n_0\, \hat{p}_{ijk}}$$

$$+ 2 \sum_t \sum_{t(ijk)} y_{t(ijk)}{}^{t(ABC)} \ln \frac{y_{t(ijk)}{}^{t(ABC)}}{n_t \sum_{\sim t(ijk)} \hat{p}_{ijk}}$$

Nonparametric tests of consistency

For $S_1, S_2 \in \mathbb{S}$ with $S_1 \cap S_2 \neq \emptyset$, and $\ell \in \{1, 2\}$, let $P_{S_\ell}^{S_1 \cap S_2}$ denote the marginal distribution of P_{S_ℓ} on $\mathcal{X}_{S_1 \cap S_2}$.

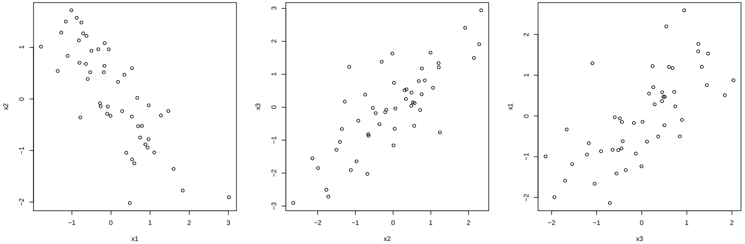
We say that $P_{\mathbb{S}}$ is *consistent* if $P_{S_1}^{S_1 \cap S_2} = P_{S_2}^{S_1 \cap S_2}$ for all $S_1, S_2 \in \mathbb{S}$ with $S_1 \cap S_2 \neq \emptyset$.



We can rule out H_0 if $P_{\mathbb{S}}$ is not consistent, and this motivates two-sample tests of consistency (Li & Yu, 2015; Michel et al., 2021).

Testing consistency is not sufficient

There exist non-MCAR settings where all consistency tests have trivial power.



$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}\right), \quad \begin{pmatrix} X_2 \\ X_3 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad \begin{pmatrix} X_1 \\ X_3 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

We can rule out MCAR if $\rho > 1/2$.

Our methodological aims

We would like to introduce methods that:

- Do not rely on parametric assumptions;
- Can be used for any $\mathbb{S}$, without the need for complete cases (or data on each pair of variables);
- Have power against all detectable alternatives.

Our methodological aims

We would like to introduce methods that:

- Do not rely on parametric assumptions;
- Can be used for any $\mathbb{S}$, without the need for complete cases (or data on each pair of variables);
- Have power against all detectable alternatives.

If $X \sim N(0,1)$ and $\Omega = \mathbb{1}_{\{X \geq 0\}}$, then $X \circ \Omega \overset{\mathrm{d}}{=} X' \circ \Omega'$, where X' and Ω' are independent, with X' having a folded normal distribution and $\Omega' \sim \mathrm{Bern}(1/2)$.

Compatibility

We say $P_{\mathbb{S}}$ is *compatible* if there exists a distribution P on $\mathcal{X}$ whose marginal distribution on $\mathcal{X}_S$ is P_S, for each $S \in \mathbb{S}$.

- If $\mathbb{S} = \big\{\{1\}, \ldots, \{d\}\big\}$, then any $P_{\mathbb{S}}$ is compatible.
- If $[d] \in \mathbb{S}$ then compatibility is equivalent to consistency*.
- If $\mathbb{S} = \big\{\{1,2\}, \{2,3\}, \{1,3\}\big\}$, then consistency is not sufficient for compatibility.

*More generally, compatibility is equivalent to consistency if $\mathbb{S}$ is *decomposable* (Lauritzen & Spiegelhalter, 1988).

Compatibility

We say $P_{\mathbb{S}}$ is *compatible* if there exists a distribution P on $\mathcal{X}$ whose marginal distribution on $\mathcal{X}_S$ is P_S, for each $S \in \mathbb{S}$.

- If $\mathbb{S} = \big\{\{1\}, \ldots, \{d\}\big\}$, then any $P_{\mathbb{S}}$ is compatible.
- If $[d] \in \mathbb{S}$ then compatibility is equivalent to consistency*.
- If $\mathbb{S} = \big\{\{1,2\}, \{2,3\}, \{1,3\}\big\}$, then consistency is not sufficient for compatibility.

Write $\mathcal{P}^0_{\mathbb{S}}$ for the set of compatible $P_{\mathbb{S}}$.

*More generally, compatibility is equivalent to consistency if $\mathbb{S}$ is *decomposable* (Lauritzen & Spiegelhalter, 1988).

Compatibility

If H_0 holds, then $X_S \overset{\mathrm{d}}{=} X_S | \{\Omega = \mathbb{1}_S\} \sim P_S$ for each $S \in \mathbb{S}$, so the distribution of X is compatible.

Compatibility

If H_0 holds, then $X_S \overset{\mathrm{d}}{=} X_S | \{\Omega = \mathbb{1}_S\} \sim P_S$ for each $S \in \mathbb{S}$, so the distribution of X is compatible.

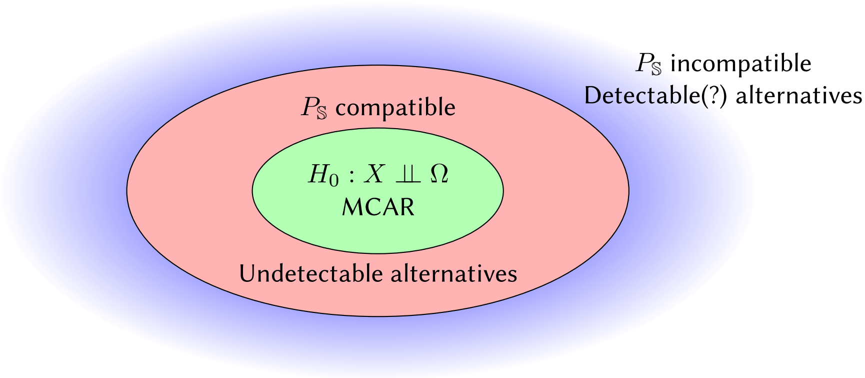
On the other hand, if $P_{\mathbb{S}} \in \mathcal{P}^0_{\mathbb{S}}$, then there exists a distribution P on $\mathcal{X}$ such that, if $\tilde{X} \sim P$ is independent of (X, Ω), then

$$\tilde{X} \circ \Omega \overset{\mathrm{d}}{=} X \circ \Omega.$$

But the distribution of $(\tilde{X}, \Omega)$ satisfies H_0.

Compatibility

$$H_0 \implies P_{\mathbb{S}} \in \mathcal{P}^0_{\mathbb{S}} \quad \text{and} \quad P_{\mathbb{S}} \in \mathcal{P}^0_{\mathbb{S}} \implies \text{cannot rule out } H_0.$$



The best we can do is test the compatibility of $P_{\mathbb{S}}$.

Testing compatibility

We slightly change our model. For fixed $\mathbb{S} \subseteq 2^{[d]}$, distributions $(P_S : S \in \mathbb{S})$ with P_S on $\mathcal{X}_S$, and deterministic sample sizes $n_{\mathbb{S}} := (n_S : S \in \mathbb{S})$ we observe

$$X_{S,1}, \ldots, X_{S,n_S} \overset{\text{iid}}{\sim} P_S \quad \forall S \in \mathbb{S}, \text{ independently.}$$

With this data we aim to test

$$H_0' : P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0.$$

Testing compatibility

We slightly change our model. For fixed $\mathbb{S} \subseteq 2^{[d]}$, distributions $(P_S : S \in \mathbb{S})$ with P_S on $\mathcal{X}_S$, and deterministic sample sizes $n_{\mathbb{S}} := (n_S : S \in \mathbb{S})$ we observe

$$X_{S,1}, \ldots, X_{S,n_S} \overset{\text{iid}}{\sim} P_S \quad \forall S \in \mathbb{S}, \text{ independently.}$$

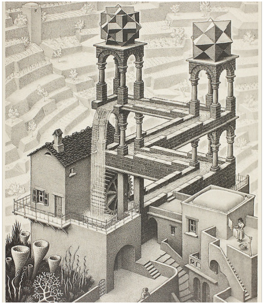
With this data we aim to test

$$H_0' : P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0.$$

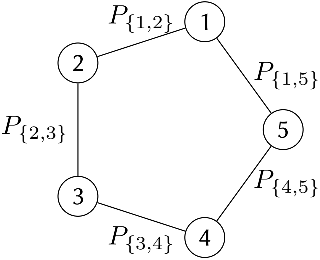
In fact, tests of compatibility are needed in other areas beyond missing data.

Quantum contextuality

'...measurements of quantum observables cannot simply be thought of as revealing pre-existing values' (Wikipedia); see Bell (1966).



M. C. Escher (Cunha, 2019)



Other relevant areas include expert systems (Lauritzen & Spiegelhalter, 1988), meta analysis (Massa & Lauritzen, 2010), relational database theory (Abramsky, 2013) and quantitative risk management (Puccetti & Rüschendorf, 2012).

Kellerer characterisation

Let $\mathcal{G}_{\mathbb{S}}$ be the set of sequences $(f_S : S \in \mathbb{S})$, where $f_S : \mathcal{X}_S \to [-1, \infty)$ is bounded and upper semi-continuous. Take

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

Theorem (Kellerer, 1984). We have $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{0}$ if and only if

$$\sum_{S \in \mathbb{S}} \int_{\mathcal{X}_S} f_S(x_S)\, dP_S(x_S) \geq 0 \text{ for all } f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^{+}.$$

Kellerer characterisation

Let $\mathcal{G}_{\mathbb{S}}$ be the set of sequences $(f_S : S \in \mathbb{S})$, where $f_S : \mathcal{X}_S \to [-1, \infty)$ is bounded and upper semi-continuous. Take

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

Theorem (Kellerer, 1984). We have $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{0}$ if and only if

$$\sum_{S \in \mathbb{S}} \int_{\mathcal{X}_S} f_S(x_S)\, dP_S(x_S) \geq 0 \text{ for all } f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^{+}.$$

This can be regarded as a generalisation of Farkas's lemma (Farkas, 1902), which underpins the theory of linear programming.

Incompatibility index

Definition. Define the *incompatibility index*

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

where

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

Incompatibility index

Definition. Define the *incompatibility index*

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

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 we may take $f_{\mathbb{S}} \equiv 0 \in \mathcal{G}_{\mathbb{S}}^{+}$, we have $R(P_{\mathbb{S}}) \geq 0$, and by Kellerer's characterisation, $R(P_{\mathbb{S}}) = 0$ iff $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{0}$.

Definition. Define the *incompatibility index*

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

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 we may take $f_{\mathbb{S}} \equiv 0 \in \mathcal{G}_{\mathbb{S}}^{+}$, we have $R(P_{\mathbb{S}}) \geq 0$, and by Kellerer's characterisation, $R(P_{\mathbb{S}}) = 0$ iff $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{0}$.

We also have $R(P_{\mathbb{S}}) \leq 1$.

Dual form of incompatibility index

Let $\mathcal{P}_{\mathbb{S}}$ denote the set of all sequences $(P_S : S \in \mathbb{S})$, where P_S is a distribution on $\mathcal{X}_S$.

Theorem. 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 $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$,

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

Dual form of incompatibility index

Let $\mathcal{P}_{\mathbb{S}}$ denote the set of all sequences $(P_S : S \in \mathbb{S})$, where P_S is a distribution on $\mathcal{X}_S$.

Theorem. 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 $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$,

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

When $\mathcal{X}$ is discrete, $R(P_{\mathbb{S}}) < 1$ iff there exists $x \in \mathcal{X}$ with $P_S(\{x_S\}) > 0$ for all $S \in \mathbb{S}$.

A simple test for discrete $\mathcal{X}$

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.

Since $R(P_{\mathbb{S}}, \cdot)$ is linear, we can compute $R(P_{\mathbb{S}})$ using efficient linear programming techniques.

A simple test for discrete $\mathcal{X}$

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.

Since $R(P_{\mathbb{S}}, \cdot)$ is linear, we can compute $R(P_{\mathbb{S}})$ using efficient linear programming techniques.

Letting $\hat{P}_{\mathbb{S}}$ denote the sequence of empirical distributions, our test statistic is $\hat{R} := R(\hat{P}_{\mathbb{S}})$. We propose to reject H_0' at level $\alpha \in (0,1)$ if $\hat{R} \geq C_\alpha$, where

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

A simple test for discrete $\mathcal{X}$

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.

Since $R(P_{\mathbb{S}}, \cdot)$ is linear, we can compute $R(P_{\mathbb{S}})$ using efficient linear programming techniques.

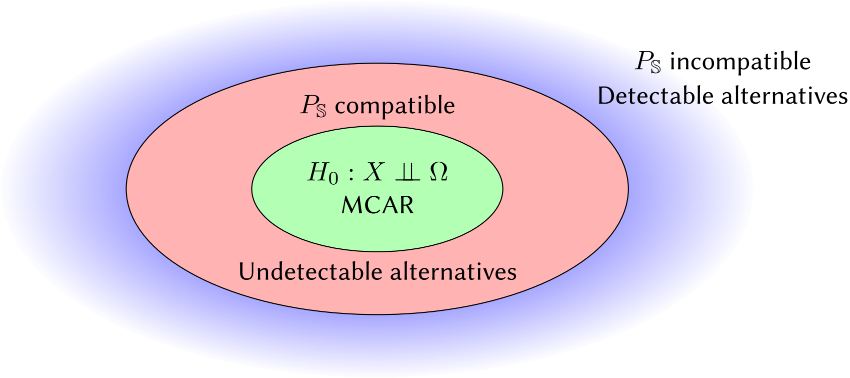
Letting $\hat{P}_{\mathbb{S}}$ denote the sequence of empirical distributions, our test statistic is $\hat{R} := R(\hat{P}_{\mathbb{S}})$. We propose to reject H_0' at level $\alpha \in (0,1)$ if $\hat{R} \geq C_\alpha$, where

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

Proposition. Fix $\alpha, \beta \in (0,1)$. If $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{0}$, then $\mathbb{P}_{P_{\mathbb{S}}}(\hat{R} \geq C_\alpha) \leq \alpha$. Moreover, for any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$ with $R(P_{\mathbb{S}}) \geq C_\alpha + C_\beta$, we have

$$\mathbb{P}_{P_{\mathbb{S}}}(\hat{R} \geq C_\alpha) \geq 1 - \beta.$$

Compatibility and detectability

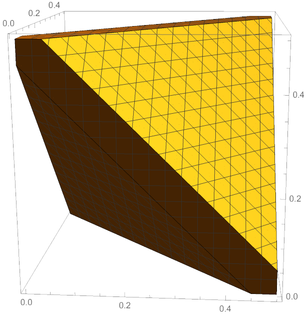


Testing membership of a convex polytope

The null space $\mathcal{P}^0_{\mathbb{S}}$ is the convex hull of the columns of $\mathbb{A} \in \{0,1\}^{\mathcal{X}_{\mathbb{S}} \times \mathcal{X}}$, with

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

In fact, $\mathcal{P}^0_{\mathbb{S}}$ is a full-dimensional subset of $\mathcal{P}^{\mathrm{cons}}_{\mathbb{S}}$, the set of consistent sequences.



Optimal testing over convex polyhedra depends on the specific geometry (Blanchard, Carpentier & Gutzeit, 2018; Wei, Wainwright & Guntuboyina, 2019).

Improved tests

In determining the critical value C_α we used the bound

$$\sup_{f_{\mathbb{S}} \in \mathcal{G}_{\mathbb{S}}^+} R(\hat{P}_{\mathbb{S}}, f_{\mathbb{S}}) \leq \sup_{-1 \leq f_{\mathbb{S}} \leq |\mathbb{S}|-1} R(\hat{P}_{\mathbb{S}}, f_{\mathbb{S}}).$$

This ignores the constraints

$$\min_{x \in \mathcal{X}} (\mathbb{A}^T f_{\mathbb{S}})_x = \min_{x \in \mathcal{X}} \sum_{S \in \mathbb{S}} f_S(x_S) \geq 0.$$

Our strategy is to seek to understand $R(\cdot)$ better, to derive improved tests.

Understanding $R(\cdot)$

Define the *marginal cone* $\mathcal{P}_{\mathbb{S}}^{0,*} := \{\lambda \cdot \mathcal{P}_{\mathbb{S}}^{0} : \lambda \geq 0\}$ and *consistent ball* $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**} := \{\lambda \cdot \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}} : \lambda \in [0,1]\}$.

The Minkowski sum $\mathcal{P}_{\mathbb{S}}^{0,*} + \mathcal{P}_{\mathbb{S}}^{\mathrm{cons},**}$ is a convex polyhedral subset of $[0,\infty)^{\mathcal{X}_{\mathbb{S}}}$, so let F denote its number of *essential facets* (i.e. ignoring non-negativity conditions).

Proposition. There exist $f_{\mathbb{S}}^{(1)}, \ldots, f_{\mathbb{S}}^{(F)} \in \mathcal{G}_{\mathbb{S}}^{+}$, depending only on $\mathbb{S}$ and $\mathcal{X}_{\mathbb{S}}$, such that for $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$,

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

More generally, for any $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$,

$$R(P_{\mathbb{S}}) \asymp_{\mathbb{S}} \max_{\ell \in [F]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{(\ell)})_{+} + \max_{S_1, S_2 \in \mathbb{S}} d_{\mathrm{TV}}\bigl(P_{S_1}^{S_1 \cap S_2}, P_{S_2}^{S_1 \cap S_2}\bigr).$$

Improved test

If F is known, then we can choose a critical value

$$C'_\alpha \asymp_{\mathbb{S}} \frac{\log(F/\alpha)}{\min_{S\in\mathbb{S}} n_S} + \max_{S_1\neq S_2, S_1\cap S_2\neq\emptyset} \frac{|\mathcal{X}_{S_1\cap S_2}|}{n_{S_1}\wedge n_{S_2}}.$$

Proposition. Fix $\alpha, \beta \in (0,1)$. If $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^0$, then $\mathbb{P}_{P_{\mathbb{S}}}(\hat{R} \geq C'_\alpha) \leq \alpha$. Moreover, there exists $M \equiv M(\mathbb{S}) > 0$ such that whenever $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$ has

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

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

Taking $\min(C_\alpha, C'_\alpha)$ as the critical value gives the best of both worlds.

Improved test

If F is known, then we can choose a critical value

$$C'_\alpha \asymp_{\mathbb{S}} \frac{\log(F/\alpha)}{\min_{S\in\mathbb{S}} n_S} + \max_{S_1\neq S_2, S_1\cap S_2\neq\emptyset} \frac{|\mathcal{X}_{S_1\cap S_2}|}{n_{S_1}\wedge n_{S_2}}.$$

Proposition. Fix $\alpha, \beta \in (0,1)$. If $P_{\mathbb{S}} \in \mathcal{P}^0_{\mathbb{S}}$, then $\mathbb{P}_{P_{\mathbb{S}}}(\hat{R} \geq C'_\alpha) \leq \alpha$. Moreover, there exists $M \equiv M(\mathbb{S}) > 0$ such that whenever $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}$ has

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

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

Taking $\min(C_\alpha, C'_\alpha)$ as the critical value gives the best of both worlds.

$d = 3$ example

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

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

where, e.g., $p_{AB\bullet} := P_{\{1,2\}}(A \times B)$. Moreover,

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

In particular, we may take $F = (2^r - 2)(2^s - 2)$. In this case, when $n_{\{1,2\}} = n_{\{2,3\}} = n_{\{1,3\}} = n/3$, we have

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

Proof idea

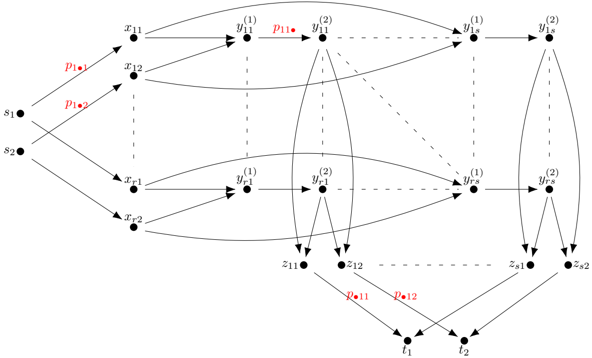
Lower bound via primal problem $R(P_{\mathbb{S}}) \geq \max_{A \subseteq [r], B \subseteq [s]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{A,B})$.

Proof idea

Lower bound via primal problem $R(P_{\mathbb{S}}) \geq \max_{A \subseteq [r], B \subseteq [s]} R(P_{\mathbb{S}}, f_{\mathbb{S}}^{A,B})$.

Upper bound via dual, relating $R(P_{\mathbb{S}})$ to a maximal two-commodity flow:

$$R(P_{\mathbb{S}}) = 1 - \max\left\{ \sum_{i,j,k} p_{ijk} : \sum_{i=1}^{r} p_{ijk} \leq p_{\bullet jk}, \sum_{j=1}^{s} p_{ijk} \leq p_{i\bullet k}, p_{ij1} + p_{ij2} \leq p_{ij\bullet} \right\}.$$



Minimax testing framework

Given $\rho \in [0,1]$, it is convenient to write

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

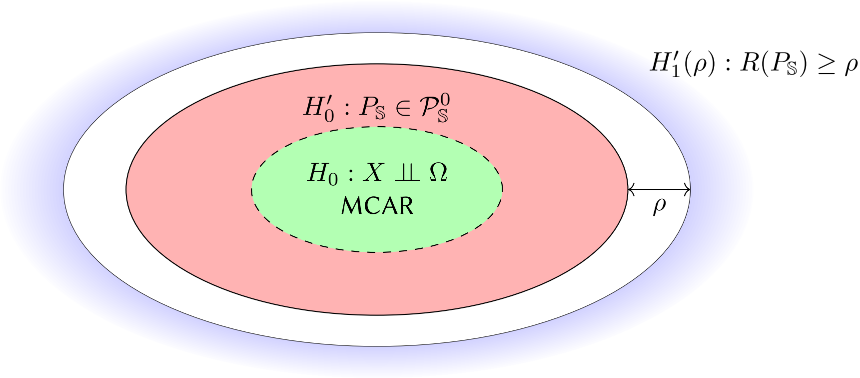
so that $\mathcal{P}_{\mathbb{S}}(0) = \mathcal{P}_{\mathbb{S}}$ and $\mathcal{P}_{\mathbb{S}}^0 = \mathcal{P}_{\mathbb{S}} \setminus \cup_{\rho \in (0,1]} \mathcal{P}_{\mathbb{S}}(\rho)$. The minimax risk at separation ρ in this problem is defined as

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

Finally, the minimax testing radius is defined as

$$\rho^*(n_{\mathbb{S}}) := \inf\bigl\{\rho \in [0,1] : \mathcal{R}(n_{\mathbb{S}}, \rho) \leq 1/2\bigr\}.$$

Illustration of framework



$d = 3$ example again

Theorem. Let $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ and $\mathcal{X} = [r] \times [s] \times [2]$ for $r, s \geq 2$. Then

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

Moreover, when $n_{\{1,2\}} \geq (r+s)\log(r+s)$, $n_{\{1,3\}} \geq r\log r$ and $n_{\{2,3\}} \geq s\log s$ we have a minimax lower bound:

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

The sequences of distributions in the lower bound contruction belong to $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$, so the same lower bound holds for testing against consistent alternatives.

Theorem. Let $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$ and $\mathcal{X} = [r] \times [s] \times [2]$ for $r, s \geq 2$. Then

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

Moreover, when $n_{\{1,2\}} \geq (r+s)\log(r+s)$, $n_{\{1,3\}} \geq r \log r$ and $n_{\{2,3\}} \geq s \log s$ we have a minimax lower bound:

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

The sequences of distributions in the lower bound contruction belong to $\mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$, so the same lower bound holds for testing against consistent alternatives.

Reductions (I)

For other $(\mathbb{S}, \mathcal{X})$, analytic expressions for $R(P_{\mathbb{S}})$ can be difficult, but we can sometimes reduce to simpler problems.

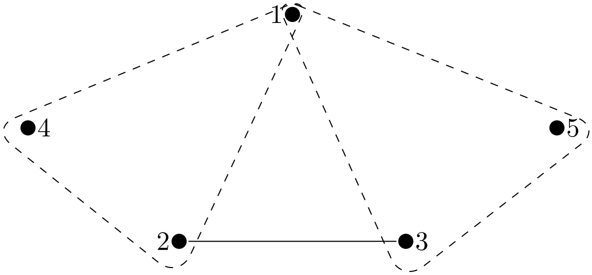
Reductions (I)

For other $(\mathbb{S}, \mathcal{X})$, analytic expressions for $R(P_{\mathbb{S}})$ can be difficult, but we can sometimes reduce to simpler problems.

If there exists $J \subseteq [d]$ and $S_0 \in \mathbb{S}$ with $J \subseteq S_0$ and $J \cap S = \emptyset$ for all $S \in \mathbb{S} \setminus \{S_0\}$, then

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

E.g., $\mathbb{S} = \{\{1,2,4\},\{2,3\},\{1,3,5\}\}$ reduces to $\mathbb{S} = \{\{1,2\},\{2,3\},\{1,3\}\}$.



Reductions (II)

If there exists $J \subseteq [d]$ such that $J \subseteq S$ and $P_S^J = P^J$ for all $S \in \mathbb{S}$, then

$$R(P_{\mathbb{S}}) = \sum_{x_J \in \mathcal{X}_J} R(P_{\mathbb{S}|X_J = x_J}) p^J(x_J)$$

when $\mathcal{X}$ is discrete.

Reductions (II)

If there exists $J \subseteq [d]$ such that $J \subseteq S$ and $P_S^J = P^J$ for all $S \in \mathbb{S}$, then

$$R(P_{\mathbb{S}}) = \sum_{x_J \in \mathcal{X}_J} R(P_{\mathbb{S}|X_J = x_J}) p^J(x_J)$$

when $\mathcal{X}$ is discrete.

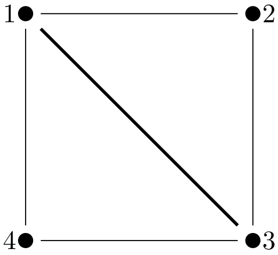
E.g., if $\mathbb{S} = \big\{\{1,2,3\},\{1,3,4\},\{1,2,4\}\big\}$ with $\mathcal{X} = [r] \times [s] \times [t] \times [2]$, then for $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$,

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

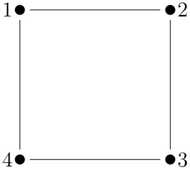
Reductions (III)

If $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{S}$ are such that there exists $J \in \mathbb{S}$ with $\mathbb{S}_1 \cap \mathbb{S}_2 = \{J\}$ and $(\cup_{S \in \mathbb{S}_1} S) \cap (\cup_{S \in \mathbb{S}_2} S) = J$, then

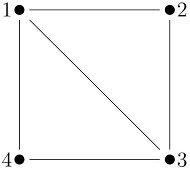
$$\max\{R(P_{\mathbb{S}_1}), R(P_{\mathbb{S}_2})\} \leq R(P_{\mathbb{S}}) \leq R(P_{\mathbb{S}_1}) + R(P_{\mathbb{S}_2}).$$



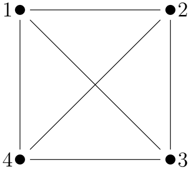
Irreducible $d = 4$ examples



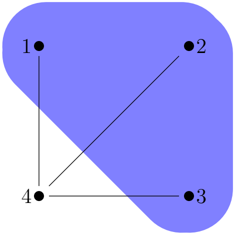
(a) Chain pairs



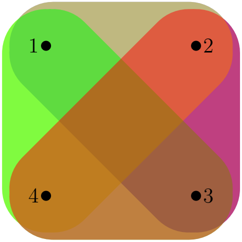
(b) All pairs except one



(c) All pairs



(d) Single triple



(e) All triples

Mixed discrete and continuous variables

By binning continuous variables we can apply our tests designed for the discrete setting.

In particular, when $\mathcal{X} = [0,1)^2 \times \{1,2\}$ and the densities on $\mathcal{X}_j$ are (r_j, L)-Hölder smooth, with $r_j \in (0,1]$ for $j = 1,2$,

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

Numerical results

Our tests have uniform, finite-sample Type I error control, but could be conservative. An alternative, Monte Carlo test appears to perform well in practice.

For $|\mathcal{X}| < \infty$, we can solve the dual program for $R(\hat{P}_{\mathbb{S}})$ to find a decomposition

$$\hat{P}_{\mathbb{S}} = \{1 - R(\hat{P}_{\mathbb{S}})\}\hat{Q}_{\mathbb{S}} + R(\hat{P}_{\mathbb{S}})\hat{T}_{\mathbb{S}} \in \{1 - R(\hat{P}_{\mathbb{S}})\}\mathcal{P}_{\mathbb{S}}^{0} + R(\hat{P}_{\mathbb{S}})\mathcal{P}_{\mathbb{S}}.$$

Here $\hat{Q}_{\mathbb{S}}$ can be thought of as a closest compatible sequence of marginal distributions to $\hat{P}_{\mathbb{S}}$.

We can generate bootstrap empirical distributions $\hat{Q}_{\mathbb{S}}^{(1)}, \ldots, \hat{Q}_{\mathbb{S}}^{(B)}$ from $\hat{Q}_{\mathbb{S}}$ and reject H_0' if and only if

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

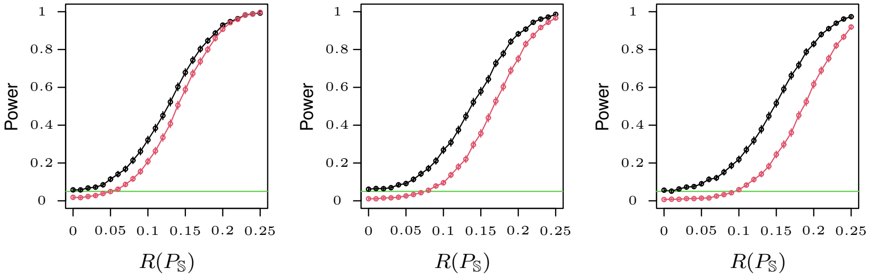
Numerical results

We compare with Fuchs's LR test. For $\mathbb{S} = \big\{\{1,2\},\{2,3\},\{1,3\}\big\}$, with $\mathcal{X} = [r] \times [2]^2$ for $r \in \{2,4,6\}$ and with $P_{\mathbb{S}} \in \mathcal{P}_{\mathbb{S}}^{\mathrm{cons}}$ defined by

$$p_{i\bullet\bullet} = \frac{1}{r}, \;\; p_{\bullet 1\bullet} = p_{\bullet\bullet 1} = \frac{1}{2}, \;\; p_{i\bullet 1} = \frac{1}{2r}, p_{i\bullet 1} = \frac{1+(-1)^i}{2r}$$

and $p_{\bullet 21} \in [0.25, 0.375]$, we take $n_{\mathbb{S}} = (200, 200, 200)$, $B = 99$, $\alpha = 0.05$.

Fuchs's test requires complete cases, so we allow it access to 200 observations from a closest compatible sequence to $P_{\mathbb{S}}$.



Numerical results

Now take $d = 5$, $\mathcal{X} = [2]^5$ and

$$\mathbb{S} = \big\{\{1,2,3,4\},\{1,2,3,5\},\{1,2,4,5\},\{1,3,4,5\},\{2,3,4,5\}\big\}.$$

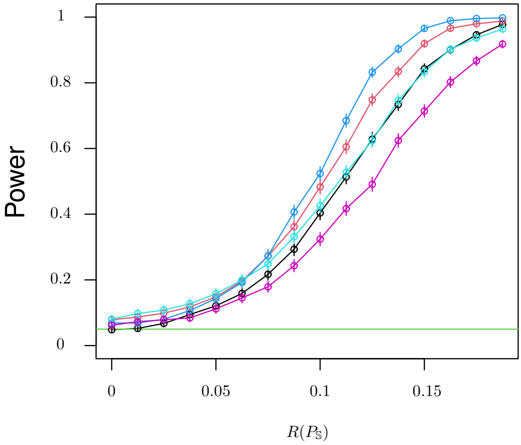
For $\epsilon \in [0.2, 0.35]$ and $i, j, k, \ell, m \in [2]$, we set

$$\begin{aligned} p_{ijk\ell\bullet} = p_{ijk\bullet\ell} = p_{ij\bullet k\ell} = p_{i\bullet jk\ell} &= \frac{1 + \epsilon(-1)^{i+j+k+\ell}}{16}, \\ p_{\bullet ijk\ell} &= \frac{1 - \epsilon(-1)^{i+j+k+\ell}}{16}, \end{aligned}$$

for which $R(P_{\mathbb{S}}) = (5\epsilon - 1)_+/4$.

Numerical results

Allow Fuchs's test $\{25, 50, 100, 200\}$ complete cases. Our test is in black.



Summary

- Testing MCAR is equivalent to testing compatibility;
- We propose a general test with asymptotic power 1 against fixed alternatives for discrete/discretisable data;
- Improved tests are possible given knowledge of underlying geometry (and are rate-optimal in certain cases);
- A Monte Carlo critical value yields good empirical power.

Berrett, T. B. and Samworth, R. J. (2022) Optimal nonparametric testing of Missing Completely At Random, and its connections to compatibility. `arXiv:2205.08627`.

R package: `MCARtest`.

‘Congratulations, Peter B’; ‘Thanks, Peter B’

Happy birthday, Peter!

References

Abramsky, S. (2013) Relational databases & Bell's theorem. *In Search of Elegance in the Theory & Practice of Computation*, 13–15.

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

Belloni, A., Rosenbaum, M. and Tsybakov, A. B. (2017) Linear and conic programming estimators in high dimensional errors-in-variables models. *J. Roy. Statist. Soc., Ser. B*, **79**, 939–956.

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

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

Cunha, M. T. (2019) On measures & measurements: a fibre bundle approach to contextuality. *Philos. Trans. R. Soc.* **377.2157**, 20190146.

Elsener, A. & van de Geer, S. (2019) Sparse spectral estimation with missing and corrupted measurements. *Stat*, **8**, e229.

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

Follain, B., Wang, T. & Samworth, R. J. (2022) High-dimensional changepoint estimation with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B, to appear.*

Fuchs, C. (1982) Maximum likelihood estimation & model selection in contingency tables with missing data. *J. Amer. Statist. Assoc.*, **77**, 270–278.

Kellerer, H. G. (1984) Duality theorems for marginal problems. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **67**, 399–432.

Lauritzen, S. L. & 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.

References

Li, J. & 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.

Loh, P.-L. & Tan, X. L. (2018) High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electr. J. Statist.*, **12**, 1429–1467.

Loh, P.-L. & Wainwright, M. J. (2012) High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, **40**, 1637–1664.

Massa, M. S. & Lauritzen, S. L. (2010) Combining statistical models. *Contemporary Mathematics: Algebraic Methods in Statistics & Probability II*, 239–260.

Michel, L., Näf, J., Spohn, M.-L. & Meinshausen, N. (2021) PKLM: A flexible MCAR test using Classification. *arXiv preprint arXiv:2109.10150.*

Puccetti, G. & Rüschendorf, L. (2012) Bounds for joint portfolios of dependent risks. *Statistics & Risk Modeling*, **29**, 107–132.

Wei, Y., Wainwright, M. J. & Guntuboyina, A. (2019) The geometry of hypothesis testing over convex cones: Generalized likelihood ratio tests and minimax radii. *Ann. Statist.*, **47**, 994–1024.

Zhu, Z., Wang, T. & Samworth, R. J. (2019) High-dimensional principal component analysis with heterogeneous missingness. *arXiv preprint arXiv:1906.12125.*