Diagnosing and Correcting Balance Assessments in High Dimensions

Mark M. Fredrickson (mfredric@umich.edu)
Joint work with Ben Hansen

2020-01-07

13th International Conference on Health Policy Statistics

Balance Assessments

In both **randomized controlled trials** and **observational studies**, unconfounded analysis depends on the treatment and control groups having the **equivalent distributions on background variables x**.

Balance Assessments

In both **randomized controlled trials** and **observational studies**, unconfounded analysis depends on the treatment and control groups having the **equivalent distributions on background variables x**.

In randomized trials, this condition is only **guaranteed in expectation**. In observational studies, researchers attempt to **enforce this condition with matching or weighting**.

Balance Assessments

In both **randomized controlled trials** and **observational studies**, unconfounded analysis depends on the treatment and control groups having the **equivalent distributions on background variables x**.

In randomized trials, this condition is only **guaranteed in expectation**. In observational studies, researchers attempt to **enforce this condition with matching or weighting**.

In either case, **assessing the similarity of treatment and control groups** ("balance") is an important component of study design.

Notation

- There are n **units** (people, clinics, schools)
- The p **variables** are collected in $\mathbf{x}$
- Each unit is assigned to **treatment** $Z_i = 1$ or **control** $Z_i = 0$.
- It will be useful to define

$$J_i = \frac{Z_i - \mathrm{P}(Z_i = 1)}{\mathrm{P}(Z_i = 1)\,\mathrm{P}(Z_i = 0)}$$

Mahalanobis Distance

Hansen and Bowers (2008) and Morgan and Rubin (2012) assess balance with a **Mahalanobis distance**, a normalized difference of group means $\mathbf{D}$:

$$M = \mathbf{D}'\text{Cov}\left(\mathbf{D}\right)^{-}\mathbf{D} = \mathbf{J}'\mathbf{x}\,\text{Cov}\left(\mathbf{x}'\mathbf{J}\right)^{-}\mathbf{x}'\mathbf{J}$$

Mahalanobis Distance

Hansen and Bowers (2008) and Morgan and Rubin (2012) assess balance with a **Mahalanobis distance**, a normalized difference of group means $\mathbf{D}$:

$$M = \mathbf{D}' \text{Cov}\left(\mathbf{D}\right)^{-} \mathbf{D} = \mathbf{J}'\mathbf{x}\, \text{Cov}\left(\mathbf{x}'\mathbf{J}\right)^{-}\, \mathbf{x}'\mathbf{J}$$

A common special case is **complete random assignment** with n_1 assigned to treatment and $n_0 = n - n_1$ to control:

$$M = \frac{n_1 n_0}{n}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)' \, S^2(\mathbf{x})^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)$$

where $\mathbf{X}_1$ and $\mathbf{X}_0$ are group means and $S^2(\mathbf{x})$ is the sample covariance matrix.

Interpretation via PCA

Since we are **scaling by the inverse covariance matrix** there is a convenient representation of the Mahalanobis distance using **the principal components of $\mathbf{x}$**:

$$M = \mathbf{J}'\mathbf{x}(\mathbf{x}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')\,\mathbf{x})^{-}\mathbf{x}'\mathbf{J} = \mathbf{J}'\mathbf{u}\mathbf{u}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')^{-}\,\mathbf{u}\mathbf{u}'\mathbf{J},$$

where $\mathbf{u}$ is from the SVD of $\mathbf{x} = \mathbf{u}\mathbf{d}\mathbf{v}'$.

Interpretation via PCA

Since we are **scaling by the inverse covariance matrix** there is a convenient representation of the Mahalanobis distance using **the principal components of $\mathbf{x}$**:

$$M = \mathbf{J}'\mathbf{x}(\mathbf{x}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')\,\mathbf{x})^{-}\mathbf{x}'\mathbf{J} = \mathbf{J}'\mathbf{u}\mathbf{u}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')^{-}\,\mathbf{u}\mathbf{u}'\mathbf{J},$$

where $\mathbf{u}$ is from the SVD of $\mathbf{x} = \mathbf{u}\mathbf{d}\mathbf{v}'$.

Going forward, we'll think of p **as the rank of** $\mathbf{u}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')\,\mathbf{u}$.

Interpretation via PCA

Since we are **scaling by the inverse covariance matrix** there is a convenient representation of the Mahalanobis distance using **the principal components of $\mathbf{x}$**:

$$M = \mathbf{J}'\mathbf{x}(\mathbf{x}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')\,\mathbf{x})^{-}\mathbf{x}'\mathbf{J} = \mathbf{J}'\mathbf{u}\mathbf{u}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')^{-}\,\mathbf{u}\mathbf{u}'\mathbf{J},$$

where $\mathbf{u}$ is from the SVD of $\mathbf{x} = \mathbf{u}\mathbf{d}\mathbf{v}'$.

Going forward, we'll think of p **as the rank of** $\mathbf{u}'\,\mathrm{E}\,(\mathbf{J}\mathbf{J}')\,\mathbf{u}$.

As an added bonus, we now have a natural way of **ordering the covariates**.

Inference when $n > p$

In an **observational trial** M is a useful **test statistic** for $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ (Hansen and Bowers, 2008).

Inference when $n > p$

In an **observational trial** M is a useful **test statistic** for $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ (Hansen and Bowers, 2008).

In **randomized trials**, M can be used to **select particular Z** that have controlled imbalance (Morgan and Rubin, 2012).

Inference when $n > p$

In an **observational trial** M is a useful **test statistic** for $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$ (Hansen and Bowers, 2008).

In **randomized trials**, M can be used to **select particular Z** that have controlled imbalance (Morgan and Rubin, 2012).

With the number of covariates p **being fixed** and n **tending to infinity**,

$$M \xrightarrow{\text{D}} \chi^2(p)$$

(Hansen and Bowers, 2008; Li et al., 2018)

Example: Homicide rates and public infrastructure

Cerdá et al. (2012) reported an **natural experiment** in Medellín, Colombia:

- Intervention: Public works **transportation program**.

Example: Homicide rates and public infrastructure

Cerdá et al. (2012) reported an **natural experiment** in Medellín, Colombia:

- Intervention: Public works **transportation program**.
- Units: **48 neighborhoods** (25 with stations, 23 without) matched in **pairs and two triples**.

Example: Homicide rates and public infrastructure

Cerdá et al. (2012) reported an **natural experiment** in Medellín, Colombia:

- Intervention: Public works **transportation program**.
- Units: **48 neighborhoods** (25 with stations, 23 without) matched in **pairs and two triples**.
- Covariates: Survey responses, governmental data, mix of types (**48 with interactions**)

Example: Homicide rates and public infrastructure

Cerdá et al. (2012) reported an **natural experiment** in Medellín, Colombia:

- Intervention: Public works **transportation program**.
- Units: **48 neighborhoods** (25 with stations, 23 without) matched in **pairs and two triples**.
- Covariates: Survey responses, governmental data, mix of types (**48 with interactions**)
- Balance Assessment: Validating matching strategy created **comparable pairs and triples**.

Cerdá et al. (2012) Empirical Mahalanobis Distance

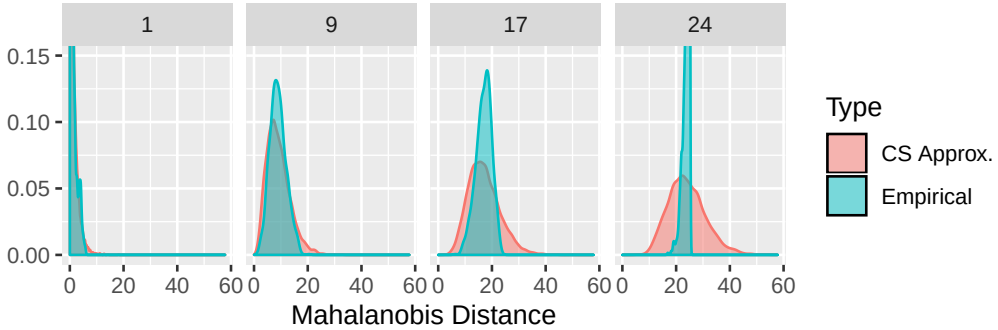


Figure 1: Mahalanobis distance distribution for 48 matched neighborhoods in reported in Cerdá et al. (2012)

Reduced Rank Mahalanobis Distance

A **degenerate** M provides no information on **suitability of observational studies** or **equivalence at baseline in randomize trials**.

Reduced Rank Mahalanobis Distance

A **degenerate** M provides no information on **suitability of observational studies** or **equivalence at baseline in randomize trials**.

As the previous slides demonstrated, **selecting** $k < n$ **principal components** provides a **non-degenerate distribution**.

Reduced Rank Mahalanobis Distance

A **degenerate** M provides no information on **suitability of observational studies** or **equivalence at baseline in randomize trials**.

As the previous slides demonstrated, **selecting** $k < n$ **principal components** provides a **non-degenerate distribution**.

We define **reduced rank Mahalanobis distance** as:

$$M_k = \mathbf{J}'\mathbf{u}_k\mathbf{u}_k'\,\mathrm{E}\left(\mathbf{J}\mathbf{J}'\right)^{-}\mathbf{u}_k\mathbf{u}_k'\mathbf{J}$$

where $\mathbf{u}_k$ collects the first k columns of $\mathbf{u}$.

Picking k

Typical rules for picking k (e.g., including 80% of variance in $\mathbf{x}$) do not take into account the role of assignment mechanism $\mathbf{Z}$ (Chang, 1983).

Picking k

Typical rules for picking k (e.g., including 80% of variance in $\mathbf{x}$) do not take into account the role of assignment mechanism $\mathbf{Z}$ (Chang, 1983).

A natural analog in this case is **maximizing the (null) variance of M_k**.

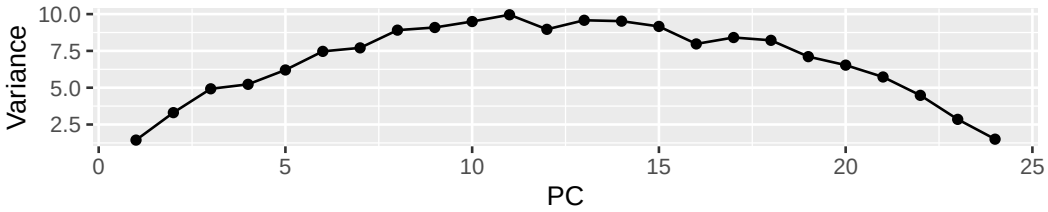


Figure 2: Variance of M_k for matched neighborhoods in Medellín

Satterthwaite Approximation

Rather than approximate the distribution of M_k as $\chi^2(k)$, we could use a correction that matches both **mean and variance** to a **scaled** χ^2:

$$M_k \approx a\chi^2(v) \Rightarrow v = \frac{2k^2}{\text{Var}\,(M_k)},\ a = \frac{k}{v}$$

Satterthwaite Approximation

Rather than approximate the distribution of M_k as $\chi^2(k)$, we could use a correction that matches both **mean and variance** to a **scaled** χ^2:

$$M_k \approx a\chi^2(v) \Rightarrow v = \frac{2k^2}{\text{Var}\,(M_k)},\ a = \frac{k}{v}$$

- Provides approximation to tail probabilities $\text{P}\,(M_k \geq m_k)$

Satterthwaite Approximation

Rather than approximate the distribution of M_k as $\chi^2(k)$, we could use a correction that matches both **mean and variance** to a **scaled** χ^2:

$$M_k \approx a\chi^2(v) \Rightarrow v = \frac{2k^2}{\text{Var}\left(M_k\right)},\ a = \frac{k}{v}$$

- Provides approximation to tail probabilities $\text{P}\left(M_k \geq m_k\right)$
- Provides approximate cutoff for selecting "rerandomization" RCT designs

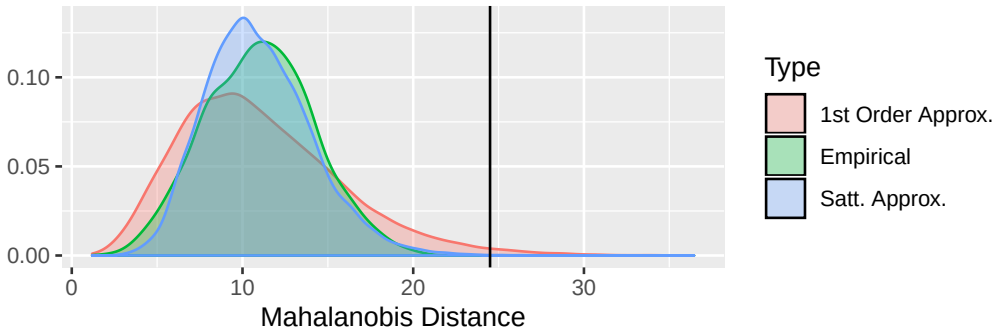
Satterthwaite Approximation

Rather than approximate the distribution of M_k as $\chi^2(k)$, we could use a correction that matches both **mean and variance** to a **scaled** χ^2:

$$M_k \approx a\chi^2(v) \Rightarrow v = \frac{2k^2}{\text{Var}\,(M_k)},\ a = \frac{k}{v}$$

- Provides approximation to tail probabilities $\text{P}\,(M_k \geq m_k)$
- Provides approximate cutoff for selecting "rerandomization" RCT designs
- Alternatively, use Monte Carlo sampling from $\mathbf{Z}$

Corrected Medellín Distribution



$m_{11} = 24.5$; $\mathrm{P}\left(M_{11} \geq m_{11}\right)$: 0.01088 (1st order approx.), 0.00048 (2nd order approx.), 0.001 (empirical, 5k samples)

Summary and Future Directions

- **Balance assessments** compare treated and control groups for similar distributions.

Summary and Future Directions

- **Balance assessments** compare treated and control groups for similar distributions.
- The **Mahalanobis distance statistic** is useful when $n \gg p$, but **becomes degenerate when** $p \approx n$.

Summary and Future Directions

- **Balance assessments** compare treated and control groups for similar distributions.
- The **Mahalanobis distance statistic** is useful when $n \gg p$, but **becomes degenerate when** $p \approx n$.
- **Reduced rank Mahalanobis distance** guarantees non-degenerate distribution.

Summary and Future Directions

- **Balance assessments** compare treated and control groups for similar distributions.
- The **Mahalanobis distance statistic** is useful when $n \gg p$, but **becomes degenerate when** $p \approx n$.
- **Reduced rank Mahalanobis distance** guarantees non-degenerate distribution.
- **Satterthwaite χ^2 approximation** superior and relatively simple.

Summary and Future Directions

- **Balance assessments** compare treated and control groups for similar distributions.
- The **Mahalanobis distance statistic** is useful when $n \gg p$, but **becomes degenerate when** $p \approx n$.
- **Reduced rank Mahalanobis distance** guarantees non-degenerate distribution.
- **Satterthwaite χ^2 approximation** superior and relatively simple.
- **In-progress**: extend `RItools` package for R, formalize justification for variance maximization, implications for design (both observational and RCTs)

Thank You!