Robust Weighting and Matching Techniques for Causal Inference in Observational Studies with Continuous Treatment

Universität Stuttgart

Universität Stuttgart

Ioan Scheffel

December 18, 2022

Contents

1 Introduction

2 Balancing Weights

2.1 Introduction

2.2 Estimating the Population Mean of Potential Outcomes

2.3 Application of Convex Optimization

Assumption 2.1. *Assume that the map $f : \mathbb{R} \to \overline{\mathbb{R}}$ has the following properties.*

(i) *f is strictly convex.*

(ii) *f is lower-semicontinuous and continuously differentiable on $\mathrm{int}(\mathrm{dom}(f))$.*

(iii) *The derivative of f on $\mathrm{int}(\mathrm{dom}(f))$ is a diffeomorphism.*

(iv) *The Legendre transformation f^* of f is finite.*

(v) *The function $x \mapsto xt - f(x)$ takes its supremum on $\mathrm{int}(\mathrm{dom}(f))$ for all $t \in \mathbb{R}$.*

We consider the following optimization problem.

Problem 2.1.

$$\underset{w_1,\ldots,w_n \in \mathbb{R}}{\text{minimize}} \qquad \sum_{i=1}^{n} T_i f(w_i)$$

subject to the constraints

$$w_i T_i \geq 0, \qquad i = 1, \ldots, n,$$

$$\sum_{i=1}^{n} w_i T_i = 1$$

$$\left| \sum_{i=1}^{n} w_i T_i B_k(X_i) - \frac{1}{n} \sum_{i=1}^{n} B_k(X_i) \right| \leq \delta_k, \qquad k = 1, \ldots, K$$

Theorem 2.1. *Under Assumption, the dual of the above Problem is the unconstrained optimization problem*

$$\underset{\lambda \in \mathbb{R}^K}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} n T_i f^*(\langle B(X_i), \lambda \rangle) - \langle B(X_i), \lambda \rangle + \langle \delta, |\lambda| \rangle,$$

where $t \mapsto f^*(t) = t \left(f'\right)^{-1}(t) - f\left(\left(f'\right)^{-1}(t)\right)$ is the Legendre transformation of f, $B(X_i) = [B_1(X_i), \ldots, B_K(X_i)]^\top$ denotes the K basis functions of the covariates of unit $i \in \{1, \ldots, n\}$ and $|\lambda| = [|\lambda_1|, \ldots, |\lambda_K|]^\top$, where $|\cdot|$ is the absolute value of a real-valued scalar. Moreover, if $\lambda^\dagger$ is an optimal solution then

$$w_i^* = (f')^{-1}(\langle B(X_i), \lambda^\dagger \rangle), \quad i \in \{1, \ldots, n\} \tag{2.1}$$

are the unique optimal solutions to (P) .

Proof. We prove the following Lemma at the end of the section.

Lemma 2.1. *The dual of the optimization problem is*

$$\underset{\lambda \in \mathbb{R}^{2K}}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^{n} n T_i f^*(\langle Q_{\bullet i}, \lambda \rangle) - \langle Q_{\bullet i}, \lambda \rangle + \langle d, \lambda \rangle$$

subject to

$$\lambda_k \geq 0 \quad \text{for all } k \in \{1, \ldots, K\}, \tag{2.2}$$

where

$$\mathbf{Q} := \begin{bmatrix} \mathbf{I}_n \\ \mathbf{B}(\mathbf{X}) \\ -\mathbf{B}(\mathbf{X}) \end{bmatrix}, \qquad \mathbf{B}(\mathbf{X}) := \begin{bmatrix} B(X_1), \ldots, B(X_n) \end{bmatrix}, \qquad \text{and} \qquad d := \begin{bmatrix} 0_n \\ \delta \\ \delta \end{bmatrix}. \tag{2.3}$$

$\square$

2.4 Application of Matrix Concentration Inequalities

3 Convex Analysis

3.1 Basic Notions

3.2 Relative Interior

3.3 Conjugate Calculus

3.4 Tseng Bertsekas

We present the relevant parts of the paper [BT03].
    Consider the following optimization problem

$$\underset{x \in \mathbb{R}^m}{\text{minimize}} \quad f(x)$$

subject to the constraints

$$\mathbf{A}x \geq b, \tag{3.1}$$

Where $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, $\mathbf{A}$ is a given $n \times m$ matrix, and b is a vector in $\mathbb{R}^n$.

Assumption 3.1. *Assume that the map $f : \mathbb{R}^m \to \overline{\mathbb{R}}$ has the following properties.*

(i) *f is strictly convex.*

(ii) *f is lower-semicontinuous and continuous $\text{dom}(f)$.*

(iii) *The convex conjugate f^* of f is finite.*

4 Random Matrix Inequalities

4.1 Matrix Analysis

The **trace** of a square matrix, denoted by tr, is the sum of its diagonal entries, i.e. $\mathrm{tr}(\mathbf{B}) = \sum_{j=1}^{d} b_{jj}$ for $\mathbf{B} \in \mathbb{M}_d$. The trace is unitarily invariant, i.e. $\mathrm{tr}(\mathbf{B}) = \mathrm{tr}(\mathbf{Q}\mathbf{B}\mathbf{Q}^*)$ for all $\mathbf{B} \in \mathbb{M}_d$ for all unitary $\mathbf{Q} \in \mathbb{M}_d$. In particular, the existence of an eigenvalue value decomposition shows that the trace of a Hermitian matrix equals the sum of its eigenvalues. Let $f : I \to \mathbb{R}$ where $I \subseteq \mathbb{R}$ is an interval. Consider a matrix $\mathbf{A} \in \mathbb{H}_d$ whose eigenvalues are contained in I. We define the matrix $f(\mathbf{A}) \in \mathbb{H}_d$ using an eigenvalue decomposition of $\mathbf{A}$:

$$f(\mathbf{A}) = \mathbf{Q} \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_d) \end{bmatrix} \mathbf{Q}^* \quad \text{where} \quad \mathbf{A} = \mathbf{Q} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_d \end{bmatrix} \mathbf{Q}^*. \quad (4.1)$$

The definition of $f(\mathbf{A})$ does not depend on which eigenvalue decomposition we choose. Any matrix function that arises in this fashion is called a **standard matrix function**.

Proposition 4.1. *Let $f, g : I \to \mathbb{R}$ be real-valued functions on an interval $I \subseteq \mathbb{R}$, and let $\mathbf{A} \in \mathbb{H}_d$ be a Hermitian matrix whose eigenvalues are contained in I.*

(i) If λ is an eigenvalue of of $\mathbf{A}$, then $f(\lambda)$ is an eigenvalue of $f(\mathbf{A})$.

(ii) $f(a) \le g(a)$ for all $a \in I$ implies $f(\mathbf{A}) \preccurlyeq g(\mathbf{A})$.

4.2 Matrix Khintchin Inequality

Theorem 4.1. [MJC$^+$14, Corollary 7.3] *Suppose that $p = 1$ or $p \ge 3/2$. Consider a finite sequence $(\mathbf{Y}_k)_{k\ge 1}$ of independent, random, Hermitian matrices and a deterministic sequence $(\mathbf{A}_k)_{k\ge 1}$ for which*

$$\mathbf{E}[\mathbf{Y}_k] = 0 \quad and \quad \mathbf{Y}_k^2 \preccurlyeq \mathbf{A}_k^2 \qquad almost\ surely\ for\ all\ k \ge 1. \quad (4.2)$$

Then

$$\mathbf{E}\left[\left\| \sum_{k\ge 1} \mathbf{Y}_k \right\|_{2p}^{2p} \right]^{1/(2p)} \le \sqrt{p - \frac{1}{2}} \left\| \left(\sum_{k\ge 1} (\mathbf{A}_k^2 + \mathbf{E}[\mathbf{Y}_k^2]) \right)^{1/2} \right\|_{2p}. \quad (4.3)$$

In particular, when $(\xi_k)_{k\geq 1}$ is an independent sequence of Rademacher random variables,

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\xi_k\mathbf{A}_k\right\|_{2p}^{2p}\right]^{1/(2p)} \leq \sqrt{2p-1}\left\|\left(\sum_{k\geq 1}\mathbf{A}_k^2\right)^{1/2}\right\|_{2p}. \qquad (4.4)$$

4.3 Matrix Moment Inequality

Theorem 4.2. *Assume $n \geq 3$*

(i) Suppose that $p \geq 1$, and fix $r \geq p \vee 2\log(n)$. Consider a finite sequence $(\mathbf{S}_k)_{k\geq 1}$ of independent, random, positive-semidefinite matrices with dimension $n \times n$. Then

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{S}_k\right\|^p\right]^{1/p} \leq \left[\left\|\sum_{k\geq 1}\mathbf{E}[\mathbf{S}_k]\right\|^{1/2} + 2\sqrt{er}\mathbf{E}[\max_{k\geq 1}\|\mathbf{S}_k\|^p]^{1/(2p)}\right]^2. \qquad (4.5)$$

(ii) Suppose that $p \geq 2$, and fix $r \geq p \vee 2\log(n)$. Consider a finite sequence $(\mathbf{Y}_k)_{k\geq 1}$ of independent, symmetric, random, self-adjoint matrices with dimension $n \times n$. Then

$$\mathbf{E}\left[\left\|\sum_{k\geq 1}\mathbf{Y}_k\right\|^p\right]^{1/p} \leq \sqrt{er}\left\|\left(\sum_{k\geq 1}\mathbf{E}[\mathbf{Y}_k^2]\right)^{1/2}\right\| + 2er\mathbf{E}[\max_{k\geq 1}\|\mathbf{S}_k\|^p]^{1/p}. \qquad (4.6)$$

4.4 Intrinsic Dimension

Definition 4.1. *For a positive-semidefinite matrix $\mathbf{S}$, the **intrinic dimension** is the quantity*

$$\text{intdim}(\mathbf{A}) := \frac{\text{tr}\mathbf{A}}{\|\mathbf{A}\|}.$$

Lemma 4.1. *(Intrinsic dimenision) Let $\varphi : [0,\infty) \to \mathbb{R}$ be a convex function with $\varphi(0) = 0$. For any positive-semidefinite matrix $\mathbf{S}$ it holds that*

$$\text{tr}(\varphi(\mathbf{S})) \leq \text{intdim}(\mathbf{S}) \cdot \varphi(\|\mathbf{S}\|).$$

Proof. *[Tro15, Lemma 7.5.1]* Since φ is convex on any interval $[0,L]$ with $L > 0$ and $\varphi(0) = 0$, it holds

$$\varphi(a) \leq \left(1-\frac{a}{L}\right)\varphi(0) + \frac{a}{L}\varphi(L) = \frac{a}{L}\varphi(L) \qquad \text{for all } a \in [0,L]. \qquad (4.7)$$

Since $\mathbf{S}$ is positive-semidefinite, the eigenvalues of $\mathbf{S}$ fall in the interval $[0,L]$, where $L = \|\mathbf{S}\|$.

$$\text{tr}(\varphi(\mathbf{S})) = \sum_{i=1}^{d}\varphi(\lambda_i) \leq \frac{\sum_{i=1}^{d}\lambda_i}{\|\mathbf{S}\|}\varphi(\|\mathbf{S}\|) = \frac{\text{tr}(\mathbf{S})}{\|\mathbf{S}\|}\varphi(\|\mathbf{S}\|) = \text{intdim}(\mathbf{S}) \cdot \varphi(\|\mathbf{S}\|). \qquad (4.8)$$

$\square$

5 Empirical Processes

Let $(\Omega, \mathcal{A}, \mathbf{P})$ be a probability space and $(\mathcal{X}, \Sigma)$ a measurable space. Let $X_j : (\Omega, \mathcal{A}, \mathbf{P}) \to (\mathcal{X}, \Sigma), j = 1, \ldots, n$ be independent and identically-distributed (i.i.d.) random variables with probability distribution $\mathbf{P}_X$ and $\mathcal{F}$ a family of measurable functions $f : (\mathcal{X}, \Sigma) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Consider the map

$$f \mapsto G_n f := \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbf{P}_X f \right), \tag{5.1}$$

where $\mathbf{P}_X f := \int_{\mathcal{X}} f \mathrm{d} \mathbf{P}_X$. We call $(G_n f)_{f \in \mathcal{F}}$ the empirical process indexed by $\mathcal{F}$. Furthermore

$$\|G_n f\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |G_n f| . \tag{5.2}$$

Lemma 5.1. (Bernstein Inequality for Empirical Processes) *For any bounded, measurable function f it holds for all $t > 0$*

$$\mathbf{P} (|G_n f| > t) \leq 2 \exp \left(-\frac{1}{4} \frac{t^2}{\mathbf{P}_X(f^2) + t \|f\|_{\infty} / \sqrt{n}} \right) \tag{5.3}$$

Proof. By the Markov inequality it holds for all $\lambda > 0$

$$\mathbf{P} (G_n f > t) \leq e^{-\lambda t} \mathbf{E} \exp (\lambda G_n f) \tag{5.4}$$

$\square$

Lemma 5.2. *For any finite class $\mathcal{F}$ of bounded, measurable, square-integrable functions, with $|\mathcal{F}|$ elements, it holds*

$$\mathbf{E} \|G_n f\|_{\mathcal{F}} \lesssim \max_{f \in \mathcal{F}} \frac{\|f\|_{\infty}}{\sqrt{n}} \log (1 + |\mathcal{F}|) + \max_{f \in \mathcal{F}} \|f\|_{\mathbf{P},2} \sqrt{\log (1 + |\mathcal{F}|)}. \tag{5.5}$$

6 Simple yet useful Calculations

Theorem 6.1. (Multivariate Taylor Theorem) *Let $f \in C^2(\mathbb{R}^n, \mathbb{R})$. Then for all $x, \Delta \in \mathbb{R}^n$ there exists $\xi \in [0,1]$ such that it holds*

$$
\begin{aligned}
f(x + \Delta) = f(x) + \sum_{i=1}^{n} \frac{\partial f(x)}{\partial x_i} \Delta_i + \sum_{\substack{i,j=1 \\ i \neq j}} \frac{\partial^2 f(x + \xi\Delta)}{\partial x_i \partial x_j} \Delta_i \Delta_j \\
+ \frac{1}{2} \sum_{i=1}^{n} \frac{\partial^2 f(x + \xi\Delta)}{\partial x_i^2} \Delta_i^2
\end{aligned}
\tag{6.1}
$$

Corollary 6.1.1. *Let $f \in C^2(\mathbb{R})$. Then for all $a, x, \Delta \in \mathbb{R}^n$ there exist $\xi \in [0,1]$ such that it holds*

$$
f(a^T(x + \Delta)) - f(a^T x) = f'(a^T x)\, \Delta^T a + \frac{1}{2} f''(a^T(x + \xi\Delta))\, \Delta^T A\, \Delta,
\tag{6.2}
$$

where $A := aa^T \in \mathbb{R}^{n \times n}$.

Proof. By the chain rule we have for all $a, x, \Delta \in \mathbb{R}^n$ and $\xi \in [0,1]$

$$
\frac{\partial^2 f(a^T(x + \xi\Delta))}{\partial x_i \partial x_j} = f''(a^T(x + \xi\Delta))\, a_i a_j.
\tag{6.3}
$$

Since $A := aa^T$ is symmetric we have

$$
\Delta^T A\, \Delta = 2 \sum_{\substack{i,j=1 \\ i \neq j}}^{n} a_i a_j \Delta_i \Delta_j + \sum_{i=1}^{n} a_i^2 \Delta_i^2.
\tag{6.4}
$$

Plugging (6.3) and (6.4) into (6.1) yields (6.2). $\qquad\square$

Proposition 6.1. *For all $x, y \in \mathbb{R}$ it holds*

$$
|x + y| - |x| \geq -|y|
\tag{6.5}
$$

Proof. Checking all 6 combinations of $x + y, x, y$ being nonnegative or negative yields the result. $\qquad\square$

Notation Index

$\#A$ cardinality of the set A

$\mathbf{E}[X|Y]$ conditional expectation of the random variable X with respect to $\sigma(Y)$

$\mathbf{E}[X]$ expectation of the random variable X

$\mathbf{Var}[X]$ variance of the random variable X

$\overline{\overline{\mathbb{R}}} = \mathbb{R} \cup \{+\infty\}$ extension of the real numbers

$\xrightarrow{\mathcal{D}}$ convergence of distributions

$\mathbf{P}$ generic probability measure

$\mathbf{P}_X = \mathbf{P} \circ X^{-1}$ distribution of the random variable X

$\mathbb{R}$ set of real numbers

$x \vee y, x \wedge y, x^+, x^-$ maximum, minimum, positive part, negative part of real numbers

$X \sim \mu$ the random variable has distribution μ

Bibliography

[BT03]   Dimitri P. Bertsekas and John N. Tsitsiklis. Parallel and Distributed Computa-
            tion:Numerical Methods. November 2003.

[MJC+14] Lester Mackey, Michael I. Jordan, Richard Y. Chen, Brendan Farrell, and Joel A.
            Tropp. Matrix concentration inequalities via the method of exchangeable pairs. *The
            Annals of Probability*, 42(3), May 2014.

[Tro15]   Joel A. Tropp. An Introduction to Matrix Concentration Inequalities, January 2015.