Matrix Constraints and Multi-Task Learning for Covariate Balance

Eli Ben-Michael and Avi Feller (U.C. Berkeley)

Berkeley UNIVERSITY OF CALIFORNIA

Summary

- Many estimators use multiple sets of weights
- Can jointly find all weights in a single optimization problem
- Dual view as multi-task learning and hierarchical modeling
- Can combine design-based weights with outcome modeling (AIPW)

Motivation: Estimating Multiple Means

Typical to use more than one set of weights in an estimator

- Estimating the ATE (vs. ATT)
- Heterogeneous treatment effects and subgroup effects
- Obs. study with missing outcomes (IPTW + IPMW; SUR)
- Multilevel observational study

Instead of fitting weights separately, we can exploit natural structure

Vector Constraints for Covariate Balance

Choose weights that balance covariates and are close to uniform

$$\min_{\gamma \in \mathbb{R}^{n_0}} \sum_{T_i=0} f(\gamma_i) + h(\bar{X}_1 - X_0'\gamma)$$

$X_1 \in \mathbb{R}^{n_1 \times d}$, $X_0 \in \mathbb{R}^{n_0 \times d}$: treated/control; $\bar{X}_1 \in \mathbb{R}^d$: treated means
General formulation encompasses several estimators, [1, 2, 3, 4, 5]

- **Dispersion Function**: f penalizes large weights
- **Balance Criterion**: h measures covariate balance

Dual representation as *regularized M-estimator* of p-score

$$\min_{\theta \in \mathbb{R}^d} \sum_{T_i=0}^{n} f^*(X_i'\theta) - \bar{X}_1'\theta + h^*(\theta)$$

Where g^* is the *convex conjugate* of g.

- **Odds Function**: $f^{*\prime}(X'\theta) = \frac{\pi(X)}{1-\pi(X)}$
- **Regularization**: h^* penalizes complex models

Example 1: Entropy Balancing [1] with L^∞ constraint [5]

$$\min_{\gamma \in \mathbb{R}^{n_0}} \sum_{T_i=0} \gamma_i \log \gamma_i$$
$$\text{s.t. } \|\bar{X}_1 - X_0'\gamma\|_\infty \le \delta$$

Dual fits logit p-score with LASSO penalty

$$\min_{\theta \in \mathbb{R}^d} \sum_{T_i=0} \exp(X_i'\theta - 1) - \bar{X}_1'\theta + \delta\|\theta\|_1$$

Matrix Constraints: Special Cases

Example 2: Estimating subgroup ATT

$$\tau_k = \mathbb{E}[Y(1) - Y(0) \mid T = 1, G = k]$$

Primal: Weighted Frobenius *soft* constraint

$$\min_{\Gamma \in \mathbb{R}^{n_0 \times m}} \sum_{k=1}^{m} \sum_{G_i=k, T_i=0} \Gamma_{ik} \log \Gamma_{ik} + \frac{\lambda}{2} \text{tr}((\bar{X}_1 - X_0'\Gamma)\Omega(\bar{X}_1 - X_0'\Gamma))$$

$\bar{X}_1 \in \mathbb{R}^{d \times m}$ subgroup treated means

Dual: Fully-interacted logit p-score with **Hierarchical Prior**

$$\min_{\Theta \in \mathbb{R}^{d \times m}} \sum_{k=1}^{m} \sum_{G_i=k, T_i=0} \exp(X_i'\Theta_k - 1) - \text{tr}(\bar{X}_1\Theta') + \frac{\lambda}{2} \text{tr}(\Theta\Omega^{-1}\Theta')$$

$\Omega \in \mathbb{R}^{m \times m}$ corresponds to *prior covariance*

$$\Theta_{j1}, \ldots, \Theta_{jm} \sim \text{Normal}(0, \Omega)$$

Example 3: Estimating individual-level CATE

$$\tau_i = \mathbb{E}[Y(1) - Y(0) \mid X = x_i], \quad T_i = 1$$

Primal: Spectral norm *hard* constraint

$$\min_{\Gamma \in \mathbb{R}^{n_0 \times n_1}} \sum_{T_k=1} \sum_{T_i=0} \Gamma_{ik} \log \Gamma_{ik}$$
$$\text{s.t. } \sup_{\|u\|_2=1} \|(X_1 - X_0'\Gamma)u\|_2 \le \delta$$

Dual: Individual logit p-score models have **Low Rank**

$$\min_{\Theta \in \mathbb{R}^{d \times n_1}} \sum_{T_k=1} \sum_{T_i=0} [\exp(X_i'\Theta_k - 1) - X_k'\Theta_k] + \delta\|\Theta\|_*$$

Corresponds to assuming that

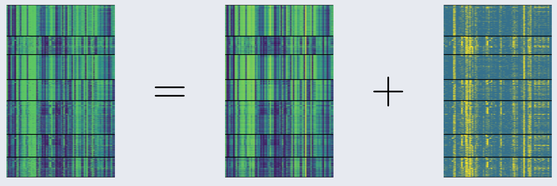
- There are $p \ll n_1$ archetypal models $U \in \mathbb{R}^{d \times p}$
- Each realization: weighted average $\Theta = UV'$, $V \in \mathbb{R}^{n_1 \times p}$

We can also **combine vector and matrix constraints**

Primal: Spectral norm and L^∞ *hard* constraint

$$\min_{\Gamma \in \mathbb{R}^{n_0 \times n_1}} \sum_{T_k=1} \sum_{T_i=0} \Gamma_{ik} \log \Gamma_{ik}$$
$$\text{s.t. } \sup_{\|u\|_2=1} \|(X_1 - X_0'\Gamma)u\|_2 \le \delta_1$$
$$\|X_1 - X_0'\Gamma\|_\infty \le \delta_2$$

Dual: Robust PCA decomposition, used in image processing [6]



Full = Low Rank + Sparse

Matrix Constraints and Multi-Task Learning

Primal: Measure dispersion separately, **balance** jointly

$$\min_{\Gamma} \sum_{k=1}^{m} \sum_{i} (1 - T_i) f(\Gamma_{ik}) + h\left(\bar{X}_1 - X_0'\Gamma\right)$$

Dual: Fit p-score models separately, **regularize** jointly

$$\min_{\Theta} \sum_{k=1}^{m} \sum_{i} \left[(1 - T_i) f^*(\Theta_k' X_i) - \frac{1}{n_{1k}} T_i X_i' \Theta_k \right] + h^*(\Theta)$$

Sim Study: Matrix vs. Vector Constraints

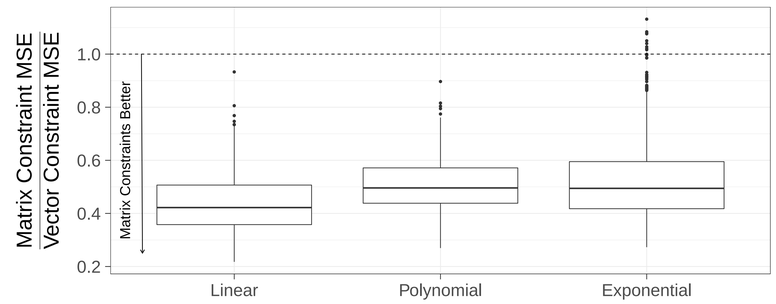


Fig: MSE ratio of Nuclear (Matrix) vs L1 (Vector) penalties, following [7]

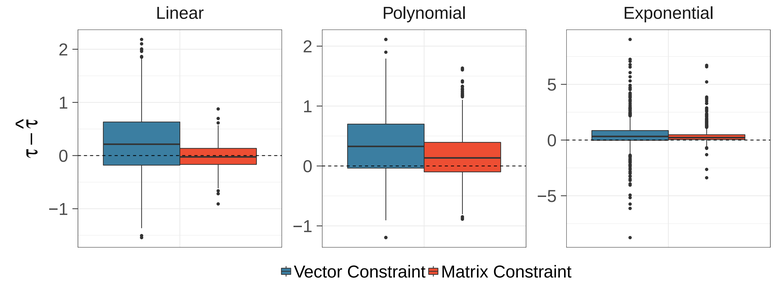


Fig: ATT estimates with Nuclear (Matrix) vs L1 (Vector) penalties, following [7]

References

[1] J. Hainmueller, "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies," *Political Analysis*, vol. 20, pp. 25–46, 2011.

[2] J. R. Zubizarreta, "Stable Weights that Balance Covariates for Estimation With Incomplete Outcome Data," *Journal of the American Statistical Association*, vol. 110, no. 511, pp. 910–922, 2015.

[3] Q. Zhao and D. Percival, "Entropy Balancing is Doubly Robust," *Journal of Causal Inference*, 2016.

[4] Q. Zhao, "Covariate Balancing Propensity Score by Tailored Loss Functions," 2017.

[5] Y. Wang and J. R. Zubizarreta, "Minimal Approximately Balancing Weights: Asymptotic Properties and Practical Considerations." 2018.

[6] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis," *Journal of the ACM*, vol. 58, no. 3, pp. 11:1—11:37, 2011.

[7] J. L. Hill, "Bayesian nonparametric modeling for causal inference," *Journal of Computational and Graphical Statistics*, vol. 20, no. 1, pp. 217–240, 2011.