Coarse Probabilistic Matching for Causal Inference through Bayesian Fellegi Sunter

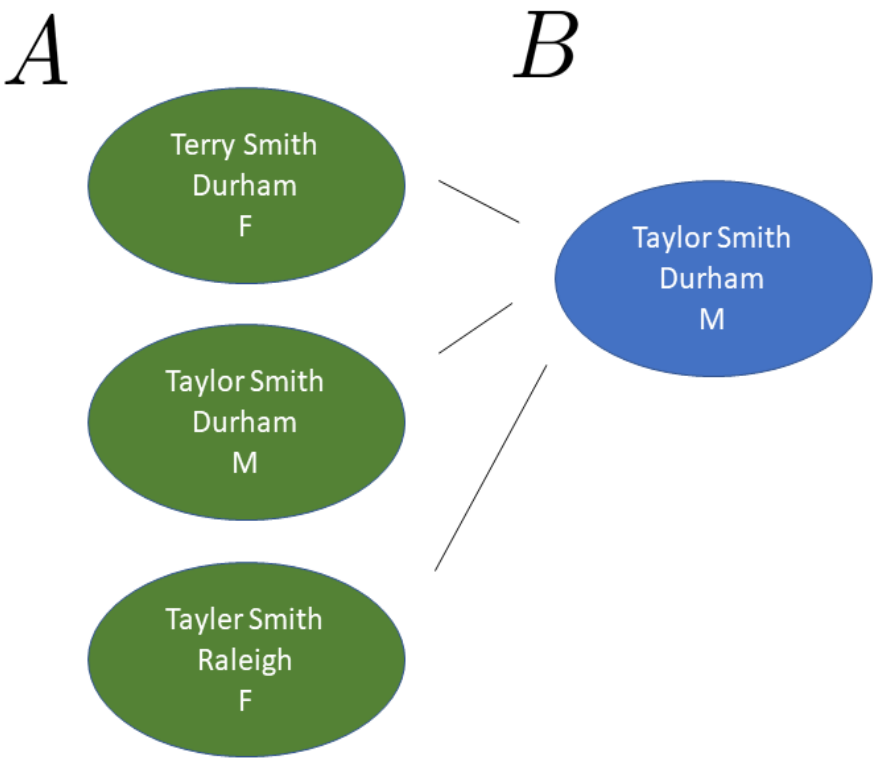
Brian Kundinger

Duke University

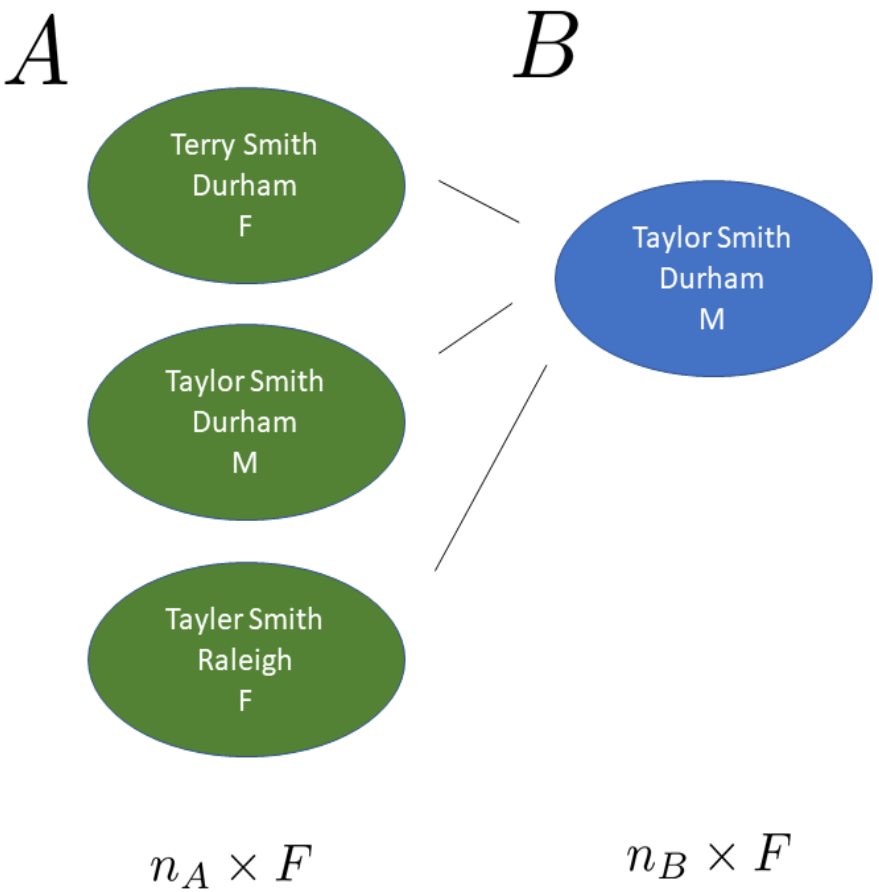
July 10, 2023

Table of Contents

Linkage through Comparison Vectors

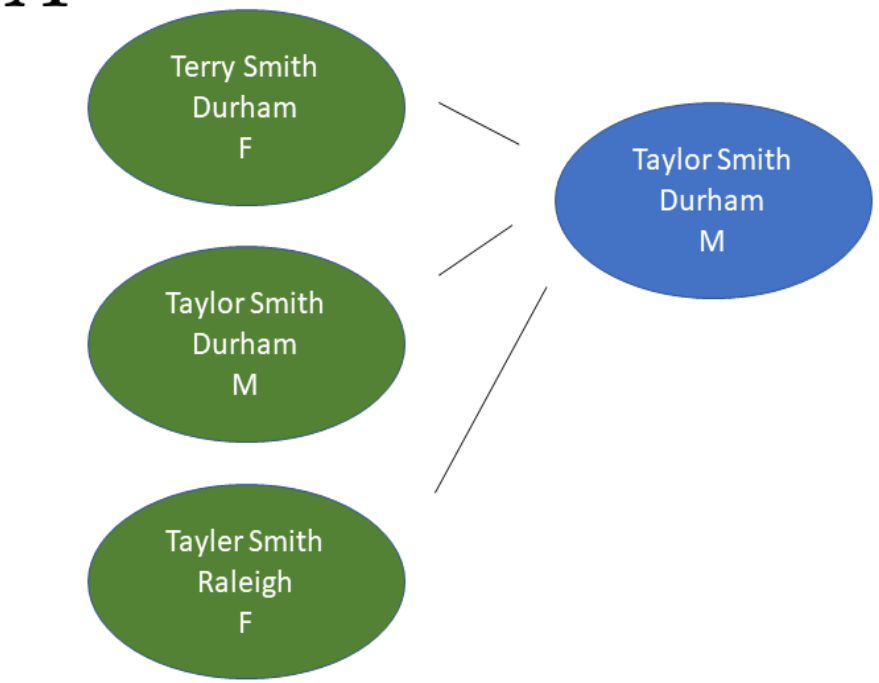


Linkage through Comparison Vectors



- n_A, n_B records in A, B
- $F = 4$ features for comparison
  - First name
  - Last name
  - City
  - Gender
- $L = \{3, 3, 2, 2\}$ levels of comparison

Linkage through Comparison Vectors



$n_A \times F$ $\quad$ $n_B \times F$

Γ

$$\gamma_{11} = [3, 1, 1, 2]$$

$$\gamma_{21} = [1, 1, 1, 1]$$

$$\gamma_{31} = [2, 1, 2, 2]$$

$n_A n_B \times F$

Review of Fellegi Sunter

Let $\Delta_{ij} \sim$ Bernoulli(λ) be the indicator that (A_i, B_j) is a match.
Let $m_{fl} = p(\gamma_{ij}^f = l | \Delta_{ij} = 1)$, and $u_{fl} = p(\gamma_{ij}^f = l | \Delta_{ij} = 0)$.

The Fellegi Sunter Model is:

$$p(\Gamma | m, u, \Delta) = \prod_i^{n_A} \prod_j^{n_B} \lambda \left[\prod_{f=1}^{F} \prod_{l=1}^{L_f} m_{fl}^{\gamma_{ij}^f = l} \right] + (1 - \lambda) \left[\prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{\gamma_{ij}^f = l} \right]$$

Bayesian Fellegi Sunter

- Different Bayesian priors on Δ can enforce many-to-many, many-to-one, or one-to-one matchings.
- Bayesian methods average results over sets of plausible matchings, and provide uncertainty quantification.
- Extensions for enhanced speed and scalability

Table of Contents

Matching in Causal Inference

- When estimating causal effects using observational data, it is desirable to replicate a randomized experiment as closely as possible by obtaining treated and control groups with similar covariate distributions.
- This goal can often be achieved by choosing well-matched samples of the original treated and control groups, thereby reducing bias due to the covariates ¡2-¿
- If Fellegi Sunter works for record linkage, why not for causal inference?

Mahalonobis Distance and Propensity Matching

- One approach to matching is to use the Mahalonobis distance, which measures distance between units purely in the covariate space, without regard to which features are related to the outcome of interest. It is defined by

$$d(X_1, X_2) = \sqrt{|X_1 - X_2| S^{-1} |X_1 - X_2|}$$

- Another method is to estimate propensity scores $p(Z_n = 1|X_n)$ for each unit in the data, and match units based on this score.
- Both methods are Equal Percent Bias Reducing (EPBR), meaning that

$$E(\bar{X}_{m_T} - \bar{X}_{m_C}) \gamma E(\bar{X}_{n_T} - \bar{X}_{n_C})$$

for $0 \leq \gamma < 1$, where m_T, m_C are the matched treated and control data, and n_T, n_C are the full data

Critique of Propensity Score Matching

- David King published a strong critique of EPBR methods in general, and PS matching in particular
- The theoretical garauntees for EPBR occur in expectation across many experiments (expectation on the RHS)
- Only one parameter γ for all fields, no way to match more precisely for specific fields
- PS matching reduces all of the information in X to single score, often matches units that do not actually appear similar
- PS attempts to recreate a *completely* randomized experiment, but it is better to recreate a *blocked* randomized experiment

King's Coarse Exact Matching (CEM)

- Coarsen each variable to integer values based on reasonable knowledge (or using automated recommendations)
- Each data unit has effectively been transformed to a set of integer values.
- Call treated and control units a match if they match at the coarsened level for *all* variables
- This coarsening is controversial, but it allows CEM to be Monotonic Imbalance Bounding (MIB). Shown to have better covariate balance, at expense of sample size.

King's Coarse Exact Matching (CEM)

- Coarsen each variable to integer values based on reasonable knowledge (or using automated recommendations)
- Each data unit has effectively been transformed to a set of integer values.
- Call treated and control units a match if they match at the coarsened level for *all* variables
- This coarsening is controversial, but it allows CEM to be Monotonic Imbalance Bounding (MIB). Shown to have better covariate balance, at expense of sample size.
- THOUGHT: This coarsening sounds a lot like Fellegi-Sunter...

Table of Contents

Monotonic Imbalance Bounding

As defined by Iacus, King, Porro (2012), a matching method is *Monotonic Imbalance Bounding* on a function f with respect to a distance $D(\cdot, \cdot)$, if for a monotonically increasing function $\gamma_{f,D}(\cdot)$ and any $\pi \in \mathbb{R}_+^K$, we have

$$D(f(\mathcal{X}_{m_T(\pi)}), f(\mathcal{X}_{m_C(\pi)})) \leq \gamma_{f,D}(\pi)$$

Here, π is a vector of scalar parameters relating the coarsening of each feature. Also $m_T(\pi)$ and $m_C(\pi)$ are the sizes of the matched sets, importantly determined by π, and not determined by the modeller. Lastly, $\mathcal{X}_{m_T}$ and $\mathcal{X}_{m_C}$ are the matched subsets of T and C.

Expected Imbalance Bounding

King's CEM is deterministic, and therefore is able to satisfy stronger criteria. No probabilistic method can fulfill this. However, I propose we think about a class of Expected Imbalance Bounding (EIB) matching methods, as somewhat of a middle ground between MIB and EPBR. Such methods would satisfy:

$$E\left[D(f(\mathcal{X}_{m_T(\pi)}), f(\mathcal{X}_{m_C(\pi)}))\right] \leq \gamma_{f,D}(\pi)$$

Review of Fellegi Sunter

Let $\Delta_{ij} \sim$ Bernoulli(λ) be the indicator that (A_i, B_j) is a match.
Let $m_{fl} = p(\gamma_{ij}^f = l | \Delta_{ij} = 1)$, and $u_{fl} = p(\gamma_{ij}^f = l | \Delta_{ij} = 0)$.

The Fellegi Sunter Model is:

$$p(\Gamma | m, u, \Delta) = \prod_{i=1}^{n_A} \prod_{j=1}^{n_B} \lambda \left[\prod_{f=1}^{F} \prod_{l=1}^{L_f} m_{fl}^{\gamma_{ij}^f = l} \right] + (1 - \lambda) \left[\prod_{f=1}^{F} \prod_{l=1}^{L_f} u_{fl}^{\gamma_{ij}^f = l} \right]$$

MIB for Binary Covariates

Let X_{m_T} and X_{m_C} be vectors of binary covariates for units matched through Fellegi-Sunter (and with matchings encoded in Δ)

$$\begin{aligned} E[|\bar{X}_{m_T} - \bar{X}_{m_C}|] &= \frac{1}{n} E\left[|\sum X_{m_T,i} - \sum X_{m_C,j}||\Delta\right] \\ &\leq \frac{1}{n} E\left[\sum |X_i - X_j||\Delta_{ij} = 1\right] \\ &= 0 \cdot m_1 + 1 \cdot m_2 \\ &= m_2 \end{aligned}$$

MIB for Continuous Covariates

I can prove this for continuous covariates as well. Suppose X has range $R = X_{\max} - X_{\min}$, and bin the data according to $\epsilon = \frac{R}{K}$, where $\pi_j = K$ is our coarsening parameter. Then the upper bound becomes $\epsilon \sum_{k=1}^{K} k m_k$. Note that this is increasing in K.

Note that we can directly compute this bound, and check that it holds on actual matched samples. Therefore, we can use desired bounds on covariate imbalance as a way to set priors for the m parameters!

Table of Contents

Course Probabilistic Matching

In this context, I'd like to propose Course Probabilistic Matching (CPM). As in Fellegi-Sunter, we create rules to characterize the level of agreement between different values. For example, we could have $\gamma_{ij}^f = |C(X_{i,f}) - C(X_{j,f})|$, where $C(\cdot)$ is a coarsening function mapping bins in the covariate space to integer values.

We calculate probabilities exactly as we would under FS:

$$P(\gamma_{ij}|\Delta_{ij} = 1) = \prod_{f=1}^{F} p(\gamma_{ij}^f|\Delta_{ij} = 1)$$

$$P(\gamma_{ij}|\Delta_{ij} = 0) = \prod_{f=1}^{F} p(\gamma_{ij}^f|\Delta_{ij} = 0)$$

These are just products of relevant m and u probabilities.

Course Exact Matching

Then, CEM can be viewed as a special case of CPM, where comparison vectors are constructed through $\gamma_{ij}^{f} = I(C(X_{i,f}) = C(X_{j,f}))$, and probabilities are assigned deterministically through

$$P(\Delta_{ij} = 1 | C(X_i) = C(X_j)) = P(\Delta_{ij} = 1 | \gamma_{ij}^{f} = 1, \forall f) = 1$$
$$P(\Delta_{ij} = 1 | C(X_i) \neq C(X_j)) = 0$$

Additionally, if the m parameter gives all its probability mass to exact agreement, then the MIB bound for FS, $\epsilon \sum_{k=1}^{K} k m_k$, reduces down to just ϵ, as provided in the paper.

One Difficulty

Ordinarily, we use $m^f \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_{L_f})$ as a prior for each field. However, I sense that this will be too unstructured, and the sampler would be too flexible in choosing matches.

However, if we decide on acceptable bounds for covariate imbalance, we can use these bounds to construct these priors.

Stacked Truncated Beta Priors

Sadinle 2014 used a series of beta distributions as follows. Define $m^*_{f0} = P(\Gamma^f_{ij} = 0)$ and $m^*_{fl} = P(\Gamma^f_{ij} = l | \Gamma^f_{ij} > l - 1)$.

These are related to the standard Fellegi Sunter parameters through:

$$m_{fl} = P(\Gamma^f_{ij} = l | \Delta_{ij} = 1) = \prod_{l=1}^{L_f - 1} m^{*I(\gamma_{ij}=l)}_{fl} (1 - m_{fl})^{*I(\gamma_{ij}>l)} \quad (1)$$

Prior Specification

This parameterization is more useful because we can more directly control the kinds of matches we want to make through setting the support of the prior. We would use $m^*_{f0} \sim \text{TBeta}(\alpha_{fl}, \beta_{fl}, \lambda_{fl}, 1)$ prior distribution, where λ_{fl} is a lower bound for the support.

This may seem overly informative, but these λ effectively set bounds for the allowable covariate imbalance, and it is common for this to be chosen by the matcher.

Table of Contents

Concluding Thoughts

- I suspect the Coarse Probabilistic Matching would outperform Course Exact Matching pretty easily. It would find all of the coarsened "exact" matches, plus additional near matches, resulting in mild increases in covariate imbalance with substantial increases in sample size after matching
- Since this matching method is fully Bayesian and probabilistic, we can create joint models for the matching and treatment effect estimation, propagating uncertainty between the steps, and averaging over many plausible matchings.

Tons of Ways Forward

- Establish theory for CPM in the context of other matching methods, compare out of box performance. Particularly interested in comparisons with King's Matching Frontier and Morucci's Matching Bounds.
- Incorporate propensity scores in the CPM method (construct a comparison feature based on that score)
- Variable selection prior for propensity score, feeding into a record linkage model on the selected covariates
- Many other ideas!