Coarse Probabilistic Matching for Causal Inference through Bayesian Fellegi Sunter

Brian Kundinger

Duke University

May 15, 2023

Table of Contents

1 Stuff I've Been Thinking About

2 Fellegi Sunter as Monotone Imbalance Bounding

3 Fellegi Sunter as Course Probabilistic Matching

4 Conclusion

Fellegi Sunter

- Let datasets A and B have records indexed $i \in \{1, \ldots, n_A\}$ and $j \in \{1, \ldots, n_B\}$
- For each record pair, compute comparison vector γ_{ij}, where each component γ_{ij}^f takes on value $\ell \in \{1, \ldots, L_f\}$ according to predefined distance function
    - Categorical: $\gamma_{ij}^f = 1$ if (i, j) agree on f, $\gamma_{ij}^f = 2$ otherwise
    - Names: 1 for exact agreement, 2 for close agreement, 3 for no agreement

Fellegi Sunter

- Every time I've presented this model, I'm embarassed by this coarsening of data.
- Still, I think it might be interesting to see if it could be useful for matching in 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.
- Most matching methods in causal inference then use deterministic algorithms to match treated units to their nearest control unit based on the chosen metric
- David King publishes a strong critique of PS matching In any given matching, no guarantee of reducing covariate imbalance

Coarse Exact Matching

- 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 allows CEM to be Monotonic Imbalance Bounding (MIB). Shown to have have better covariate balance, at expense of sample size ¡2¿
- THOUGHT: If coarsening was good in CEM, perhaps it can be good in Fellegi Sunter

Review of Fellegi Sunter

Let $\Delta_{ij} \sim \text{Bernoulli}(\lambda)$ 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:

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

Table of Contents

1 Stuff I've Been Thinking About

2 Fellegi Sunter as Monotone Imbalance Bounding

3 Fellegi Sunter as Course Probabilistic Matching

4 Conclusion

MIB for Binary Covariates

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.

MIB for Binary Covariates

I've been struggling with the exact best way to interpret this definition, so I've editted my proof a couple times, and I don't have a finished version. But for now, I can show you that they have this flavor:

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)

$$\begin{aligned}
|\bar{X}_{m_T} - \bar{X}_{m_C}| &= E[|\bar{X}_{m_T} - \bar{X}_{m_C}||\Delta] \\
&\leq E[|X_i - X_j||\Delta_{ij} = 1] \\
&= 0\dot{m}_1 + 1\dot{m}_2 \\
&= m_2
\end{aligned}$$

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 $\epsilon \sum_{k=1}^{K} km_k$. Note that this is increasing in K.

Note that we can directly compute this bound, and check that it holds on actual matched samples.

Table of Contents

1 Stuff I've Been Thinking About

2 Fellegi Sunter as Monotone Imbalance Bounding

3 Fellegi Sunter as Course Probabilistic Matching

4 Conclusion

Course Probabilistic Matching

In this context, I'd like to propose Course Probabilistic Matching (CPM). Instead coarsening the*distance* between covariates, we will calculate the distance between coarsened covariates. For example, we would 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:

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

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

These are just products of relevant *m* and *u* probabilities.

CEM

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.

Table of Contents

1 Stuff I've Been Thinking About

2 Fellegi Sunter as Monotone Imbalance Bounding

3 Fellegi Sunter as Course Probabilistic Matching

4 Conclusion

Concluding Thoughts

- Last week, I thought that FS would perform about the same as Mahalanobis distance, and that we would need to modify it a bunch to make it useful for causal inference
- Now I suspect the Coarse Probabilistic Matching would outperform Course Exact Matching, right out of the box. 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
- We can measure match quality through posterior probabilities, and discard matches we think are too weak
- 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
- Incorporate propensity scores in the CPM method (estimate PS through draws of β, construct a comparison feature based on that score)
  - More sophisticated, use scores in a dual latent class model.
- Variable selection prior for propensity score, feeding into a record linkage model on the selected covariates
- Empirical Process for deriving bounds on covariate imbalance