Unpaired measurement integration using Markov Link Intervals

Jackson Loper[a,b*], Trygve Bakken[c], Uygar Sumbul[c],
Gabe Murphy[c], Hongkui Zeng[c], David Blei[b,d], Liam Paninski[a,b]

[a]Grossman Center for the Statistics of Mind, Zuckerman Mind-Brain Behavior Institute,
Department of Neuroscience, and Center for Theoretical Neuroscience, Columbia University, NY NY

[b]Department of Statistics, Columbia University, NY NY

[c]Allen Institute for Brain Science, Seattle WA

[d]Department of Computer Science, Columbia University, NY NY

April 17, 2019

Abstract

We consider two datasets, each measuring a different of observations on a common population. The first dataset is constructed by drawing individuals randomly from this population and measuring a quantity about each individual. The second dataset is constructed similarly, but a different quantity is measured. "Unpaired measurement integration" attempts to use this data to infer the joint distribution of the two different quantities. We overcome the unidentifiability problems in this task by measuring an additional quantity, similar to an instrumental variable. This quantity is observed in both datasets and satisfies a conditional independence assumption. We characterize identifiability in the presence of these variables and develop a consistent estimator. The performance of the estimator is tested in simulations. The method is applied to transcriptomic measurements to uncover the relationship between two different technologies for measuring gene expression in mouse neurons.

1 Introduction

Unpaired measurement integration seeks to infer the joint distribution on two variables (X, Y) using knowledge of the marginal distributions on X and Y. This problem is ill-posed, but practical algorithms have nonetheless been developed. An early approach by Andrey Kolmogorov noted that the expectation for any linear function $f(X, Y) = \alpha X + \beta Y$ could be effectively bounded using only the marginal distributions [1]. Bounds on arbitrary univariate functions $f(X, Y)$ have also been developed in the context of causal inference [2, 3]. More recently, the CycleGan algorithm has gained prominence in the computer graphics community for succesfully estimating a joint distribution between photographs (X) and watercolor paintings (Y) of the same subject; CycleGan accomplishes this unpaired measurement integration using two completely separate datasets, one of photographs and one of watercolor paintings [4].

To resolve the fundamental ill-posedness of unpaired dataset integration, some form of extra information is often employed. We assume this side-information variable can be observed for each sample in both datasets. The work of Fan *et al* (2017) gives a detailed overview for this case [3]. Fan characterizes the asymptotics of estimators for $\mathbb{E}[f(X, Y)]$, where f can takes a wide variety of forms. Estimator asymptotics are established under many different assumptions on the relationship between the measurements and the side-information

Inspired by problems in the study of cell populations, this paper considers a new kind of assumption about the side-information. Let L denote the side-information. Suppose that L and Y are conditionally independent given X. This assumption holds naturally whenever both L and Y can both be understood as

*To whom correspondence should be addressed: email jl5116@columbia.edu.

independent noisy measurements of X. We give a complete real-world example in Section 5. The variable L bears a formal similarity to an instrumental variable, but its function is different: instrumental variables resolve causal effects in the presence of unobserved confounders, whereas these variables permit joint inference when only marginal observations are available.

Inference under the conditional independence assumption must overcome two obstacles: the datasets only allow approximate estimation for the marginal distributions, and the marginal distributions only give a partial picture of the joint. These problems cannot be considered separately. If we treat statistical estimates as exact truth and proceed with a purely algebraic analysis of the model, inference can go badly wrong. It turns out that even infinitesimal estimation errors can result in substantially incorrect conclusions, as we show in Lemma 2. This marks a stark contrast from the conditions considered in Fan *et al* (2017), where additional statistical noise results in a graceful degradation of inference accuracy.

In this paper, we overcome both of these challanges for the discrete setting. When L, X, Y are categorical distributions it is possible to make a detailed study of what are known as the observational equivalence classes [5, 6, 7]. Each class corresponds to a set of parameters that give rise to the same distribution on the data we observe. We can achieve consistent inference by taking a carefully chosen dilation of these sets. We design a simple optimization algorithm to find extremal points in these dilated sets. These extremal points yield bounds for the parameters which we dub "Markov Link Intervals," because they use the Markov conditional independence assumption to bound the parameter that links X to Y. Performance is investigated through a variety of simulation studies. Finally, the estimator is applied to uncover the relationship between two different technologies for measuring gene expression in mouse neurons.

2 Formal problem statement and notation

Consider a triple of discrete random variables variables, $(L, X, Y) \sim p$. We assume p has support on $\{1 \cdots N_L\} \times \{1 \cdots N_X\} \times \{1 \cdots N_Y\}$ and L, Y are conditionally independent given X. We parameterize p as

$$p(\ell, x, y; \pi, \alpha, \beta) = \pi_\ell \alpha_{\ell,x} \beta_{x,y}$$

where π lies strictly in the interior of the N_L-simplex, α is an $N_L \times N_X$ right-stochastic matrix, and β is an $N_X \times N_Y$ right-stochastic matrix. We will also find it useful to let $\gamma_{\ell y}$ denote $p(y|\ell) = \sum_x \alpha_{\ell,x}\beta_{x,y}$.

Suppose we are given independent samples from p, but some data is missing. Each sample includes L, X or L, Y but no sample ever includes both X and Y. Can we still learn p? Let $n, m > 0$ and $(L_1, X_1, Y_1), (L_2, X_2, Y_2) \cdots (L_{n+m}, X_{n+m}, Y_{n+m})$ denote samples of independent copies of (L, X, Y). Let

$$\mathcal{D}^X = (L_1, X_1), \cdots (L_n, X_n)$$
$$\mathcal{D}^Y = (L_{n+1}, Y_{n+1}), \cdots (L_{n+m}, Y_{n+m})$$

Our task is to use the data $(\mathcal{D}^X, \mathcal{D}^Y)$ to estimate the parameters of p, namely (π, α, β). Estimating π, α from this data is fairly straightforward. However, existing estimation theory is less helpful for the parameter β. In this paper we develop theory and practice for estimating this parameter.

3 Markov Link Intervals (MLI) for estimating β

Correct inferences about β require careful treatment of its identifiability. Our first result is a negative one:

Lemma 1 (Unidentifiability)**.** *When $N_L < N_X$ there is no generally consistent estimator to estimate β from the data $(\mathcal{D}^X, \mathcal{D}^Y)$.*

Unfortunately, the $N_L < N_X$ regime is precisely the one which motivated our present work, inspired by investigation of cell-types in neuroscience. This result shows that perfect estimation of β may not be possible in our case.

When a parameter cannot be consistently estimated, one common approach is to bound the unidentifiable parameter using "observational equivalence classes" [5, 6, 7]. In our case, this equivalence class is defined as the set of all the different parameters (π', α', β') which would give rise to the true distribution on the observable variables $(\mathcal{D}^X, \mathcal{D}^Y)$. Even though β is unidentifiable, we might hope to identify this set of possible parameters. Our second result shows this approach is unlikely to succeed:

Lemma 2 (Discontinuity of observational equivalence classes)**.** *Endow the space of observable distributions with the total variation metric. Endow the space of sets of parameters π, α, β with the Pompeiu-Hausdorff metric. The map from possible distributions on $\mathcal{D}^X, \mathcal{D}^Y$ to observational equivalence classes of parameters is not continuous. This mapping is sometimes discontinuous even around points where the rows of α are linearly independent and all parameters lie in the interior of the natural parameter space.*

Even infinitesimal error in the estimation of the observable distribution may thus yield large errors in our estimates of the observational equivalence class.

A partial resolution to this difficulty can be obtained if it is known that β is strictly positive and the rows of α are linearly independent. Under these conditions we can first estimate α and then estimate the observational equivalence class for β for the observations $\mathcal{D}^Y$ under a fixed value for α. It turns out that the distance between this class and β does converge to zero:

Theorem 3 (Consistency for positive β)**.** *Let the rows of α be linearly independent and let $\beta_{x,y} \geq c > 0$. Define γ as the $N_L \times N_Y$ right-stochastic matrix*

$$\gamma_{\ell y} \triangleq \sum_x \alpha_{\ell x} \beta_{x,y}$$

Let $\hat{\pi}, \hat{\alpha}, \hat{\gamma}$ denote any consistent estimators for π, α, γ. Consider $\{\beta' : \hat{\alpha}\beta' = \hat{\gamma}\}$, the equivalence class for the observations $\mathcal{D}^Y$ after fixing the estimate $\hat{\alpha}$. The distance between the point β and this set converges to zero in probability.

A more general solution can be obtained by seeking a small dilation of the observational equivalence classes. Let $\mathcal{L}(\mathcal{D}^X, \mathcal{D}^Y; \hat{\pi}, \hat{\alpha}, \hat{\beta})$ denote the likelihood of the observed data under the model under the parameters $\hat{\pi}, \hat{\alpha}, \hat{\beta}$. Let $\mathcal{L}^*$ denote the maximum value of $\mathcal{L}$ over all possible parameters. Define

$$\hat{\Theta} = \left\{ \hat{\pi}, \hat{\alpha}, \hat{\beta} : \ \mathcal{L}(\mathcal{D}^X, \mathcal{D}^Y; \hat{\pi}, \hat{\alpha}, \hat{\beta}) > \mathcal{L}^* - \log(n+m) \right\}$$

This log-scaled likelihood envelope has two desirable properties: it both concentrates around the true distribution on the observables and is asymptotically guaranteed to contain the true model parameters.

Theorem 4 (Consistency of $\hat{\Theta}$)**.** *Let $n, m \to \infty$ such that $|\log(n/m)| < M$. Then:*

1. *$\hat{\Theta}$ is asymptotically guaranteed to contain the true parameter values.*
2. *$\hat{\Theta}$ concentrates for $\pi, \alpha, \alpha\beta$. Let $\xi(\pi', \alpha', \beta')$ denote the maximum of the Euclidean differences $|\pi - \pi'|$, $|\alpha - \alpha'|$, $|\alpha\beta - \alpha'\beta'|$. Then $\max_{\theta \in \hat{\Theta}} \xi(\theta)$ converges to zero in probability.*

This theorem indicates a practical solution for estimating β. For each x, y, we define the *Markov Link Interval* as

$$[\hat{\beta}^-_{x,y}, \hat{\beta}^+_{x,y}] \triangleq \left[\min_{\pi', \alpha', \beta' \in \hat{\Theta}} \beta'_{x,y}, \max_{\pi', \alpha', \beta' \in \hat{\Theta}} \beta'_{x,y} \right]$$

Theorem 4 shows that $\beta_{x,y}$ is asymptotically included in the interval $[\hat{\beta}^-_{x,y}, \hat{\beta}^+_{x,y}]$.

To calculate the Markov Link Interval we must solve the associated optimization problems. Due to the high dimensionality of the problem and the ill-conditioning of the matrices involved, we were unable to tune off-the-shelf optimization methods to reliably solve these problems. We developed an alternative optimization procedure which relies on a minorization-maximization approach and requires no tuning parameters. Appendix A provides a detailed outline of the algorithm.[1]

[1]Code is published at https://github.com/jacksonloper/markov-link-method, including a tutorial ipython notebook that details every computation made in this paper.

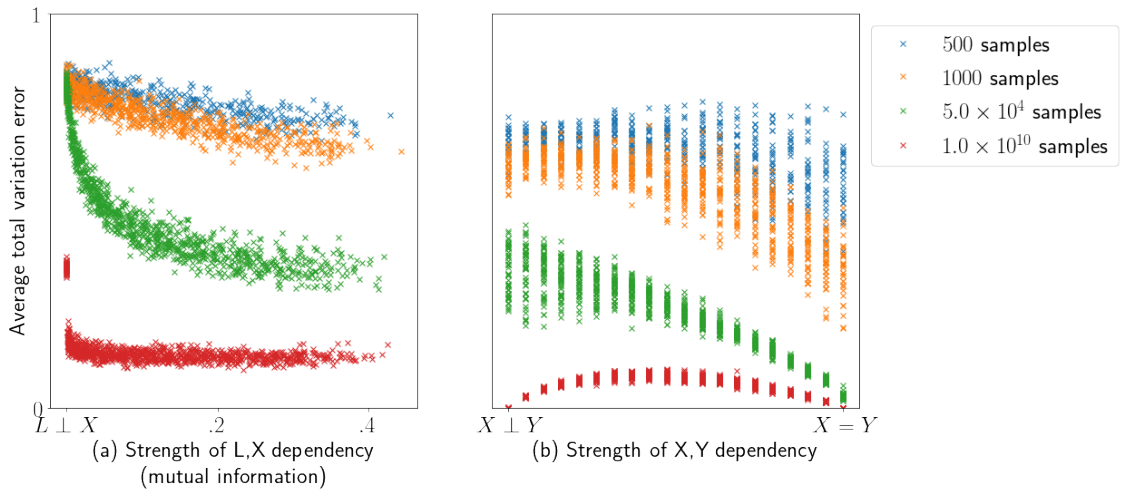

Figure 1: **Depending upon the exact model parameters, maximum likelihood estimation for β may fail.** We conducted a series of trials; in each trial we picked a sample size and ground-truth model parameters, simulated a dataset, calculated a maximum likelihood estimator for β, and compared it against the ground-truth. Even when the sample size is effectively infinite, this estimator can fail. Here we explore these failure modes. To produce subfigure (a) each row of β is drawn uniformly from the simplex; we explore how estimation error changes as we vary α to control the strength of the dependency between the variable L and X. When $L \perp X$, we see that maximum likelihood estimation fails significantly even with an essentially infinite number of samples. To produce subfigure (b), we draw each row of α uniformly from the simplex and vary β to control the dependency between X and Y. When $X \perp Y$ or $X = Y$ it is possible to get accurate estimates of β. However, in a more typical case where the relationship between X, Y is neither independent nor deterministic, we see that the maximum likelihood estimator fails to accurately determine β, even in the limit of essentially infinite data. Both plots are consistent with Lemma 1, which states that there is no estimator for β which will be consistent for every ground-truth parameter.

4 Simulations

The previous section gives generic theoretical results about inference for the parameter β. In this section we use simulations to get a more detailed understanding.

Our first simulations investigate Lemma 1, which states that there are cases where $N_L < N_X$ and β cannot be identified even in the limit of infinite data. We test this Lemma by exploring the performance of a maximum likelihood estimator for β on different simulated datasets. By varying the sample size and the ground-truth parameters of the model, we can explore this unidentifiability in detail:

- Varying the sample size and the strength of the L, X dependency. Fix $N_X = N_Y = 20$ and $N_L = 15$. We perform a series of trials. In each trial we select π and the rows of β uniformly from the simplex. The rows of α are chosen more carefully. We will first generate $\tilde{\alpha}_1$ by sampling the rows uniformly from the simplex. We then generate τ uniformly from the simplex and define $\tilde{\alpha}_2 = \mathbf{1}\tau^T$, i.e. every row of α_2 is given by τ. Finally, we select $c \in [0, 1]$ and define $\alpha = (1 - c)(\tilde{\alpha}_1) + c\tilde{\alpha}_2$. By varying c, we can control the mutual information between L and X. When $c = 0$ the variable L gives significant information about X; when $c = 1$ we have that L is independent of X. Once π, α, β are selected, we select a sample size $n = m$, simulate a dataset, compute a maximum likelihood estimator $\hat{\beta}$ (see Appendix A for our exact algorithm), and compute the average variation distance between β and our estimate $\hat{\beta}$. This average total variation distance is computed by averaging the row-by-row total

variations between β and $\hat{\beta}$. After performing a series of such trials, we visualize the results to obtain qualitative understanding of the performance of the maximium likelihood estimator $\hat{\beta}$.

- Varying the sample size and the strength of the X, Y dependency. Fix $N_X = N_Y = 20$ and $N_L = 15$. We perform a series of trials. For each trial, we select π and the rows of α uniformly from the simplex. To determine β, we first generate τ uniformly from the N_Y-simplex. We then select $c \in [0, 1]$ and define $\beta = c(\mathbf{1}\tau^T) + (1 - c)I$. By varying c, we can control the strength of dependency between X and Y. When $c = 0$ we have that $X = Y$ (degenerate dependency); when $c = 1$ we have that $X \perp Y$ (complete independence). As above, for each trial we then compute a maximum likelihood estimator and compare the estimator to the ground-truth parameter.

Figure 1 shows the results. We see that β is much harder to estimate in some situations. For example, the first simulation (Figure 1a) demonstrates that β is harder to estimate when L, X have low mutual information. The second simulation (Figure 1b) shows a more complex picture; in some cases it is actually possible to get perfect estimation of β, even though $N_L < N_X$. If X, Y are independent or deterministically related, we see that perfect estimation becomes possible in the limit of infinite data. However, in the more typical case where the relatioship between X and Y neither independent nor deterministic, it appears that consistent estimation is not possible.

We next investigate Theorem 4, which implies that the Markov Link Intervals are asymptotically guaranteed to include the true parameter values for β. Fix $N_X = N_Y = 20$ and $N_L = 15$. We perform a series of trials. In each trial we select π and the rows of α, β uniformly from the simplex. We then select a sample size $n = m$, simulate a dataset, and calculate the Markov Link Interval for the parameter $\beta_{1,1}$. For each sample size n we can plot a histogram of the differences between the upper bound $\hat{\beta}^+_{1,1}$ and the true parameter $\beta_{1,1}$. A small positive value is ideal, indicating a tight bound on the true parameter. Negative values are undesirable, since they indicate that the true parameter was in fact greater than the upper bound of the Markov Link Interval. Large values are also undesirable since they indicate the bound was loose. We can apply the same procedure to the differences between β and $\hat{\beta}^-$. By inspecting these histograms we obtian a qualitative sense for the performance of these intervals, complementing the theoretical guarantees of Theorem 4. The results are shown in Figure 2. We see that the true parameters are reliably bounded by the Markov Link Intervals, and these intervals shrink as we get more data. However, even in the limit of essentially infinite data ($n = 10^{10}$) we note that the width of the intervals does not vanish entirely. This is consistent with the unidentifiability of β demonstrated in Lemma 1.

These simulations are by no means exhaustive. However, they yield a high-level overview of the estimation problem for this model. First, estimation is easier when:

- L is highly informative about X.
- X, Y are either highly dependent or nearly independent
- Sample size is high

Second, even when estimation is difficult, the Markov Link Intervals give bounds which reliably contain the true parameter values.

5 Experiments for cell-types

We now apply Markov Link Intervals to study the scientific problem which motivated their development.

Every cell in a human body has the same DNA (to a first approximation), but some cells behave differently from others. The biomolecular processes that drive this diversity are an area of active research [8, 9, 10]. Efforts such as the Human Cell Atlas project seek to map out a taxonomy of cell-types [11], thereby enabling a more systematic study of cellular diversity.

At least at a coarse level, the cell-type X of a given cell can be reliably determined by applying RNA sequencing techniques [12]. However, these techniques only measure one property of the cell: RNA expression.

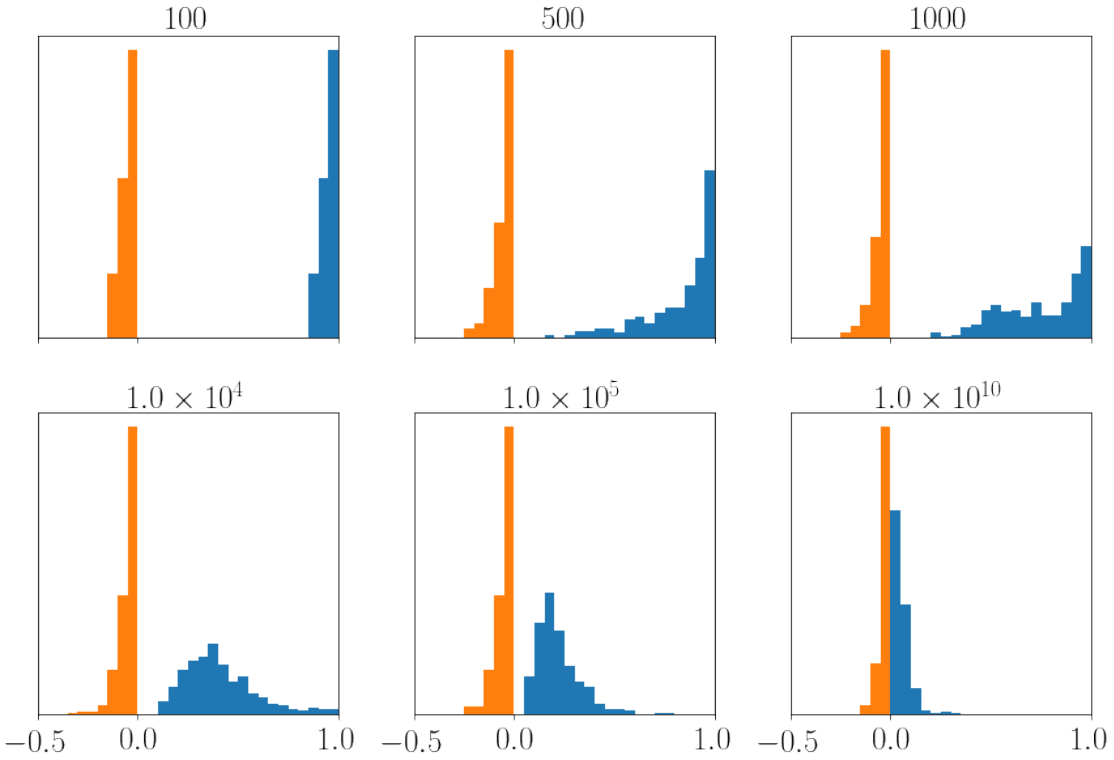

Figure 2: **The Markov Link Intervals are conservative bounds around the true parameter estimates.** We conducted a series of trials; in each trial we picked a sample size and ground-truth model parameters, simulated a dataset, calculated the Markov Link Intervals $\hat{\beta}^-, \hat{\beta}^+$, and compared them against the truth. For each trial we calculated both an upper slack $(\hat{\beta}^+ - \beta)$ and a lower slack $(\hat{\beta}^- - \beta)$. For each sample size $n \in \{100, 500, 1000, 10^4, 10^5, 10^{10}\}$, we computed a histogram of all the upper and lower slack values in all of the simulations performed with that sample size. We observe first that for every n the upper slack is always positive and the lower slack is always negative, indicating the true value was contained inside our bound for every simulation. Moreover, as the number of samples increases, we see that we can obtain tighter bounds; the magnitude of the upper and lower slacks decreases. However, even in the limit of essentially infinite data ($n = 10^{10}$) we note that the slack does not converge to zero. This is consistent with the unidentifiability of β demonstrated in Lemma 1.

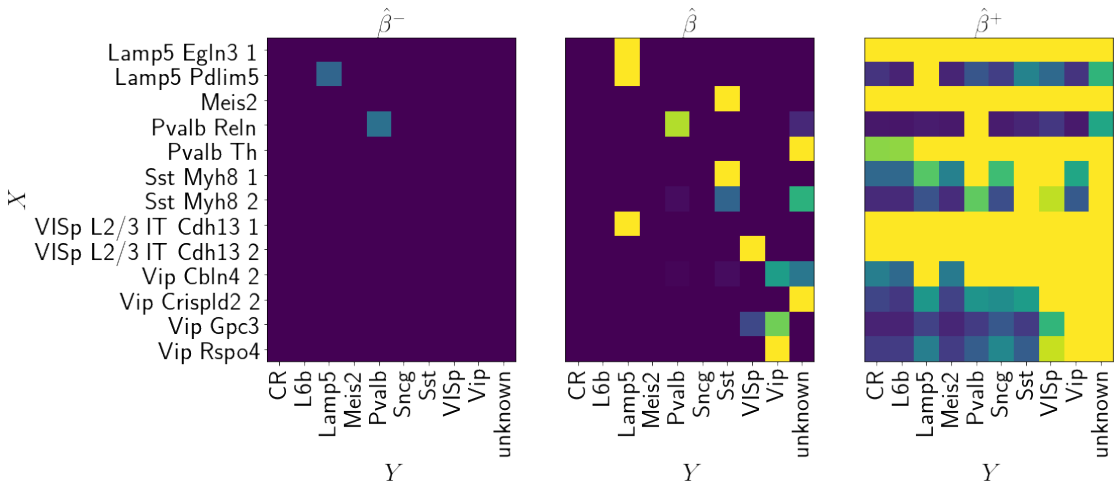

Figure 3: **Cell-type science: maximum likelihood estimation of β suggests unlikely results about cell biology; Markov Link Intervals indicate plausible bounds.** We performed unpaired measurement integration for two ways of measuring the 'type' of mouse cells. The first method uses gene expression to obtain a definitive measurement of the cell-type, X. The second method yields a considerably less reliable estimate for the cell-type, Y. For each cell we also know the value of L, which corresponds to another noisy indicator about the cell-type X. In one dataset we can observe (L, X) for many cells; in another dataset we can observe (L, Y) many other cells. Using this data we construct three estimates for β, the matrix of parameters which defines the relationship between X and Y. For each possible index (x, y), $\hat{\beta}^-_{x,y}$ estimates a lower-bound estimate for $\beta_{x,y}$, $\hat{\beta}_{x,y}$ is a maximum likelihood estimator, and $\hat{\beta}^+$ estimates an upper bound. Due to identifiability issues $\hat{\beta}_{x,y}$ is not guaranteed to be consistent. However, Theorem 4 shows that the Markov Link Interval $[\hat{\beta}^-_{x,y}, \hat{\beta}^+_{x,y}]$ is asymptotically guaranteed to include the truth. Above we show all three of these estimators for a selection of values of (x, y); we visualize these values in a grid, where rows correspond to values of x, columns correspond to values of y, and the color of the square in the (x, y) cell indicates the value of the corresponding parameter. The values of $\hat{\beta}$ are essentially nonsensical in many cases. On the other hand, the Markov Link Interval suggests plausible bounds for the parameters and also indicates further experiments which would help refine these bounds.

To determine the relationship between cell type and other properties of the cell, alternative techniques are required. "Patch-seq" is one such approach [13]; this method obtains RNA, electrophysiological, and morphological properties at the single-cell level. Unfortunately, the richness of the data comes at the cost of a degraded RNA signal, leading to a noisy estimate of the cell-type. We will denote this noisy estimate as Y.

For any cell it is thus impossible to measure a rich variety of cell properties *and* reliably estimate the cell-type. To understand the relationship between cell-type and these other cell properties, we therefore need some form of unpaired measurement integration. This integration would allow us to determine the reliability of the noisy estimates Y, and thereby determine whether the relationship between cell-type and other cell properties can be reliably determined from the Patch-seq approach. To facilitate this integration, we can construct experiments which make restrictions on which cells are are picked out, using a Cre/Lox system [14]. In each experiment, the probability that a particular cell will be picked out is entirely determined by the cell type X, and is thus statistically isolated from the noisy cell-type estimate Y. By varying the experimental parameters we can change the distribution on which cells are picked out. Letting L denote these corresponding experimental parameters, we see that L can act as a noisy independent indicator about the cell-type. Markov Link Intervals should therefore give asymptotically consistent bounds on the reliability of the noisy estimate Y.

The Allen Institute recently conducted a series of experiments using both traditional RNA sequencing as well as the Patch-seq technique [14]. The same Cre/Lox systems were used for both techniques, thereby allowing us to observe the indicator L for each sample in both datasets. For each sample in the traditional RNA sequencing dataset ("Standard") the cell-type was determined. For each sample in the Patch-seq dataset we also have an estimated cell-type, although the Patch cell-type estimates are only given at a coarse level of the taxonomy of cells. For example, in the Standard dataset a distinction is made between cell-type 'Lamp5 Egln 1' and cell-type 'Lamp5 Pdlim5,' whereas in the Patch-seq dataset we group all GABAergic neuronal cell-types characterized by high expression of 'Lamp5' into a single coarse 'cell class.'

We applied the Markov Link Intervals to the Allen Institute data try to determine the reliability of the noisy cell-type estimators Y arising from Patch-seq approach. We can visualize a maximum likelihood estimator as well as the Markov Link Intervals through heatmaps (see Figure 3)). We see that the maximum likelihood estimator sometimes gives essentially nonsensical results. For example, the maximum likelihood estimator suggests that a particular group of excitatory glutamatergic neurons from visual cortex are *always* assigned to a GABAergic inhibitory type by the Y variable. This is quite unlikely, since even the noisy measurements of Y can reliably determine between excitatory and inhibitory neurons [13, 15].

The Markov Link Intervals yield plausible bounds on the true model parameters. For example, the lower bounds of the Markov Link Interval suggests that a sizeable fraction of 'Lamp5 Pdlim5' cells are indeed classified as 'Lamp5' cells by the Patch method (note that the Patch method does not give fine-grained cell-type predictions, such as 'Lamp5 Pdlim5', but instead only gives coarse class-level cell-type predictions, such as 'Lamp5'). The upper bounds furthermore show that these cells are seldom classified as anything *but* 'Lamp5' cells or possibly the 'unknown' type which the Patch method assigns to cells which it cannot identify. For cells of type 'Sst Myh8-1', Markov Link Intervals give us considerably less certainty regarding the related type; the intervals indicate that the Patch method does *not* classify these cells as 'CR,' 'L6b,' 'Meis2,' 'Sncg,' or 'Vip' – but it does not indicate how they *are* being classified. At the extreme end, we have cells of type 'Lamp5 Elgn3-1,' for which the Markov Link Intervals are completely vacuous. These 'Lamp5 Elgn3-1' cells might get perfectly classified or they may get poorly classified; the available data cannot tell us either way.

Parameters with large intervals suggest new experiments to perform in order to more closely determine the model parameters. For example, the significant ambiguity for cells with standard type 'Lamp5 Egln 1' suggests we need a series of experiments with more distinct Cre/Lox experimental parameters that pick out cells of this kind. If we could find a Cre/Lox procedure that obtained many 'Lamp5 Egln-1' cells but no cells which measure as Patch type 'Pvalb,' this would show that the Standard 'Lamp5 Egln 1' is not associated with Patch type 'Pvalb.' On the other hand, if we could find Cre/Lox parameters that select many 'Lamp5 Egln 1' cells but *only* cells with Patch type 'Pvalb,' this would show the opposite. If these

additional experiments were performed, new Markov Link Intervals could be computed on the new data to determine what aspects of the link are still ambiguous.

6 Conclusions

We here study 'unpaired measurement integration,' the problem of estimating a joint distribution (X, Y) from two separate datasets, one with samples of X and another with samples of Y. Rigorous statistical analysis of this ill-posed problem are complicated by statistical error and identifiability issues. Here we characterize both issues in the discrete settings when we can observe a third variable L for each sample in both dataset, under the assumption that L and Y are independent conditioned in X. This assumption arises naturally if L gives noisy measurement about X and that noise is independent of Y. Under this assumption, we construct "Markov Link Intervals" which give provably consistent bounds on the joint distribution. Simulations bear out the consistency of these bounds and also suggest estimation is easiest performs best when L is predictive of X and the relationship between X and Y is either deterministic or completely independent. An application of the method to cell-type populations demonstrates the method's practical utility for unpaired measurement integration problems. One can imagine that these results could be extended to the continuous setting, though additional regularity conditions would be required.

References

[1] GD Makarov. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, 26(4):803–806, 1982.

[2] Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pages 46–54. Morgan Kaufmann Publishers Inc., 1994.

[3] Yanqin Fan, Emmanuel Guerre, and Dongming Zhu. Partial identification of functionals of the joint distribution of potential outcomes. *Journal of Econometrics*, 197(1):42–59, 2017.

[4] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[5] Carlos Daniel Mimoso Paulino and Carlos Alberto de Bragança Pereira. On identifiability of parametric statistical models. *Journal of the Italian Statistical Society*, 3(1):125–151, 1994.

[6] Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 13(Aug):2409–2464, 2012.

[7] Lang Tong, R-W Liu, Victor C Soon, and Y-F Huang. Indeterminacy and identifiability of blind identification. *IEEE Transactions on circuits and systems*, 38(5):499–509, 1991.

[8] Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, 2018.

[9] Camille Boudreau-Pinsonneault and Michel Cayouette. Cell lineage tracing in the retina: Could material transfer distort conclusions? *Developmental Dynamics*, 247(1):10–17, 2018.

[10] Li He, Guangwei Si, Jiuhong Huang, Aravinthan DT Samuel, and Norbert Perrimon. Mechanical regulation of stem-cell differentiation by the stretch-activated piezo channel. *Nature*, 555(7694):103, 2018.

[11] Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature News*, 550(7677):451, 2017.

[12] Benjamin W Okaty, Ken Sugino, and Sacha B Nelson. Cell type-specific transcriptomics in the brain. *Journal of Neuroscience*, 31(19):6939–6943, 2011.

[13] Cathryn R Cadwell, Athanasia Palasantza, Xiaolong Jiang, Philipp Berens, Qiaolin Deng, Marlene Yilmaz, Jacob Reimer, Shan Shen, Matthias Bethge, Kimberley F Tolias, et al. Electrophysiological, transcriptomic and morphologic profiling of single neurons using patch-seq. *Nature biotechnology*, 34(2):199, 2016.

[14] Bosiljka Tasic, Zizhen Yao, Kimberly A Smith, Lucas Graybuck, Thuc Nghi Nguyen, Darren Bertagnolli, Jeff Goldy, Emma Garren, Michael N Economo, Sarada Viswanathan, et al. Shared and distinct transcriptomic cell types across neocortical areas. *bioRxiv*, page 229542, 2017.

[15] Cathryn R Cadwell, Federico Scala, Shuang Li, Giulia Livrizzi, Shan Shen, Rickard Sandberg, Xiaolong Jiang, and Andreas S Tolias. Multimodal profiling of single-cell morphology, electrophysiology, and gene expression using patch-seq. *Nature protocols*, 12(12):2531, 2017.

A Markov Link Interval details

Let $n, m > 0$ and $(L_1, X_1, Y_1), (L_2, X_2, Y_2) \cdots (L_{n+m}, X_{n+m}, Y_{n+m}) \sim p$, where $p(\ell, x, y) = \pi_\ell \alpha_{\ell,x} \beta_{x,y}$. However, for $i \in 1 \cdots n$ we cannot observe Y_i and for $i \in n+1 \cdots n+m$ we cannot observe X_i. Let $\mathcal{L}(\theta)$ denote the likelihood of the observed data under the model under any particular choice of the parameters $\theta = \pi', \alpha', \beta'$. Let $\mathcal{L}^*$ denote the maximum value of $\mathcal{L}$ over all possible parameters. Define

$$\hat{\Theta} = \{\theta : \ \mathcal{L}(\theta) > \mathcal{L}^* - \log(n+m)\}$$

Our task is to calculate

$$[\hat{\beta}^-_{x,y}, \hat{\beta}^+_{x,y}] \triangleq \left[\min_{\pi',\alpha',\beta' \in \hat{\Theta}} \beta'_{x,y}, \max_{\pi',\alpha',\beta' \in \hat{\Theta}} \beta'_{x,y} \right]$$

Proceed as follows:

1. Summarize the data into two matrices:

$$\mathcal{D}^X_{\ell x} = \#\{i \le n : L_i = \ell, X_i = x\}$$
$$\mathcal{D}^Y_{\ell y} = \#\{i > n : L_i = \ell, Y_i = y\}$$

2. Estimate the maximum likelihood. The likelihood is given as

$$\mathcal{L}(\pi, \alpha, \beta) = \sum_{\ell x} \mathcal{D}^X_{\ell x}(\pi_\ell + \log \alpha_{\ell x}) + \sum_{\ell y} \mathcal{D}^X_{\ell y} \left(\pi_\ell + \log \sum_x \alpha_{\ell x} \beta_{x,y} \right)$$

   We optimize this likelihood using Expectation Maximization. Let

$$\tau_{\ell x y} = \alpha_{\ell x} \beta_{x,y} / \sum_x \alpha_{\ell x'} \beta_{x',y}$$

   denote a posterior summary for the distribution of the unobserved values of the measurement X. We then update the parameters by solving the surrogate problem

$$\max \left(\sum_{\ell x} \mathcal{D}^X_{\ell x}(\pi_\ell + \log \alpha_{\ell x}) + \sum_{\ell y} \mathcal{D}^X_{\ell y} \left(\tau_\ell + \sum_x \tau_{\ell x y} \log \alpha_{\ell x} \beta_{x,y} \right) \right)$$

Note that this surrogate problem can be solved in closed form. We proceed by iteratively updating our posterior summary then solving the the surrogate problem. We terminate the algorithm when all parameters change less than 10^{-7} in a single iteration.

3. Define

$$\hat{\Theta} = \left\{\theta : \ \mathcal{L}(\theta) > \mathcal{L}(\hat{\theta}) - \log(n+m)\right\}$$
$$[\hat{\beta}^-_{x,y}, \hat{\beta}^+_{x,y}] \triangleq \left[\min_{\beta' \in \hat{\Theta}} \beta'_{x,y}, \max_{\beta' \in \hat{\Theta}} \beta'_{x,y}\right]$$

Let x^*, y^* index a particular parameter for which we would like to compute $\hat{\beta}^-_{x^*,y^*}, \hat{\beta}^+_{x^*,y^*}$. We compute as follows:

- To compute $\min_{\beta' \in \hat{\Theta}} \beta'_{x^*,y^*}$. The Lagrangian of this constrained problem may be written as

$$V(\beta, \lambda) = \lambda \log \beta_{x^*,y^*} + L(\hat{\alpha}, \hat{\beta})$$

For any fixed λ, note that this problem has the same functional form as our original problem, and can be solved using the same algorithm as we used to find the maximium likelihood. The updates for α are in fact identical. The β updates are only slightly different; combining the posterior summary and objective updates into one step, we obtain the updates

$$\beta_{xy} \propto \begin{cases} \beta_{xy}\left(\frac{\lambda}{\beta_{xy}} + \sum_\ell \alpha_{\ell x} \frac{\mathcal{D}^X_{\ell y}}{\sum_x \alpha_{\ell x}\beta_{xy}}\right) & x = x^*, y = y^* \\ \beta_{xy}\left(\sum_\ell \alpha_{\ell x} \frac{\mathcal{D}^X_{\ell y}}{\sum_x \alpha_{\ell x}\beta_{xy}}\right) & \text{otherwise} \end{cases}$$

Let $\tilde{\beta}(\lambda, x^*, y^*)$ denote the optimizer of this inner problem. To solve our overall problem, we need to find the correct value of λ, i.e. $\max\{\lambda : \ \tilde{\beta}(\lambda, x^*, y^*) \in \hat{\Theta}\}$. We find this λ using line-search.

- To compute $\min_{\beta' \in \hat{\Theta}^\epsilon} \beta'_{x^*,y^*}$. This problem is equivalent to $\max_{\beta' \in \hat{\Theta}^\epsilon} \log \sum_{y \neq y^*} \beta'_{x^*,y}$. This allows us to pose the problem just as we did above. The resulting updates are given by:

$$\beta_{xy} \propto \begin{cases} \beta_{xy}\left(\frac{\lambda}{\sum_{y' \neq y} \beta_{xy'}} + \sum_\ell \alpha_{\ell x} \frac{\mathcal{D}^X_{\ell y}}{\sum_x \alpha_{\ell x}\beta_{xy}}\right) & x = x^*, y \neq y^* \\ \beta_{xy}\left(\sum_\ell \alpha_{\ell x} \frac{\mathcal{D}^X_{\ell y}}{\sum_x \alpha_{\ell x}\beta_{xy}}\right) & \text{otherwise} \end{cases}$$

B Proofs

Lemma (Unidentifiability)**.** *When $N_L < N_X$ there is no generally consistent estimator to estimate β from the data $(\mathcal{D}^X, \mathcal{D}^Y)$.*

Proof. It suffices to show that there exist parameter choices (π, α, β) and (π', α', β') where $\beta \neq \beta'$ yet the distribution on $\mathcal{D}^X, \mathcal{D}^Y$ remains the same. In this case there would be not way to determine which of these paramer choices was correct, so no consistent estimator could be produced. Note that the distribution on $\mathcal{D}^X, \mathcal{D}^Y$ is entirely determined by π, α, and $\gamma = \alpha\beta$. Thus it suffices to find an example where $\alpha = \alpha', \beta \neq \beta', \alpha\beta = \alpha'\beta'$. For a simple example, consider the case $N_L = 1, N_X = N_Y = 2$ with

$$\alpha = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \end{pmatrix} \qquad \beta = \begin{pmatrix} 0.5 & 0.5 \\ 0.1 & 0.9 \end{pmatrix} \qquad \beta' = \begin{pmatrix} 0.8 & 0.2 \\ 0 & 1 \end{pmatrix}$$

□

Lemma (Discontinuity of observational equivalence classes)**.** *Endow the space of observable distributions with the total variation metric. Endow the space of sets of parameters π, α, β with the Pompeiu-Hausdorff metric. The map from possible distributions on $\mathcal{D}^X, \mathcal{D}^Y$ to observational equivalence classes of parameters is not continuous. This mapping is sometimes discontinuous even around points where the rows of α are linearly independent and all parameters lie in the interior of the natural parameter space.*

Proof. We start by inspecting the two relevant spaces: the space of distributions on $\mathcal{D}^X, \mathcal{D}^Y$ and the space of sets of parameters π, α, β.

- The possible distributions on $\mathcal{D}^X, \mathcal{D}^Y$. Each distribution is parameterized by the triple (π, α, β), where π lies in the probability simplex and α, β are right-stochastic matrices. However, note that this parameterization is degenerate with respect to our chosen metric; any two points in this space, (π, α, β), (π', α', β'), are considered equal as long as $\pi = \pi', \alpha = \alpha'$ and $\alpha\beta = \alpha'\beta'$. In particular, the total variation distance between these two points is topologically equivalent to the $\mathscr{L}^1$ distance between for triples $(\pi, \alpha, \alpha\beta)$, $(\pi', \alpha', \alpha'\beta')$.
- Sets of parameters. Let Θ, Θ' denote any two sets of parameters. The Pompeiu-Hausdorff distance is the largest Euclidean distance from any point in either set to the closest point in the other set.

We are interested in the mapping from the first space to the second induced by observational equivalence classes, i.e.

$$\pi, \alpha, \beta \mapsto \{(\pi, \alpha, \beta') : \ \alpha\beta = \alpha\beta'\}$$

To show that this is discontinuous at some point π, α, β with respect to our chosen metrics, it suffices to find a sequence $\tilde{\pi}(\epsilon), \tilde{\alpha}(\epsilon), \tilde{\beta}(\epsilon)$ indexed by ϵ such that that

- $\tilde{\pi}(\epsilon), \tilde{\alpha}(\epsilon), \tilde{\alpha}(\epsilon)\tilde{\beta}(\epsilon) \to \pi, \alpha, \alpha\beta$ in the $\mathscr{L}^1$ metric as $\epsilon \downarrow 0$
- $\Theta(\epsilon) \triangleq \{(\pi(\epsilon), \alpha(\epsilon), \beta') : \ \alpha(\epsilon)'\beta' = \gamma\}$ fails to converge to $\Theta(0) \triangleq \{(\pi, \alpha, \beta) : \ \alpha\beta' = \gamma\}$ in the Pompeiu-Hausdorff metric.

We exhibit such a case. Fix any π strictly on the interior of the simplex. Let $\Theta(0) = \{(\pi, \alpha, \beta') : \ \alpha\beta' = \alpha\beta\}$, where

$$\alpha = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{pmatrix} \qquad \beta = \begin{pmatrix} 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad \alpha\beta = \begin{pmatrix} \frac{1}{4} & \frac{3}{4} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

For each $\epsilon > 0$ let $\Theta(\epsilon) = \{(\pi, \tilde{\alpha}(\epsilon), \beta') : \ \tilde{\alpha}(\epsilon)\beta' = \tilde{\alpha}(\epsilon)\tilde{\beta}\}$, where

$$\tilde{\alpha}(\epsilon) = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} - \frac{\epsilon}{2} & \frac{1}{4} + \frac{\epsilon}{2} \\ \frac{1}{3} & \frac{1}{3} - \frac{\epsilon}{3} & \frac{1}{3} + \frac{\epsilon}{3} \end{pmatrix} \qquad \tilde{\beta} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \tilde{\alpha}(\epsilon)\tilde{\beta} = \begin{pmatrix} \frac{1}{4} - \frac{\epsilon}{2} & \frac{3}{4} + \frac{\epsilon}{2} \\ \frac{1}{3} - \frac{\epsilon}{3} & \frac{2}{3} + \frac{\epsilon}{2} \end{pmatrix}$$

Now observe:

- $\pi, \tilde{\alpha}(\epsilon), \tilde{\alpha}(\epsilon)\beta \to \pi, \alpha, \alpha\beta$ in $\mathscr{L}^1$ metric as $\epsilon \downarrow 0$
- $\Theta(\epsilon)$ fails to converge to $\Theta(0)$. Indeed consider any $\beta' \in S(\epsilon)$. Algebraic manipulations yield that $\beta'_{11} = 2(\beta'_{21} - 1)\epsilon$ for every $\beta' \in S(\epsilon)$. This is a negative value unless $\beta_{21} \geq 1$, ergo $\beta' \in S(\epsilon)$ is only a valid right-stochastic matrix if $\beta'_{21} = 1$. On the other hand, $\beta \in S(0)$ satisfies $\beta_{21} = 1/2$. It follows that the Pompeiu-Hausdorff from $\Theta(0)$ to $\Theta(\epsilon)$ is at least $1/2$ for every $\epsilon > 0$.

This shows that the mapping is discontinuous at π, α, β. Furthermore note that the rows of α are linearly indendent. Moreover, π lies strictly on the interior of the simplex, as does every row of α and $\alpha\beta$. Thus the distribution corresponding to π, α, β lies strictly on the interior of the natural parameter space. □

Theorem (Consistency for positive β)**.** *Let the rows of α be linearly independent and let $\beta_{x,y} \geq c > 0$. Define γ as the $N_L \times N_Y$ right-stochastic matrix*

$$\gamma_{\ell y} \triangleq \sum_x \alpha_{\ell x} \beta_{x,y}$$

Let $\hat{\pi}, \hat{\alpha}, \hat{\gamma}$ denote any consistent estimators for π, α, γ. Consider $\{\beta' : \ \hat{\alpha}\beta' = \hat{\gamma}\}$, the equivalence class for the observations $\mathcal{D}^Y$ after fixing the estimate $\hat{\alpha}$. The distance between the point β and this set converges to zero in probability.

Proof. Fix any α, β. It suffices to show that by ensuring $|\alpha' - \alpha| < \delta, |\gamma' - \alpha\beta| < \delta$ for δ sufficiently small we can ensure that β is arbitrarily close to some valid right-stochastic matrix in the set $\{\beta' : \ \alpha'\beta' = \gamma'\}$. Let $\tilde{\beta}$ denote the orthogonal projection of β to this set. Note that since the rows of α are independent we can ensure the rows of α' are also independent by taking δ sufficiently small. In particular, we can bound the spectral norm of the pseudoinverse of α', and thereby ensure that $\left|\tilde{\beta}_{x,y} - \beta_{x,y}\right| < \epsilon$ for every x, y by taking δ sufficiently small. Finally, the positivity of β allows to ensure that $\tilde{\beta}$ is a valid right-stochastic matrix. Thus $\tilde{\beta}$ is indeed a point in the observational equivalence class which is arbitrarily close to β. □

Theorem (Consistency of $\hat{\Theta}$)**.** *Let $n, m \to \infty$ such that $|\log(n/m)| < M$. Then:*

1. *$\hat{\Theta}$ is asymptotically guaranteed to contain the true parameter values.*
2. *$\hat{\Theta}$ concentrates for $\pi, \alpha, \alpha\beta$. Let $\xi(\pi', \alpha', \beta')$ denote the maximum of the Euclidean differences $|\pi - \pi'|$, $|\alpha - \alpha'|$, $|\alpha\beta - \alpha'\beta'|$. Then $\max_{\theta \in \hat{\Theta}} \xi(\theta)$ converges to zero in probability.*

Proof. A corollary of Lemma 5, below. A minor correction is necessary to deal with the fact that we have two sets of independent samples instead of a single list of independent samples. Fix any $n' > 0, m' > 0$. Arrange the $\mathcal{D}^X, \mathcal{D}^Y$ into T independent groups of samples, each containing $n' > 0$ samples from $\mathcal{D}^X$ and $m' > 0$ samples from $\mathcal{D}^Y$. Drop samples at the end that cannot be fit into even groups. Since we have required $|\log(n/m)| < M$, note that can obtain an arbitrarily large number of such groups in the limit. We can now understand our data as a list of T independent samples drawn from a distribution p^Π. Here

$$p^\Pi(\ell_1, x_1 \cdots \ell_{n'}, x_{n'}, \ell_{n'+1}, y_{n'+1} \cdots \ell_{n'+m'}, y_{n'+m'})$$

is defined as the product of the corresponding distributions arising from p.

Now apply Lemma 5 to this grouped data in the limit as $T \to \infty$, where the set S indicates the image of our parameterization of p. The lemma shows that the true parameter π, α, β is asymptotically guaranteed to lie in the set $\hat{\Theta}$; this proves the first claim of our theorem. Second, the theorem shows that π', α', β' will be asymptotically excluded as long as $p^\Pi(\cdot; \pi', \alpha', \beta') \neq p^\Pi(\cdot; \pi, \alpha, \beta)$. Observe that $p^\Pi(\cdot; \pi', \alpha', \beta') \neq p^\Pi(\cdot; \pi, \alpha, \beta)$ if and only if π', α', β' lies in the true observational equivalence class, i.e. $\pi = \pi', \alpha = \alpha', \alpha\beta = \alpha'\beta'$. Thus any set of parameters whose Kullback-Leibler divergence from the truth is bounded away from zero will asymptotically fail to intersect $\hat{\Theta}$. To conclude our proof of the second claim, note that Pinsker's inequality lower-bounds this divergence by the Euclidean distance on the space of the observable distributions $\pi, \alpha, \alpha\beta$. □

Lemma 5. *Let $(W_1 \cdots W_n)$ i.i.d $\sim p^*$, where p^* is a probability mass function on a finite number of atoms. Let $\mathcal{L}(p) = \log \sum_i \log p_{W_i}$ denote the log likelihood for any p on the simplex. Let S denote any subset of the simplex with $p^* \subset S$. Let $\hat{p} = \arg\max_{p \in S} \mathcal{L}(p)$ denote a maximum likelihood estimator. Fix any $\delta > 0$ and define*

$$\hat{\Theta} = \{\theta : \ \mathcal{L}(\theta) > \mathcal{L}(\hat{\theta}) - \delta \log n\}$$

Then:

1. *The probability that $p^* \in \hat{\Theta}$ is asymptotically unity.*

2. *Let $D(p^*||p)$ denote the Kullback-Leibler divergence from p^* to p for any p on the simplex. Let $\Gamma(\epsilon) = \{p : \ D(p^*||p) > \epsilon\}$. The probability that $\Gamma(\epsilon)$ and $\hat{\Theta}$ intersect asymptotically vanishes for any $\epsilon > 0$.*

Proof. The proof is straightforward because we have required p is a probability mass function with a finite number of atoms.

1. $p^* \in \hat{\Theta}$ with high probability. Since $\log n \to \infty$ it suffices to show that $\limsup_{n\to\infty} \mathbb{E}[\mathcal{L}(\hat{p}) - \mathcal{L}(p^*)]$ is bounded. The only slight subtlety here is that the estimator $\hat{p}$ is required to lie in the arbitrary set S – otherwise the usual asymptotic theory would immediately give that $\mathcal{L}(\hat{p}) - \mathcal{L}(p^*)$ is asymptotically χ^2 with $\#\{x : \ p(x; p^*) > 0\} - 1$ degrees of freedom. However, since $\hat{p}$ arises from optimizing over a *restricted* domain, it follows that $\mathcal{L}(\hat{p}) - \mathcal{L}(p^*)$ is actually stochastically dominated by this same χ^2 distribution. Thus the expectation in question is bounded.

2. Incorrect parameters are excluded. First observe that the result holds pointwise, i.e. for any $\epsilon > 0, p \in \Gamma(\epsilon)$ we have that p is asymptotically excluded from $\hat{\Theta}$. To see this, observe that the central limit theorem yields, asymptotically, that

$$V(p) \triangleq \frac{1}{\sqrt{n}}(\mathcal{L}(p^*) - \mathcal{L}(p) - nD(p^*||p)) \sim \mathcal{N}\left(0, \sigma_p^2\right)$$

where σ_p^2 is the variance of the corresponding single-sample estimator for the KL divergence. Thus

$$\begin{aligned}\mathbb{P}\left(p \in \hat{\Theta}\right) &= \mathbb{P}\left((\mathcal{L}(\hat{p}) - \mathcal{L}(p^*)) + (\mathcal{L}(p^*) - \mathcal{L}(p) - nD(p^*||p)) < \delta \log n - nD(p^*||p)\right) \\ &= \mathbb{P}\left(V(p) < \delta\frac{\log n}{\sqrt{n}} - \sqrt{n}D(p^*||p) - (\mathcal{L}(\hat{p}) - \mathcal{L}(p^*))/\sqrt{n}\right)\end{aligned}$$

This probability vanishes whenever $D(p^*||p) > 0$, since the right-hand-side of the inequality inside the probability expression escapes to $-\infty$. Thus every $p \in \Gamma$ is eventually excluded from $\hat{\Theta}$. To conclude our proof of the second claim, note that $\hat{\Theta}$ is a convex set whose eccentricity is bounded by the ratio between the lowest and highest nonzero value in the empirical distribution of the Xs; since there are a finite number of atoms this eccentricity is thus asymptotically bounded. Moreover, note that $p \mapsto D(p^*||p)$ is convex and uniformly continuous in neighborhoods of p^*. We can therefore find a grid of points $p \in \Gamma(\epsilon/2)$ whose convex hull contains p^*, such that any convex set which does not intersect those points must lie in the complement of $\Gamma(\epsilon)$. By applying the pointwise result to this finite collection of points, we obtain our uniform result.

□