Markov Link Method for calibrating without joint measurement, including the case of destructive measurements

Jackson Loper

April 20, 2018

Abstract

A proliferation of new experimental tools has left a serious gap: calibration. Two thermometers can be calibrated against each other by simply measuring the same bodies of water with both thermometers, but the problem is much harder for many modern tools. One common problem is that we do not have measurements from the same "body of water" for both tools. We propose the Markov Link Method (MLM) as a way to overcome this difficulty. This method produces consistent estimators that tightly bound the calibration, i.e. the conditional distribution of one tool's measurement given another tool's measurement. It achieves this without any measurement data from both tools applied to the same "bodies of water." Moreover, MLM makes zero assumptions about what calibrations we might expect to see, instead applying a subpopulation-based conditional independence assumption. We evaluate MLM on a pair of single-cell RNA techniques, obtaining precise calibrations between the tools as well as more accurate models for each tool separately.

The modern setting is rife with experimental measurement tools, and it can be very frustrating to understand how the output of these tools relate to one another. This problem is known as "calibration" or "zeroing" [1]. A calibration tells us what readings we should expect from one tool, given the reading we obtained from another tool. Calibration additionally must give uncertainty bounds for how much we can trust those expectations. Calibration between measurement tools allows us to combine experimental results from different labs and different methodologies into larger scientific theories.

Formally, a calibration is simply a conditional distribution. As input, this conditional distribution takes the measurement results from one tool on a particular specimen. As output, it yields a distribution on what a second tool would measure on the same specimen. One way to learn the calibration is to measure the same specimens under both tools. We call this "joint measurement." Unfortunately, calibrations are often required even when joint measurement is unavailable. For example, if the measurement tool alters the specimen being measured, joint measurement is simply impossible. In other cases, it may be expensive or impractical.

We here propose the Markov Link Method (MLM) to estimate calibrations between tools. The MLM can be trained without any joint measurement. The key idea is to use multiple subpopulations of specimens. If each subpopulation captures a different slice of the overall population, we can obtain tight bounds on the true calibration. This is true even if each subpopulation is highly heterogeneous. By integrating information from all the subpopulations we can make rigorous deductions about what the calibration might be. MLM also gives suggestions about which further subpopulations might be helpful to study in order to further refine our knowledge of the true calibration. The method can also be naturally extended to calibration distributions among many tools.

Some example applications include:

- Quality control for manufacturing. The surest way to test the reliability of a part is to construct a machine that pushes the part until it breaks. However, how can we test the reliability of the machine that performs the test? In each test run there will be some variability induced by the machine itself, which induces a measurement error In practice, some kind of assumptions about part homogeneity are used to approximate this error (cf. [2]). However, if we have multiple testing machines we can use the MLM to obtain a calibration between the machines, even though we can never test the same part with both machines. This enables us to bound the overall measurement error.

- Combining knowledge across experimental modalities: morphology and transcriptomics. There are different ways to think about the different types of cells in an organism. A traditional approach is to classify cells based on what they look like (cf. [3, 4]). A more modern approach is to assay the cell's transcriptome (cf. [5]). Unfortunately, modern high-resolution cell photography and single-cell sequencing technologies are both destructive. As a result, we can't get both kinds of data for the same specimens. For cells native to regions full of diverse cell-types, finding a correspondence is a real problem. The result is two completely independent classifications of cells, one for each way of looking at the cell. MLM allows us to estimate the relationship between those two classification systems, yielding a wholistic understanding of the different types of cells.

- Cancer treatment efficacy prediction. Starting from in-vivo human cancers, many cell-lines have been cultured over the years. These cell cultures live indefinitely on plates. Many experiments have been performed to see how these cancer cells respond to treatment. However, if a treatment work on a particular cultured cell-line, what can we say about which kinds of cancer might respond well to that treatment? Coarse side-information such as original cancer location is often available for both in-vivo and cultured cells, but this is often a surprisingly weak signal. Cell transcriptomes provides much more specific information about the cancer, and thus, in theory, what treatments might be appropriate (cf. [6]). However, we know that cultured cell-lines look quite different from in-vivo cells (cf. [7, 8]). Moreover, we have very little joint measurement between the human cells which originally gave rise to a cell culture and the cells that survived to beceome the cell line. The Markov Link method can leverage the common side-information together with separate transcriptome information to produce a fine-grained correspondence between in-vivo and cultured cells. This correspondence can used to propose new treatments for cancers.

- Text/image correspondence. Automatic image captioning is an ongoing effort in machine learning (cf. [9]). There are three types of data available to help develop such algorithms: text-only data, image-only data, and paired-text-and-image data. Obviously the last kind is the most useful for automatic image captioning, but there is much less of it. The Markov Link Method suggests one way to use the more plentiful text-only and image-only data. We can first apply classic machine learning techniques to get coarse labels for both kinds of data. Using this side-information to identify subpopulations, the MLM can then deduce a fine-grained correspondence between text and images by combining information from across all the subpopulations.

- Replication crisis and lab effects. Replicating a published study is not always an easy thing to do. This difficulty is commonly attributed to selective publication bias, bad design, poor description of methods, and even outright fraud [10]. A calibration would allow us to understand this problem in detail. If two labs perform identical experiments and get different data, that does not mean we need to throw out both datasets. Instead, we can use MLM to calibrate the tools. Once the tools are properly

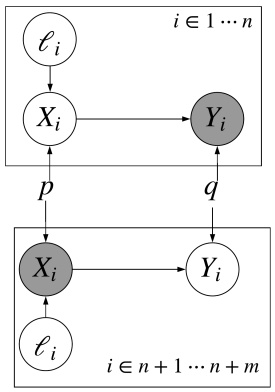

Figure 1: We assume that we have two datasets, drawn from different techniques applied to specimens from various subpopulations. In the top dataset, the results of technique II, denoted by Y_i, are hidden from us (we indicate this by making the circle gray). In the bottom dataset, the results of technique I, denoted by X_i, are hidden from us. In each dataset, we know what subpopulation of specimens we are considering, and this is designated by ℓ_i. Thus, for any given specimen i we can observe either ℓ_i, X_i or ℓ_i, Y_i but never ℓ_i, X_i, Y_i. Using observations of X_i we can certainly learn the parameters p which govern the relationship between X_i and the subpopulation, ℓ_i. Likewise we can determine the relationship between Y_i and ℓ_i. However, if we believe that the relationship between X and Y is the same regardless of the population value ℓ, we can also learn something about the parameters q which govern the relationship *between* X and Y.

calibrated, we can combine both datasets. Unlike other tools to deal with lab or batch effects (e.g. [11, 12]), MLM makes zero assumptions about what calibrations we might expect.

We apply MLM to the inferred cell-types of single neurons, as determined by two different methods for single-cell RNA sequencing. Technique I is known to have high fidelity, whereas technique II is less accurate. However, technique II also carries additional electrophysiological information which we would later like to use for other purposes. We show how the cell-type inferences from the two techniques can be connected. Our method yields results which are generally consistent with the scientist's understanding of the data and clusters, but also reveal potential gaps which could be important for further study.

This work stands on the shoulders of a long history of relating probabilistic assumptions to probabilistic inequalities. Much of this literature comes from research into causality. For example, in [13] Bonet uses polytopes not unlike the ones seen here to explore whether a variable can be used as an instrument. The famous Clauser-Horne-Shimony-Holt inequality was designed to help answer causality questions in quantum physics, but it also sheds light on what distributions are consistent with certain assumptions [14]. Indeed the physics literature has contributed many key inequalities (cf. [15], [16], and the references therein). Perhaps the closest work to this one would be [17], which used two marginal distributions to get bounds on a property of the joint distribution (namely the distribution of the sum). We advance this approach to a more general-purpose technique, both by using many subpopulations to refine our estimates and by considering the entire space of possible joint distributions instead of simply a particular aspect of the joint.

1 Mathematical formulation

Mathematically, we formulate our model as follows:

- Let us say we have $n+m$ individual members of a population. These could be cells, humans, or whatever the smallest sample unit may be for a given problem.
- For each individual member, we have some readily observable features by which we can group the members, such as the size of a cell, where an individual lives, or some other basic information. We designate these features as ℓ_i for the ith individual.
- For the first n members, we have made an observation using the first technique. This gives us the values $X_1 \cdots X_n$.
- The last m members were observed using the second technique. This gives us the values $Y_{n+1} \cdots Y_{n+m}$.

For simplicity, we here assume that ℓ_i, X_i, Y_i take values in finite sets, $\Omega_\ell, \Omega_X, \Omega_Y$. Extensions to more general cases should be straightforward, but we must leave them to future work.

We further posit a set of *counterfactual* variables – variables which could not ever be obtained in practice, but which help us organize our thinking. These are sometimes referred to as "potential outcomes" (cf. [18]).

- Let $Y_1 \cdots Y_n$ denote the observations we *would have gotten* if we had applied the second technique to the first n members.
- Let $X_{n+1} \cdots X_n$ denote the observations we would have gotten if we had applied the first technique to the last m members.

Our main assumption is a kind of Markov assumption, from which we obtain the name "Markov Link Method":

$$\mathbb{P}(X_i = x, Y_i = y | L_i = \ell) = p^*(x_i|\ell_i) q^*(y_i|x_i) \tag{1}$$

for some distributions p^*, q^*. Moreover, we assume that the vectors $\{(p^*(x_1|\ell) \cdots p^*(x_n|\ell))\}_\ell$ are linearly independent. If this seems unlikely, it may make sense to collapse similar values of ℓ together. The Markov assumption can be visualized with the plate diagram found in Figure 1.

The validity of these assumptions for a given situation should be closely contemplated. There are several key questions to answer when deciding whether this assumption is applicable. Does the process by which samples are gathered gathered depend only upon ℓ? In particular, is it not at all statistically related to which measurement technique was applied, except through ℓ? Are the particular measurement biases of both techniques the same for every value of ℓ? Is the statistical relationship between the quantities being measured by the two techniques the same for every value of ℓ?[1] If the answer to all of these questions is yes, the assumption that (X_i, Y_i) is drawn from $p(x_i|\ell_i)p(y_i|x_i)$ may apply. If some answer is no the method may still apply, but certain corrections may be necessary; we discuss these in our conclusions.

Under these key assumptions, the Markov Link Method gives us a way to uncover something about the relationship between the observations X_i from one measurement technique and the observations Y_i from the other measurement technique, i.e. the distribution q^*. Unfortunately, given the limitations of the data, q^* itself may be dramatically unidentifiable

[1] Note that this is automatically true if both techniques are measuring the same thing. More generally, if two techniques have something that they both measure, we can restrict attention to that one common phenomenon.

in many cases. Instead, the Markov Link Method yields a *set* of possible values of q which are consistent with the data. It can then be shown that the true value q^* lies close to this set with high probability.

To understand the limitations of the data more clearly, consider the case that ℓ lies in the set $\{1,2\}$, X lies in the set $\{1,\cdots 100\}$ and Y lies in the set $\{1,2\}$. As long as m,n are large enough, we can effectively estimate p^* and $h^*(Y|\ell) = \sum_x p^*(x|\ell)q(Y|x)$. However, this does not tell us very much about what q might be. Indeed, q lies in a 100-dimensional space, and the restriction that q must satisfy $h^*(Y|\ell) = \sum_x p^*(x|\ell)q(Y|x)$ for each (ℓ, Y) actually only introduces 2 new constraints. Thus there is a 98-dimensional space that q may lie in, and we simply cannot know where in that space q may lie. However, in practice, we find that the *inequality constraints*, namely that $q(y|x) \geq 0$ for every (x,y), can actually force this potentially vast space to be tightly centered around a single point.

In formal terms, the Markov Link Method produces a estimator for the set of possible values of q which are consistent with what is observable, i.e.

$$\Theta^* \triangleq \left\{ q : \sum_y p^*(x|\ell)q(y|x) = \sum_y p^*(x|\ell)q^*(y|x) \ \forall \ell, y \right\}$$

Let us see now how MLM achieves this.

2 Algorithm

The Markov Link Method begins by estimating

- p^* using the empirical distribution of the observations $\{X_i\}$, $\hat{p}(x|\ell)$
- $h^*(y|\ell) \triangleq \sum_x p^*(x|\ell)q^*(y|\ell)$ using the empirical distribution of $\{Y_i\}$, $\hat{h}(y|\ell)$

These empirical distributions may not be quite consistent with the original assumption that the distribution of $X,Y|\ell$ may be written as $p^*(x|\ell)q^*(y|x)$. In particular, it may be that there is *no* value of q such that $\sum_x \hat{p}(x|\ell)q(y|x) = \hat{h}(y|\ell)$. There may also be many such values. To obtain a well-defined estimator, we take a value of q which is decent, and use a slight regularization to ensure it is uniquely defined. Let $N_{X,\ell} = \#\{i \leq n : \ \ell_i = \ell\}, N_{Y,\ell} = \#\{i > n : \ \ell_i = \ell\}$. We take

$$\hat{q} = \arg\max_q \sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log\left(\sum_x \hat{p}(x|\ell)q(y|x) \right) + \kappa \sum_{xy} \log q(y|x)$$

In practice, we took the entropic regularizer κ to be 0.01; it is mostly useful for ensuring numerical stability of the algorithm.

This method gives us an estimate for $\hat{q}$, but we emphasize this estimate is very likely to be *inconsistent* for the true value of q^*. As we described in the previous section, this is simply a limitation of the data we have available – attempting to deduce q^* may be too much to ask for. However, subject to suitable assumptions, the estimator

$$\hat{\Theta} \triangleq \left\{ q : \sum_y \hat{p}(x|\ell)q(y|x) = \sum_y \hat{p}(x|\ell)\hat{q}(y|x) \ \forall \ell, y \right\}$$

is indeed consistent in the sense that we can be asymptotically assured that the true q^* lies very close to some point in $\hat{\Theta}$.

Theorem. *Fix any $\kappa > 0$. Let $N_{X,\ell}, N_{Y,\ell} \to \infty$ in such a way that $N_{Y,\ell'} / \sum_\ell N_{Y,\ell} \geq \rho > 0$ for each ℓ'. Let us assume that $q^*(y|x) > 0$ for every x, y and the rows of p^* are linearly independent. Then $\inf_{q \in \hat{\Theta}} \sup_{x,y} |q^*(y|x) - q(y|x)| \to 0$ in probability.*

Proof. We defer the proof to Appendix B. □

We close by remarking that the assumption $q^*(y|x) > 0$ is very likely to be unnecessary. However, it substantially simplifies the proof. We therefore leave a more complete theorem for future work.

3 Empirical results

Our motivation for this problem arose from looking at Allen Institute cell-type assignment of cells, performed using two different experimental techniques (also called experimental "modalities"). The only thing connecting the two modalities was that each modality had a variety of samples drawn from a variety of sub-populations, and the same sub-populations were used for both modalities. Thus our input was two tables: (technique I cell-type × sub-population), and (technique II cell-type × sub-population) (Figure 2). A key question immediately appears from these tables: does the way in which we have clustered cells into different types under technique I line up nicely with the way in which we have clustered cells under technique II? The Markov Link Method gives us a way of addressing this question that is completely agnostic about the underlying genetic markers that were used to produce the clusters (this in constrast to methods such as [19] or [11] which assume both that the clusters arise from observables common to both techniques and that those observables are already at least marginally calibrated between the methods). Our output was a set $\hat{\Theta}$ of possible ways that technique I cell-types could line up against the technique II cell-types. That is, each element of $\hat{\Theta}$ was itself a table of the form (technique I cell-type × technique II cell-type).

3.1 Making sense of Markov Link Method output

The Markov Link Method yields a set (indeed, a convex polytope). Each element of this set gives a particular way that technique I might be related to technique II. It is not immediately easy to understand how to use or visualize the polytope $\hat{\Theta}$ of possible values for q. Ultimately we want to know something about the correspondence between the techniques, i.e. q^* – how can we make good use of a *set* of possible corresponences? Here we suggest several main tools for analysis:

- Rotationally Uniform eXtremal samples (RUX). Every point in the polytope is a convex combination of its vertices, so by understanding the extremal vertices of the polytope we understand much about what values are possible. Unfortunatelly, there are a truly vast number of vertices, and looking at all of them is quite impossible. Instead, we propose a method for drawing samples which are illustrative of the overall character of the polytope: Rotationally Uniform eXtremal samples. These samples are drawn simply by selecting a random direction in the polytope, and finding the point in the polytope which is furthest along that direction. With probability 1 this point is unique, and provides a useful notion of the polytope's extent. Obtaining such a sample simply requires solving a linear programming problem. We show 10 such samples in Figure 3.

  Each sample shows a possible way that the two different techniques might correspond; together, these samples reveal both what the data says about the correspondence and

Figure 2: Input: two tables. The Allen Institute had access to various cre-line-based cell selection techniques. Each technique pulls out a different group of cells. Once the cells were selected, they were then either subjected to technique I ('facs') or technique II ('patch'). Technique I gives a very complete analysis of the gene expression of the cell. Technique II gives a less complete analysis, but yields additional electrophysiological data that may be of interest. The results of technique I was used to assign cells to one of 116 categories, and the results of technique II was used to assign cells to one of 14 categories. The tables above show the distribution of category assignment for each sub-population selection technique. The goal is to use these tables to be able to say something about how the 14 categories from Technique II might line up with the 116 categories from technique I.

what we still need to know. In particular, some entries of correspondence are nearly the same for every sample, and others can vary more significantly. This suggests what possible experiments might be important to do next. Since each row always sums to one, any ambiguity must lie in trading off between different columns within a row. To resolve this ambiguity, we could find a subpopulation which we believe might include, say, half of those ambiguous columns, and then perform additional experiments in both modalities, drawing from that subpopulation.

- Diameter estimation. One problem with RUX samples is that there may be some very special direction in which the polytope has tremendous width. That is, simply looking at ten or twenty RUX samples or even taking standard deviations over thousands of RUX samples, it may appear that the polytope is quite small. Yet, in fact, it may be that there is a very special direction such that if we travel along that direction we find a value of q which is radically different from the typical RUX samples. To assess this, we attempt to show the overall diameter of the polytope is fairly limited, thus convincing ourselves that this is not a problem. In this effort, we obtained Paired Rotationally Uniform eXtremal samples (PRUX): we first choose a direction uniformly at random and then found the two extremal vertices on the polytope in that direction and in the opposite direction. We then computed the distance between those vertices, as well as the distance between those vertices projected to the direction. Doing this procedure many times and making a scatterplot yields Figure 4. This figure suggests that the polytope may be only about 8 units in diameter (although we caution that general results on polytope diameter estimation are not encouraging, cf. [20]). Considering that the magnitude of this diameter would be distributed over 1624 entries of the matrix, each of which lies in $[0, 1]$, it would seem that the polytope may indeed be a fairly small set. Although there is certainly no guarantee that there isn't some very large-magnitude direction hiding within this polytope, it is encouraging that we do not find any extreme outliers in this figure.
- Random uniform samples. The maximum entropy principle provides a quite different point of view: given that we can never distinguish where q may lie inside the polytope $\hat{\Theta}$, we could treat all such values of q as equally likely. In this vein, we apply a Dikin ellipsoid sampler to obtain samples drawn approximately uniformly from $\hat{\Theta}$. We also look at the standard deviation of each entry of q across all these samples, which gives us another way to assess where the polytope has significant ambiguity. The results may be seen in 5.

Together, these approaches give us practical ways of making real use of the results of the Markov Link Method. More generally, they may be applied to many situations where a parameter of interest simply cannot be exactly identified but can be shown to lie near to some set.

3.2 Evaluation

It is difficult to evaluate the accuracy of the Markov Link Method, because ground-truth data is not available in the cases where it actually matters. Nonetheless, we can make a variety of attempts to see how the method fares.

Parametric bootstrap provides one natural way to assess our confidence. That is, we simulate a surrogate dataset according to the estimated distribution (holding the number of samples for each group fixed), apply our method to the resulting surrogate dataset, and then obtain an estimate $\hat{\Theta}^{(i)}$. We do this 100 times. For each estimate, we can then measure whether the original $\hat{q}$ was is close to the estimated polytope, i.e. for each i we can look at the Euclidean norm between $\hat{q}$ and the set $\hat{\Theta}^{(i)}$. Doing so for 100 bootstrap samples and taking a histogram of these distances yield Figure 6. The results suggest our estimator has a

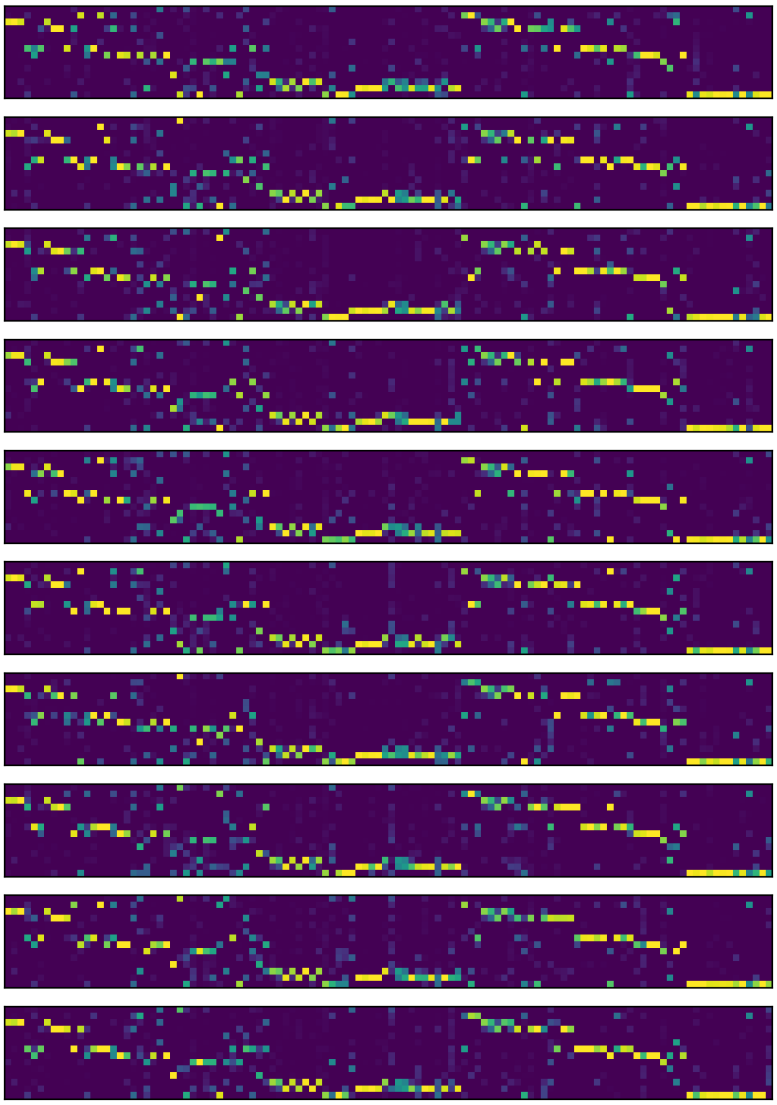

Figure 3: Output: 10 Rotationally Uniform eXtremal samples from the set $\hat{\Theta}$. Each sample shows a different way that technique I might line up with technique II; all of them are consistent with the observed data. Note that these extremal samples from $\hat{\Theta}$ have quite a lot in common. The common features are ones that we can feel confident are present in the true value of the joint distribution, q^*. Other entries have more variability; these suggest specific experiments which might help disambiguate how technique I and technique II are related through q^*.

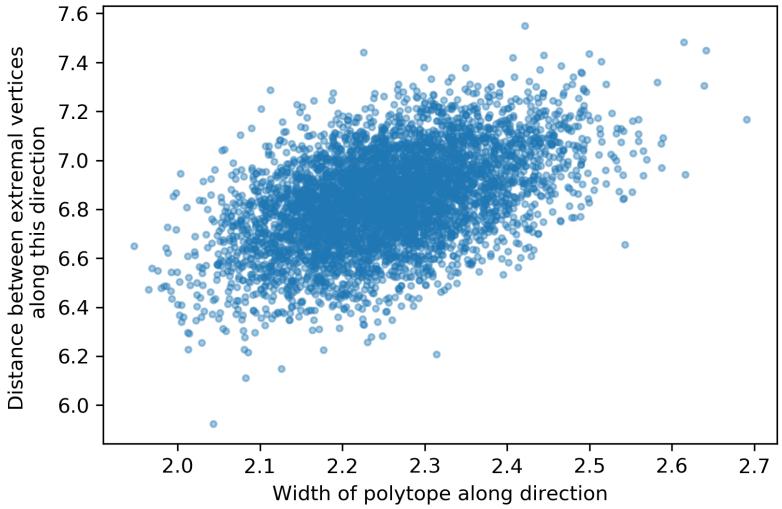

Figure 4: How wide is $\hat{\Theta}$? Each dot above corresponds to a randomly selected direction. The horizontal position of the dot indicates the width of the polytope along that direction, and the vertical position indicates the distance between the two extremal vertices along that direction. It appears that the overall diameter of the set may be less than 8 units.

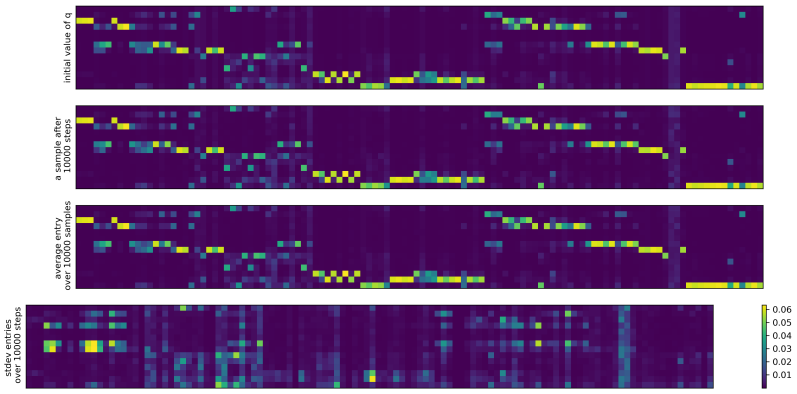

Figure 5: The first row shows the the MLM estimate of $\hat{q}$ (which is not a consistent estimator for q^*). The second row shows a sample drawn approximately uniform from $\hat{\Theta}$, obtained by 10000 iterations of the Dikin sampler. The third row shows the center of mass of $\hat{\Theta}$ and the final row shows the standard deviation of each entry (as estimated by the Dikin samples). Notice that these deviations are never more than 6%, suggesting that the polytope $\hat{\Theta}$ is closely centered around its center of mass.

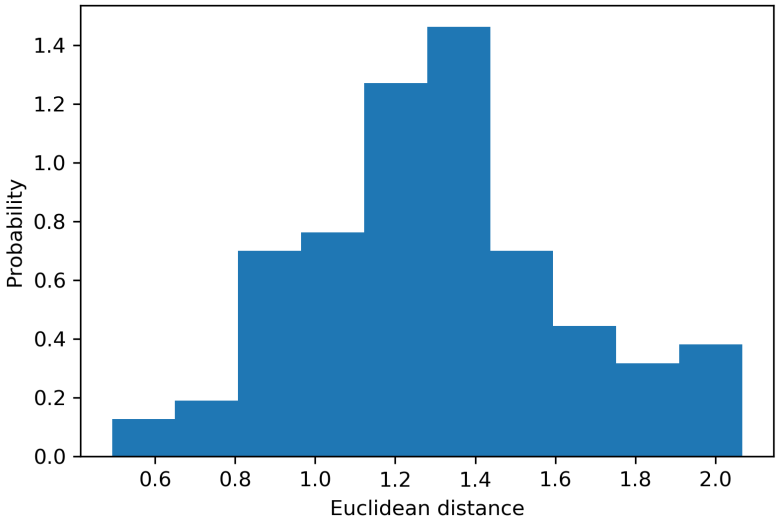

Figure 6: Bootstrap gives us some measure of confidence in the MLM estimate. None of the 100 bootstrap sample polytopes measured more than 2 units from from the original point estimate, $\hat{q}$. That is, for each sample polytope we could find some point in the polytope that was less than 2 units from $\hat{q}$. Spread out over the 1624 entries of the matrix q, each of which lies in $[0, 1]$, we are encouraged insofar as 2 units does not seem to be a very great distance.

fairly tight confidence around the true polytope Θ^* for this dataset. However, we note that theoretical properties of this are certainly lacking.

Nonparametric bootstrap is another approach. In this case, we simply sample the empirical distributions $\hat{p}, \hat{h}$ with replacement to obtain our surrogate datasets, and then apply the MLM method to each surrogate dataset. We evaluate the degree of variability amongst the MLM polytope estimates by looking at the extremal points of those polytopes. That is, we select a random direction and find the extremal vertex for each MLM estimate in that direction. We can then look at position of these extremal points projected along that direction. Looking over all the surrogates and subtracting the mean, we obtain a histogram of deviations. We can then apply this to various directions, which yields the variety of histograms seen in Figure 7.

Finally, we can look directly at whether MLM helps us model the data Y. In theory it seems possible that by pooling data between the modalities we could gain statistical power. This is particularly true in the case of the particular dataset we investigated, wherein obtaining samples for technique II is quite expensive and so there are far fewer available datapoints for the Y data. Thus the empirical distribution of $Y|\ell$ (namely $\hat{h}$) may in fact be a very poor estimator. This suggests that perhaps we might consider the estimator $\tilde{h}(y|\ell) \triangleq \sum_x \hat{p}(x|\ell)\hat{q}(y|x)$ instead, since it allows us to leverage knowledge gained from the plentiful samples of technique I. To test this estimator, we held out 10% of the Y data. We then applied the MLM method to obtain estimates $\hat{p}, \hat{q}$ and use those to produce $\tilde{h}$ as an estimator for the distributions of $Y|\ell$. We then compare $\tilde{h}$ and $\hat{h}$ by calculating the log-likelihood of each estimator on the held-out data. We found an average value of -1.1 per entry for the MLM estimator $\tilde{h}$, but the empirical distribution $\hat{h}$ yields a held-out log likelihood of $-\infty$ because there are examples in the testing set that are simply never seen in the training set. Even if we use pseudocounts to smooth the empirical distribution, we never obtain likelihoods as good as we get using the model which shared data between the two experimental modalities (See Figure 8). Thus we see that sharing data between the

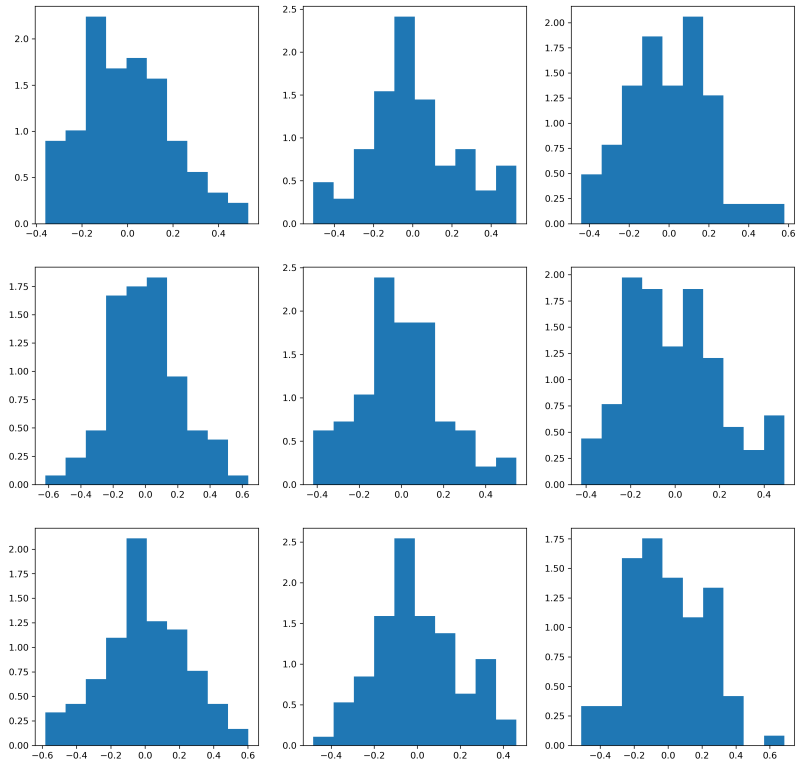

Figure 7: Nonparametric bootstrap gives us another form of confidence. For each of nine randomly selected directions and for each bootstrap sample polytope, we compute how far the polytope extends in that direction. For each direction we can then plot a histogram of how far the various polytopes extend. We see that they do not vary dramatically between samples, which gives us confidence that our estimate is somewhat close to the truth.

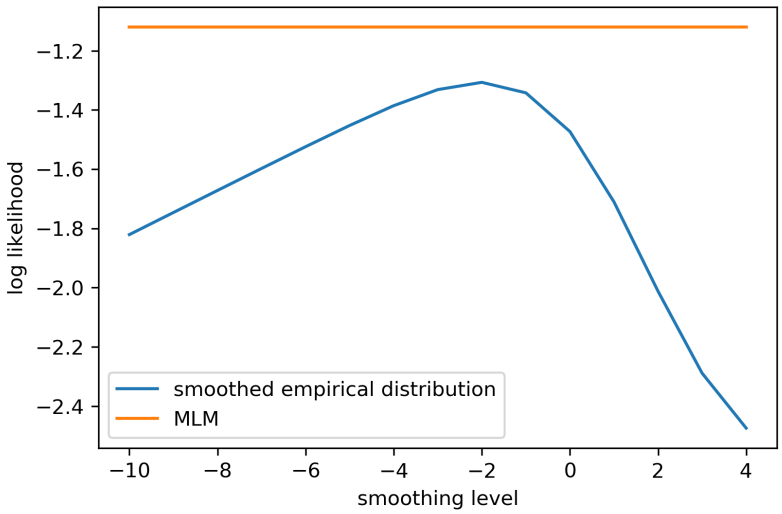

Figure 8: Here we compare two estimators of the distributions of $Y|\ell$. In the orange case, we use the Markov Link Method. In the blue case, we use the empirical distributions of $Y|\ell$, smoothed using pseudocounts. The horizontal axis indicates the size of these pseudocounts in log-scale. The MLM method performs better than the empirical distribution at any level of smoothing.

modalities helps can help us estimate each modality more accurately, as well as enabling us to understand more about the connection between the modalities.

4 Conclusions

When joint measurement is impossible, it can be difficult to calibrate two methods against each other or understand how they may be related. Here we show that a simple Markov assumption can make it possible to actually learn quite a lot. Although the exact relationship may not be identifiable, a polytope of possible relationships can be identified, and this polytope may in fact be quite small indeed. By exploring this polytope, we can understand what we know – and what we don't know – about the relationship between measurements.

The Markov assumption is of course not the only one that we could have used, and may not be valid in every case. For example, it has been speculated that some cell types tend to die more often in one experimental modality than another, and these cells are not part of the data. This violates our assumptions. However, assuming this death rate can be roughly measured, it can be adjusted for, yielding a different but equally meaningful assumption about the data. Moreover, if this method yields bizarre results, it may give useful clues as to exactly how cell death may happen differently in the two modalities.

Once we accept that what we're interested in may not be fully identifiable, any of a wide variety of assumptions can help us obtain practical bounds. Although we may not be able to learn exactly what we want, we can learn a set of possibilities. By probing this set carefully with uniform samplers and extremal tests, we can learn what the data actually has to say and what experiments we need to do to learn more.

References

[1] IEC BiPM, ILAc IFcc, IUPAC ISO, and OIML IUPAP. International vocabulary of metrology–basic and general concepts and associated terms, 2008. *JcGM*, 200:99–12, 2008.

[2] Jeroen De Mast and Albert Trip. Gauge r&r studies for destructive measurements. *Journal of Quality Technology*, 37(1):40, 2005.

[3] Ralph M Steinman and Zanvil A Cohn. Identification of a novel cell type in peripheral lymphoid organs of mice: I. morphology, quantitation, tissue distribution. *Journal of Experimental Medicine*, 137(5):1142–1162, 1973.

[4] Stewart A Bloomfield and Robert F Miller. A physiological and morphological study of the horizontal cell types of the rabbit retina. *Journal of Comparative Neurology*, 208(3):288–303, 1982.

[5] 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.

[6] Marcin Cieślik and Arul M Chinnaiyan. Cancer transcriptome profiling at the juncture of clinical translation. *Nature Reviews Genetics*, 19(2):93, 2018.

[7] Yoshinori Imamura, Toru Mukohara, Yohei Shimono, Yohei Funakoshi, Naoko Chayahara, Masanori Toyoda, Naomi Kiyota, Shintaro Takao, Seishi Kono, Tetsuya Nakatsura, et al. Comparison of 2d-and 3d-culture models as drug-testing platforms in breast cancer. *Oncology reports*, 33(4):1837–1843, 2015.

[8] Benjamin Haibe-Kains, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C Jin, Andrew H Beck, Hugo JWL Aerts, and John Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389, 2013.

[9] Gargi Srivastava and Rajeev Srivastava. A survey on automatic image captioning. In *International Conference on Mathematics and Computing*, pages 74–83. Springer, 2018.

[10] Monya Baker. Reproducibility crisis? *Nature*, 533:26, 2016.

[11] Megan Crow, Anirban Paul, Sara Ballouz, Z Josh Huang, and Jesse Gillis. Characterizing the replicability of cell types defined by single cell rna-sequencing data using metaneighbor. *Nature communications*, 9(1):884, 2018.

[12] W Evan Johnson, Cheng Li, and Ariel Rabinovic. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics*, 8(1):118–127, 2007.

[13] Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 48–55. Morgan Kaufmann Publishers Inc., 2001.

[14] John F Clauser, Michael A Horne, Abner Shimony, and Richard A Holt. Proposed experiment to test local hidden-variable theories. *Physical review letters*, 23(15):880, 1969.

[15] Rafael Chaves, Lukas Luft, Thiago O Maciel, David Gross, Dominik Janzing, and Bernhard Schölkopf. Inferring latent structures via information inequalities. *arXiv preprint arXiv:1407.2256*, 2014.

[16] Aditya Kela, Kai von Prillwitz, Johan Aberg, Rafael Chaves, and David Gross. Semidefinite tests for latent causal structures. *arXiv preprint arXiv:1701.00652*, 2017.

[17] 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.

[18] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[19] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell rna-seq data across data sets. *Nature methods*, 2018.

[20] Andreas Brieden, Peter Gritzmann, Ravi Kannan, Victor Klee, László Lovász, and Miklós Simonovits. Approximation of diameters: Randomization doesn't help. In *Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on*, pages 244–251. IEEE, 1998.

[21] Ravindran Kannan and Hariharan Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012.

A Dikin sampler

Consider a convex polytope $T = \{x : Ax \leq b\}$. We have implemented a method for sampling from this polytope, based on the paper [21]. This method makes use of the Dikin ellipsoids, $E(x)$. For any x, these are defined by

- Computing the distance from x to each facet of the polytope, i.e. $d_i = b_i - \sum_j A_{ij} X_j$.
- Constructing $\tilde{A}$ as $\tilde{A}_{ij} = A_{ij}/d_i$.
- Define $E(x) = \{y : |\tilde{A}(X - y)| \leq 1\}$.

We can use these ellpsoids to efficiently sample the polytope T. At each step, we have some point $X \in T$, and we would to use this point to obtain a new sample Y, such that by iterating this process we asymptotically obtain samples which are uniform in T. Here is how we use X to get Y:

Algorithm 1: Dikin sampler step

Data: A point $X \in T$
Result: A point $Y \in T$

Sample a proposal $\tilde{Y}$, uniformly from $E(X)$;
if $X \in E(\tilde{Y})$ **then**
  Sample $U \sim \text{Uniform}[0, 1]$;
  if $U \leq \text{Vol}(E(X))/\text{Vol}(E(\tilde{Y})) \leq 1$ **then**
    Let $Y \leftarrow \tilde{Y}$;
  else
    Let $Y \leftarrow X$;
  end
else
  Let $Y \leftarrow X$;
end

It is easy to show that the stationary distribution of the Markov chain found by iterating these Dikin sampler steps is indeed uniform on T. To ensure an numerically robust method

in the face of high-dimensional and nearly degenerate matrices, we take the following approach to robustly sampling from the ellipsoid:

Algorithm 2: Ellipsoid sampler

Data: An $n \times m$ matrix $\tilde{A}$
Result: A point X sampled uniformly from $\{x : |Ax| \leq 1\}$

Sample Z as an n-dimensional normal variables vector;
Let X denote the solution to the least squares problem $\min_x |\tilde{A}x - Z|$;
Normalize X by $X \leftarrow X/|\tilde{A}X|$;
Sample $U \sim \text{Uniform}[0,1]$;
Scale X by $X \leftarrow X \times U^{1/m}$;

B Proof of the theorem

For the benefit of the reader, we here repeat the statement of our theorem in more explicit terms.

- Let $|\cdot|_\infty$ denote the uniform norm (i.e. the maximum absolute value) and $|\cdot|$ denote the Euclidean norm (i.e. the square root of the sum of the squares). In the case of matrices, this Euclidean norm goes by the name of the Frobenius norm. Recall that in this norm matrices satisfy a Cauchy-Schwarz like equality, $|pq| \leq |p|\,|q|$. Also recall that $|a|_\infty \leq ENa \leq \sqrt{n}\,|a|_\infty$ where n is the number of entries in a.
- Let $T_{a,b}$ denote the transition matrix polytope, i.e. the set of $a \times b$ matrices whose rows sum to 1 and whose entries are all positive.
- Let $|\Omega_\ell|, |\Omega_X|, |\Omega_Y| \in \mathbb{N}$.
- Let $p^* \in T_{|\Omega_\ell|,|\Omega_X|}$.
- Let $q^* \in T_{|\Omega_X|,|\Omega_Y|}$.
- We require the matrix q^* has strictly positive entries, $q^*_{xy} \geq c > 0$.
- We require that the rows of p^* are linearly independent.
- Let $\hat{p}$ denote an empirical transition matrix drawn by obtaining $N_{X,\ell}$ samples for each row of p^*, i.e. we have samples $(\ell_1, X_1) \cdots (\ell_1, X_n)$ such that $\mathbb{P}(X_i = x) = p^*_{\ell_i,x}$, $N_{X,\ell} = \sum_{i=1}^n \mathbb{I}_{\ell_i=\ell}$, and $\hat{p}_{\ell x} = \sum_{i=1}^n \mathbb{I}_{X_i=x,\ell_i=\ell}/N_{X,\ell}$.
- Let $\hat{h}$ denote an empirical transition matrix drawn by obtainin $N_{Y,\ell}$ samples for each row of $h^* = p^*q^*$.

Now fix any $\kappa > 0$. Let

$$\hat{q} = \arg\max_q \left(\sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log \left(\sum_x \hat{p}(x|\ell) q(y|x) \right) + \kappa \sum_{xy} \log q(y|x) \right) \tag{2}$$

and $\hat{\Theta} = \{q : \hat{p}\hat{q} = \hat{p}q\} \cap T_{|\Omega_X|,|\Omega_Y|}$.

Theorem. *If $N_{X,\ell}, N_{Y,\ell} \to \infty$ in such a way that $N_{Y,\ell'}/\sum_\ell N_{Y,\ell} \geq \rho > 0$ for each ℓ', then $\inf_{q \in \hat{\Theta}} |q^* - q|_\infty \to 0$ in probability.*

Proof. It is well-known that $\hat{p} \to p^*$ in probability (in both the uniform or the Euclidean norm, which are of course equivalent in this case). It is easy to see that the same goes for $\hat{p}\hat{q} \to h^*$ (see Lemma 1). Thus, intuitively, the difficulty is this: by allowing ourselves to

ensure $|\hat{p}-p^*|_\infty, |\hat{p}-p^*|, |\hat{p}\hat{q}-h^*|_\infty, |\hat{p}\hat{q}-h^*|$ sufficiently small, can we find some $\tilde{q}\in\hat{\Theta}$ so that $|\tilde{q}-q^*|_\infty$ is arbitrarily small? It turns out we can.

Recall that $c>0$ is the smallest value of q^*_{xy}. Fix any $\epsilon < c, p^*, q^*$. Let the right inverse of a matrix be defined by $a^\dagger \triangleq a^T(aa^T)^{-1}$. Note that since p^* has linearly independent rows, this is well-defined and continuous in a small neighborhood around p^*. Let $M = \left|(p^*)^\dagger\right|$. Find δ small enough so that if $|p-p^*|_\infty < \delta$ then $\left|p^\dagger\right| < 2M$. Taking a further smaller δ if necessary, ensure that if $|p^*-p|_\infty < \delta$ then $|p^*-p|$ is less than $\epsilon/4M\sqrt{|\Omega_X||\Omega_Y|}$. Now fix any $\hat{p},\hat{q}$ with $|\hat{p}-p^*|_\infty < \delta$ and $|\hat{p}\hat{q}-p^*q^*| < \epsilon/4M$. Take

$$\tilde{q} = q^* + \hat{p}^\dagger\hat{p}(\hat{q}-q^*)$$

Then we make the following observations:

- Let us compute $|\tilde{q}-q^*|$. We have

$$\begin{aligned}|\tilde{q}-q^*| &= \left|\hat{p}^\dagger\hat{p}(\hat{q}-q^*)\right| \le 2M\,|\hat{p}\hat{q}-\hat{p}q^*| \\ &\le 2M\,|\hat{p}\hat{q}-p^*q^*| + 2M\,|(p^*-\hat{p})q^*| \\ &\le 2M\frac{\epsilon}{4M} + \frac{2M\epsilon}{4M\sqrt{|\Omega_X||\Omega_Y|}}\sqrt{|\Omega_X||\Omega_Y|}\,|q^*|_\infty \le \epsilon\end{aligned}$$

- $\hat{p}\tilde{q} = \hat{p}q^* + \hat{p}\hat{q} - \hat{p}q^* = \hat{p}\hat{q}$
- The rows of $\tilde{q}$ sum to 1. This is easy to see, because the rows of q^* sum to 1 and the rows of $\hat{q}$ sum to 1, and so $\tilde{q}\mathbf{1} = q^*\mathbf{1} + \hat{p}^\dagger\hat{p}(\hat{q}-q^*)\mathbf{1} = \mathbf{1}+0$ as desired.
- The entries of $\tilde{q}$ are positive. Indeed, the the smallest value of q^* is c, and we have already argued that $|\tilde{q}-q^*|_\infty \le \epsilon$. Thus the smallest value of $\tilde{q}$ is at least $c-\epsilon$, and we have required $\epsilon < c$.

Thus $|\tilde{q}-q^*|_\infty < \epsilon$ and $\tilde{q}\in\hat{\Theta}$.

In conclusion, we see that by taking $\hat{p}$ sufficiently close to p^* and $\hat{p}\hat{q}$ sufficiently close to p^*q^*, we can ensure that the set $\hat{\Theta}$ contains a close which is arbitrarily close to the true q^*. Since $\hat{p}$ and $\hat{p}\hat{q}$ are themselves consistent estimators, this completes the proof. □

Lemma 1. *If $N_{X,\ell}, N_{Y,\ell} \to \infty$ in such a way that $N_{Y,\ell'}/\sum_\ell N_{Y,\ell} \ge \rho > 0$ for each ℓ', then $|p^*q^*-\hat{p}\hat{q}|_\infty, |p^*q^*-\hat{p}\hat{q}| \to 0$ in probability.*

Proof. Our first task is to make a short study of the continuity of KL divergences on categorical distributions when the probabilities are bounded away from zero. Recall that we have insisted $q^*_{xy} \ge c > 0$ for every x, y – and this also means $(pq^*)_{\ell y} \ge c$ for every ℓ, y, since each row of p is itself a probability distribution. Moreover, observe that the KL-divergence on $|\Omega_Y|$-dimensional distributions, $D(\hat{r}||\tilde{r}) \triangleq \sum_y \hat{r}_y \log \hat{r}_y/\tilde{r}_y$, is *uniformly* continuous on the space of such distributions whose minimum probability is greater than any fixed positive constant. It follows that the map $h,p,q \mapsto D(h_\ell||(pq)_\ell)$ is also uniformly continuous on a space where h and q are strictly greater than some fixed positive constant.

With this in hand, the remainder of the proof follows naturally, using the well-known results that empirical distributions are consistent, i.e. $\hat{p}\to p^*$ and $\hat{h}\to p^*q^*$ in probability.

Fix any ϵ,π. Let δ the modulus of continuity in the norm $|\cdot|_\infty$ at level $\epsilon\rho$ for the map $h,p,q \mapsto D(h_\ell||(pq)_\ell)$ restricted to the domain where $h,q > c/2$. Select N large enough so that $\frac{1}{N_{Y,\ell}}\kappa|\Omega_X||\Omega_Y|\log\frac{1}{c} < \epsilon$ for each ℓ and so that with probability at least π we have that $\hat{h},\hat{p}$ so that $\left|\hat{h}-p^*q^*\right|_\infty, |\hat{p}-p^*|_\infty \le \delta, \left|\hat{h}-p^*q^*\right|_\infty < c/2$. Then, with probability π, we

must have

$$|D\left(\hat{h}_\ell||(\hat{p}\hat{q})_\ell\right) - D\left(h^*_\ell||(\hat{p}\hat{q})_\ell\right)| \leq \rho\epsilon$$
$$D\left(\hat{h}_\ell||(\hat{p}q^*)_\ell\right) = |D\left(\hat{h}_\ell||(\hat{p}q^*)_\ell\right) - D\left((p^*q^*)_\ell||(p^*q^*)_\ell\right)| \leq \rho\epsilon$$

Now, since $\hat{q}$ is defined as the maximizer of a certain quantity (Equation 2), we may be sure that it is greater than the same quantity evaluated at $q = q^*$. That is,

$$\begin{aligned} 0 &\leq \sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log \frac{\sum_x \hat{p}(x|\ell)\hat{q}(y|x)}{\sum_x \hat{p}(x|\ell)q^*(y|x)} + \kappa \sum_{xy} \log \frac{\hat{q}(y|x)}{q^*(y|x)} \\ &= \sum_\ell N_{Y,\ell} \left(D\left(\hat{h}_\ell||(\hat{p}q^*)_\ell\right) - D\left(\hat{h}_\ell||(\hat{p}\hat{q})_\ell\right)\right) + \kappa \sum_{xy} \log \frac{\hat{q}(y|x)}{q^*(y|x))} \end{aligned}$$

Applying our continuity results, it follows that

$$\sum_\ell N_{Y,\ell} \left(D\left(\hat{h}^*_\ell||(\hat{p}\hat{q})_\ell\right)\right) \leq 2\left(\sum_\ell N_{Y,\ell}\right)\epsilon + \kappa|\Omega_X||\Omega_Y| \log \frac{1}{c}$$

Note that the left-hand summands are all positive. So, in particular, it follows that for each ℓ', applying the uniformity condition ρ, we have

$$D\left(\hat{h}^*_{\ell'}||(\hat{p}\hat{q})_{\ell'}\right) \leq 2\epsilon + \frac{1}{N_{Y,\ell'}} \kappa|\Omega_X||\Omega_Y| \log \frac{1}{c} \leq 3\epsilon$$

That is, we have shown convergence of probability in the KL sense: for any ϵ, π we can find N high enough so that $D\left(\hat{h}^*_{\ell'}||(\hat{p}\hat{q})_{\ell'}\right) < \epsilon$ with at least probability π. This, in turn yields convergence in probability in the Euclidean or uniform metrics by Pinsker's inequality. □