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

Jackson Loper

May 15, 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. We will denote it by $q^*(y|x)$. As input, this conditional distribution takes the measurement results from one tool on a particular specimen, $x$. As output, it yields the probability of obtaining result $y$ from a second tool applied to measure 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.

# 1 The Markov Link Method assumption

To make these ideas rigorous, let us develop a little bit of notation. Let us say we are considering the members of a large population. For example, each "specimen" in the large population might be a human cell or a piece of metal which needs to be tested. We will assume there are three basic properties of interest for each specimen $i$:

1. $\ell_i$, the subpopulation or side information. We assume that the overall population can be split into subpopulations of interest. For example, we could define subpopulations by looking at cells in different parts of the body, or cells with different sizes.
2. $X_i$, the result of measuring a specimen with tool I. For example, perhaps tool I takes a picture of the cell with a destructive electron microscopy method.
3. $Y_i$, the result of measuinrg a specimen with tool II. For example, perhaps tool II measures the RNA expression of the cell with a destructive sequencing method.

We here consider the case that "joint measurement" is impossible or impractical. In terms of the notation above, that means that for any given specimen we can either observe $\ell_i, X_i$ or $\ell_i, Y_i$. We can never observe $\ell_i, X_i, Y_i$ for any specimen $i$. Despite this obstacle, we would like to estimate the conditional distribution $q^*(y|x) \triangleq \mathbb{P}(Y_i = y | X_i = x)$. Under suitable assumptions, we will see that such estimation is indeed possible using the side information $\ell$. However, this estimation suffers from so-called *identifiability problems*, which we will describe in detail in the next section. Fortunately, this identifiability error can be quantified, allowing us to determine exactly what our data tells us about the true calibration. If the identifiability error is large, the Markov Link method suggests future experiments based on new subpopulations which will hone down on the identifiability problem.

The key assumption of the Markov Link Method is a conditional independence assumption: that $\mathbb{P}(Y|X, \ell_i)$ is independent of $\ell$, i.e. $\mathbb{P}(Y = y | X = x, \ell) = q^*(y|x)$ for every value of $\ell$. Intuitively, this signifies that the manner in which $X$ and $Y$ are related is the same for each subpopulation. If this assumption is not met, then the method presented here is not applicable. The validity of these assumptions for a given situation should be closely contemplated.

Let us consider a few real-world examples where this assumption may apply.

- 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 two 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. In this case, $\ell$ might indicate the type of a part being tested, $X$ would indicate the reliability of a part as measured by one machine, and $Y$ would indicate the reliability of a part as measured by another machine. If the error in machine $Y$ is not correlated to the part type $\ell$, then the MLM assumption certainly holds. Even if the error is correlated, the MLM assumption may still hold. For example, imagine that the $Y$ error is correlated with the absolute reliability of the part, this may pose no problem because that reliability is measured by $X$. As long as the correlations are the result of something which is measured in $X$, the conditional independence assumption will hold.
- Combining knowledge across experimental modalities: morphology and transcriptomics. There are different ways to think about the different types of cells in an

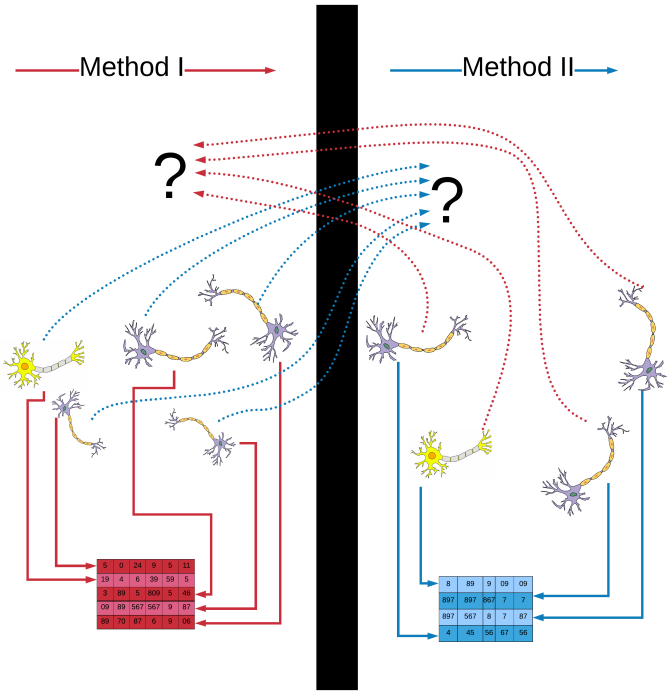


**Figure 1:** We consider the case that for each specimen there are two ways we might measure it. We divide the specimens into two groups. The specimens on the left are measured with technique I and the specimens on the right are measured with technique II. In the end, we are left wondering: what would have happened if we had measured the specimens on the left with technique II? Or what if we had measured the specimens on the right with technique I? The Markov Link Method gives a way to answer this problem using a subpopultaion analysis. For each specimen, we take $\ell$ to indicate the subpopulation (indicated here by the color of the neuron, i.e. yellow or purple), $X$ to indicate the results we would obtain from that specimen if we measured it with method I, and $Y$ to indicate the results we would obtain from that specimen if we meausured it with method II.

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. In this case, $\ell$ might indicate some side information such as where in the body the cell was found, $X$ would indicate the classification of the cell according to its transcriptomics, and $Y$ would indicate the classification of a cell according to its morphology. We expect that cell morphology is largely a function of cell transcriptomics. Thus, as long as the $X$ measurement is sufficiently detailed, we expect that any correlations between $Y$ and $\ell$ would be explained by $X$. That is, the MLM assumption holds.

- 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. Here $\ell$ might indicate cancer location, $X$ might indicate transcriptomic expression of cultured cells, and $Y$ might indicate transcriptomic expression of in-vivo cells. As the transcriptomic expression is much more informative than the cancer location, it is plausible that $X$ might be sufficient to explain any correlations between $\ell$ and $Y$. Thus the MLM assumption may hold.

- 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. Here $\ell$ would indicate coarse labels such as "cat" or "street scene." These labels could be derived from either images or text and can be trained in a supervised fashion. Then, $X$ would indicate the image and $Y$ would indicate a caption. Since the caption should be determined by the picture $X$, the MLM assumption may hold.

- 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 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. In this case, $\ell$ would indicate subpopulations which both labs could access. For example, we can take several batches of mice; for each batch we can send half to one lab and half to the other lab. $X$ will the indicate the full results from each specimen examined in one lab and $Y$ coarser information from specimens examined in the other. If the $X$ data is sufficiently detailed, the MLM assumption may hold.

The main contributions of this paper are summarized here:

- Posing of the Markov Link Method assumption as an assumption that may make it possible to perform calibrations without joint measurement.
- An analysis of some identifiability problems posed by the Markov Link Method assumption. Although it turns out precise identifiability is often impossible, we formulate an asymptotic theory for bounding the magnitude of the this problem. In particular, although we often cannot hope to exactly determine the calibration $q^*(y|x) \triangleq \mathbb{P}(Y = y|X = x)$, we can estimate a so-called polytope $\hat{\Theta}$ which captures our uncertainty about $q^*$ due to identifiability problems.
- The result of the identifiability analysis is a polytope $\hat{\Theta}$. We investigate various ways of measuring the extent of this polytope. We see how it is possible to understand what this "extent" means in terms of the identifiability problem. When the extent is small, we see that the identifiability problem is actually quite minor.
- We apply the MLM to two single-cell transcriptomic datasets. The result uncovers an important and unexpected property of one of the experimental methods.

Our identifiability analysis stands on the shoulders of a long history of relating probabilistic assumptions to probabilistic inequalities on unidentifiable parameters. Indeed, the core idea of this work is to take an assumption (the Markov Link assumption) and use it to place inequality bounds on an unidentifiable quantity (namely $q^*$). When inequalities are fairly tight on all sides, we see that much can be learned even when what we want isn't identifiable. Much of the prior literature in this kind of direction 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.

The remainder of this work proceeds as follows. First we give a careful consideration to the Markov Link Method assumption and how it allows us to use side information to estimate the calibration. We also see when this side information will and will not be helpful. Next, we outline the Markov Link Method and describe a new asymptotic consistency result concerning the identifiability issues. Finally, we consider a real dataset. We perform the method and then attempt to criticize the result three ways: through concerns about not having enough data, through concerns about modeling assumptions, and through concerns about identifiability. The method largely prevails against these criticisms, and reveals an unexpected feature of the two techniques.

## 2 Precise problem formulation and discussion of the identifiability problem

Let us now be completely rigorous. We shall make three assumptions:

1. For each specimen $i$, the distribution of $X_i, Y_i|\ell_i$ may be written

$$\mathbb{P}(X_i = x, Y_i = y|\ell_i) = p^*(x|\ell_i)q^*(y|x)$$

   This is the central assumption we have discussed at length above. For convenience we will also introduce the notation

$$\mathbb{P}(Y = y|\ell) = h^*(y|\ell) = \sum_x p^*(x|\ell)q^*(y|x)$$

2. Joint measurement is unavailable. In particular, we will assume we have $n+m$ individual specimens. Of these, we have observed $\ell_i, X_i$ for $i \in \{1 \cdots n\}$ and $\ell_i, Y_i$ for $i \in n+1 \cdots n+m$.

3. $X$ and $Y$ are discrete random variables with finite support (the general concepts here will apply more generally, but we leave it for future work). If the data is not discrete, we can always make it so by dividing it into suitable bins. In this simple case, we may summarize all of this information with two matrices:

$$N^X_{\ell x} = |\{i \le n : \ \ell_i = \ell, X_i = x\}| \qquad N^Y_{\ell y} = |\{i > n : \ \ell_i = \ell, Y_i = x\}|$$

   Thus $N^X, N^Y$ are matrices counting the number of each kind of observation for method I and method II respectively.

Our goal will be to use $N^X, N^Y$ to estimate $q^*$, the calibration. However, *the data from the matrices $N^X, N^Y$ only enable us to estimate $p^*, h^*$.* They do not enable us to directly estimate $q^*$. Therefore, we define

$$\Theta(p,h) \triangleq \left\{ q : \ \sum_x p(x|\ell)q(y|x) = h(y|\ell) \ \forall \ell, y, \ q(y|x) \ge 0 \ \forall x, y, \sum_y q_{xy} = 1 \ \forall x \right\}$$

as the set of values of the calibration $q$ which are consistent with a given value of $p$ (the conditional distribution of $X|\ell$) and $h$ (the conditional distribution of $Y|\ell$). Any effort to estimate $q^*$ must therefore overcome two fundamentally different challenges:

**Not-enough-data problems** We don't have infinite data, so we can't hope to exactly determine $p^*, h^*$.

**Identifiability problems** Even if we knew $p^*, h^*$ exactly, it is often impossible to know the value of $q^*$. We can only ever know that it lies somewhere in $\Theta(p^*, h^*)$.

Let us now take a moment to understand the set $\Theta$. This set tells us how the things we can estimate (i.e. $p^*, h^*$) inform us about what we want to know (i.e. $q^*$). To understand a bit more concretely how this works, it may be helpful to think of $p^*, h^*, q^*$ as matrices. From this point of view, one aspect of the definition of $\Theta$ can be written more concisely as a matrix equality constraint on $q^*$: if $q \in \Theta(p^*, h^*)$, then $p^*q = h^*$. The consequences of this equation depend upon whether $p^*$ has a left-pseudoinverse.

- If $p^*$ has a left-pseudoinverse then this equation allows us to uniquely determine what we want ($q^*$) in terms of what we know ($p^*, h^*$). In general, this will happen when the number of subpopulations outnumbers the number of different states that $X$ can take on. In this case, the problem should be straightforward to solve.

- If $p^*$ does not have a left-pseudoinverse, then there is an identifiability issue: we can never hope to exactly determine $q^*$.

In this paper we will focus almost entirely on the second case. In this second case there is a genuine identifiability issue; it is *impossible* to ever determine the true value of $q^*$, regardless of how much data we have. It therefore becomes of paramount importance to understand the magnitude of this identifiability problem. In particular, even though we cannot know $q^*$ exactly, can we place bounds on what $q^*$ could be?

To gain intuition, let us consider a few examples.

**Example 1.** Let $\ell \in \{1,2\}$, $X \in \{1,2,3\}$, $Y \in \{1,2\}$, and

$$p^* = \begin{pmatrix} 40\% & 50\% & 10\% \\ 10\% & 10\% & 80\% \end{pmatrix} \qquad h^* = \begin{pmatrix} 20\% & 80\% \\ 40\% & 60\% \end{pmatrix}$$

That is, for example, $p^*_{1,2} = \mathbb{P}(X = 2|\ell = 1) = 50\%$ and $h^*_{2,1}\mathbb{P}(Y = 1|\ell = 2) = 40\%$. Let us look at a single equation in the system $p^*q^* = h^*$ entailed by the MLM assumption. For example, $h^*_{2,1}$, we get

$$0.4 = h^*_{2,1} = \sum_x p^*_{2,x} q^*_{x,1} = .1q^*_{1,1} + .1q^*_{2,1} + .8q^*_{3,1}$$

This immediately tells us something about $q^*$. For example we now know that $q^*_{1,1} = 10(.4 - .1q^*_{2,1} - 0.8q^*_{3,1})$. But it actually tells us much more than that. Observe that $0 \le q^*_{xy} \le 1$ for every $x, y$; in other words, probabilities must be positive by definition. Thus, in particular,

$$0.4 = .1q^*_{1,1} + .1q^*_{2,1} + .8q^*_{3,1} \le .2 + .8q^*_{3,1}$$

It follows immediately that $q^*_{3,1} \ge .25$. Each equation in the MLM system yields insights of this kind; together these insights form the constraints that define $\Theta$. The more different subpopulations you have, the more equations of this kind you will have.

**Example 2.** Let $\ell \in \{1\}$, $X \in \{1,2,3,\cdots 11\}$, $Y \in \{1,2\}$, and

$$p^* = \begin{pmatrix} 90\% & 1\% & 1\% & \cdots & 1\% \end{pmatrix} \qquad h^* = \begin{pmatrix} 95\% & 5\% \end{pmatrix}$$

Here we have only *one* population. Still we are able to say something quite significant; using similar reasoning to that shown above we see that $q^*_{1,1} \ge 94.\bar{4}\%$.

**Example 3.** Finally, let us consider one class of examples in which the entire situation can be visualized: $\ell = 1$, $X \in \{1,2\}$, $Y \in \{1,2\}$. Let $\theta^*_1 = q_{*1,1}$ and $\theta^*_2 = q_{*2,1}$. Notice that in this simple case $\mathbb{P}(Y = 2|\ell) = 1 - \mathbb{P}(Y = 1|\ell)$, so the two values $\theta_{*1,\theta^*_2}$ completely define $q^*$. The matrix equation $p^*q^* = h^*$ corresponds to a linear constraint on $\theta^*$. On the other hand, the fact that $q^*$ must be a valid probability distribution tells us that $\theta^*_1, \theta^*_2$ must lie inside the box $[0,1] \times [0,1]$. Combining the matrix equation with the constraints of the box, we learn that $q^*$ lies inside a certain bounded, one-dimensional space. This space is precisely $\Theta$. The length of this line segment depends on the exact values of $p^*, h^*$. In some cases it is quite large, i.e. it will be difficult to know much about the true value of $q^*$. However, in other cases the positivity restrictions force $\Theta$ to be a very small region indeed. See Figure 2 to visualize this. In general, this $\Theta$ is much higher-dimensional and harder to visualize, but in this simple case we can see it clearly.

# 3 The Markov Link Method

Let $N^X, N^Y$ denote matrices describing the original observed data. The purpose of the Markov Link Method is to try to use these matrices to estimate $q^*(y|x) = \mathbb{P}(Y = y|X = x)$, the calibration between $X$ and $Y$. The Markov Link Method proceeds in two main steps:

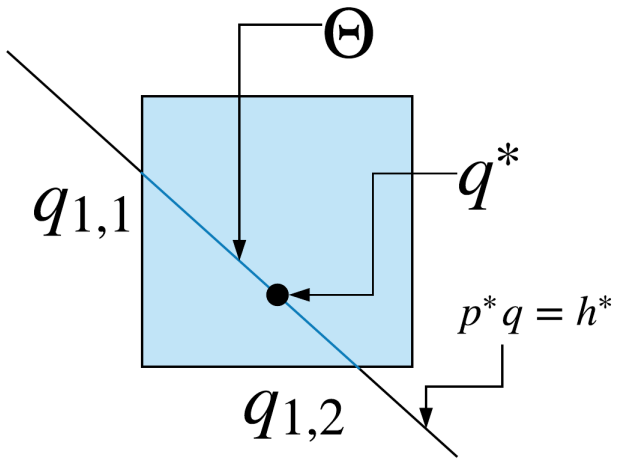


**Figure 2:** We consider the simple case $\ell = 1$, $X \in \{1,2\}$, $Y \in \{1,2\}$. Then the space of values of $q$ which are valid probability distributions can be understood as a box. The space of values of $q$ which satisfy the matrix equation $p^*q = h^*$ can be understood as a line that passes through that box. The set of possible values of $q$ which are consistent with the matrix equation and are also valid probability distributions forms the set $\Theta$, a line segment. We know that the true value of $q^*$ must lie somewhere inside this $\Theta$. Depending upon the exact values of $p^*, h^*$, this line segment may be larger or smaller. For example, if the line crosses close to a corner of the box, then $\Theta$ may be restricted to a very small region near that corner.

**Estimation** Estimate $p^*$ with robust pseudocount estimator:

$$\hat{p}(x|\ell) \triangleq \frac{1 + N^X_{\ell x}}{\sum_{x'} N^X_{\ell x'}}$$

and then calculate a possible guess for $q^*$, namely

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

The entropic regularizer $\kappa$ makes it possible to obtain a unique estimate for $q^*$, despite the identifiability problem discussed above. In practice, we took $\kappa$ to be any small constant less than 1. The optimization problem which defines $\hat{q}$ is convex and easily solved.

Note that although this gives us a unique estimate for $q^*$, the identifiability problems have not magically vanished. There is good reason to suppose that $\hat{q}$ is near to the set $\Theta(p^*, h^*)$, but if the set $\Theta$ is very large then it may be that $\hat{q}$ is very different from the true $q^*$. Therefore, there is a second step we take: a criticism step. This step should be taken for any model, but one which is particularly important in this case.

**Criticism** Level criticisms at the learned model to see whether $\hat{q}$ can be trusted.

1. Use bootstrap to determine if we have enough data.
2. Use a held-out log-likelihood method to make basic sanity checks as to whether the MLM assumption fits with the data.

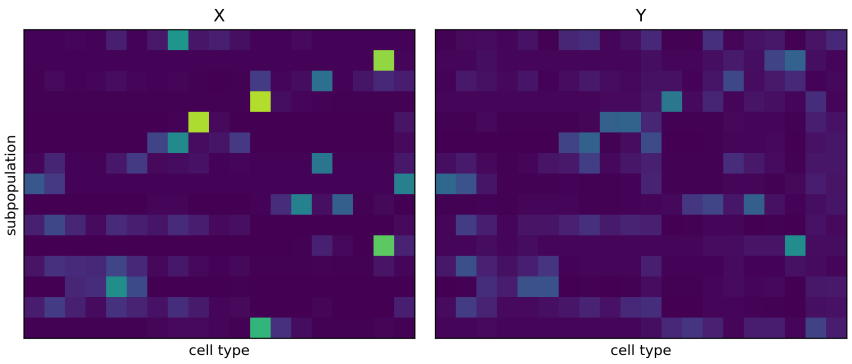


**Figure 3:** Input: two tables. The Allen Institute had access to various cre/lox cell selection techniques. Each technique pulls out a different group of cells. Once the cells were selected, they were then either subjected to technique I technique II ('patch'). Technique I gives a very complete analysis of the gene expression of the cell. Technique II gives a less complete analysis, but yields additional electrophysiological data that may be of interest. For each subpopulation and each technique, data was gathered to estimate the distribution of different cell-types. These are shown above. Each column corresponds to a different cell-type. Each row corresponds to a different subpopulation. Each entry gives the proportion of a particular cell-type found within a particular population, according to a particular technique. If the two methods were perfectly calibrated, we would expect the two tables to look identical.

3. Measure the extent of $\Theta(\hat{p}, \hat{p}\hat{q})$ to see whether we have significant identifiability issues.

The very final point merits a little bit more discussion. In the previous section, we saw that the identifiability issues could be summed up as follows: we can't generally tell what $q^*$ is, only that it lies somewhere inside the set $\Theta(p^*, h^*)$. The extent of the identifiability problems can therefore be understood in terms of the extent of this set. However, in the method above, we look at the extent of a different set: $\Theta(\hat{p}, \hat{p}\hat{q})$. This set is based on mere estimates: $p^* \approx \hat{p}$ and $h^* \approx \hat{p}\hat{q}$. It is not immediately obvious that the size of this *estimated* set would give us an indication of the size of the *true* set of interest, $\Theta(p^*, h^*)$. Fortunately, it is straightforward to show that under mild conditions this is indeed the case. This is the content of the main analytical theorem of this paper, for which a precise statement may be found in Appendix B. It is in part due to the structure of this proof that we choose to estimate $h^*$ with $\hat{p}\hat{q}$ instead of a more traditional estimator.

# 4 Empirical results

## 4.1 Background

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"). Each modality would take a cell and determine what "type" of cell it was. However, as part of that process it would destroy the cell.

Best efforts were made to use biological intuition to calibrate the methods. For example, if a cell was designated as "Lamp5 Egln3_1" celltype using one method, the hope was that it would also be given the same designation if it was processed using the other method. The two methods were designed by scientists to achieve this goal. However, each method has its own biases and errors, and it was not obvious whether this effort was successful.

In particular, it seemed clear that in some cases cells labelled one way with one method would get labelled another way with another method, but it was not clear how often this occurred.

Fortunately, there was a kind of information that seemed like it might help determine whether the two methods were properly calibrated: sub-populations. Using a cre/lox system (cf. [5]) they were able to pick out specific, overlapping subpopulations of neurons. Each subpopulation was expected to contain different proportions of the different cell-types. For each subpopulation and each method, many specimens were sampled and their cell-types determined. If the methods were perfectly calibrated, we would expect that both methods would yield the same distribution of cell-types in each subpopulation.

Towards this effort, data was collected for each subpopulation and each method. The result of this process was two tables, shown in Figure 3. Perhaps not surprisingly, it was found that the distribution of cell-types appeared different under the two modalities. In fact, in one of the methods some of the cells were designated as "unknown," so even the set of cell-types was not the same between the two groups. Clearly the methods were not perfectly calibrated – but how big of a problem was it? It is not obvious to know just by staring at Figure 3. A more quantitative method was needed.

## 4.2 Estimation

The first step of the Markov Link Method was simply to estimate the calibration $q^*$. Taking $\kappa = .1$, we obtained a calibration found in Figure **??**. While this result seemed to generally suggest the methods were well-calibrated, there are some striking divergences. For example, this calibration suggests a that celltypes designated as celltype "10" in one method are actually quite likely to be designated as celltype "7" in the other technique. Both celltypes are associated with similar neuronal types with high expression of somatostatin. However, type "10" is associated with genes such as 'C1ql3,' 'Chodl,' and 'Nts,' whereas type "7" is associated with 'Etv1' and 'Myh8.'

Notice that there are several fundamentally different reasons we could see these kinds of miscalibrations:

- The methods are actually miscalibrated.
- The subpopulations were unevenly sampled. It may be that certain techniques tend to cause cell-death in certain cell-types more than others. These cells are then discarded from the data. If this process is different between the two techniques, it could have an important impact on the calibration estimated here.
- The calibration was poorly estimated.

It is to the third point we now turn.

## 4.3 Criticism

TODO:

1. Nonparametric bootstrap (avg distance 0.66, Figure 4)
2. Parametric bootstrap (avg distance .74, Figure 5)
3. Held out likelihood (2.21 nats vs our method 2.14 nats)
4. RUX samples (avg distance .16, Figure 6)
5. Uniform samples (avg distance .02, Figure 7)

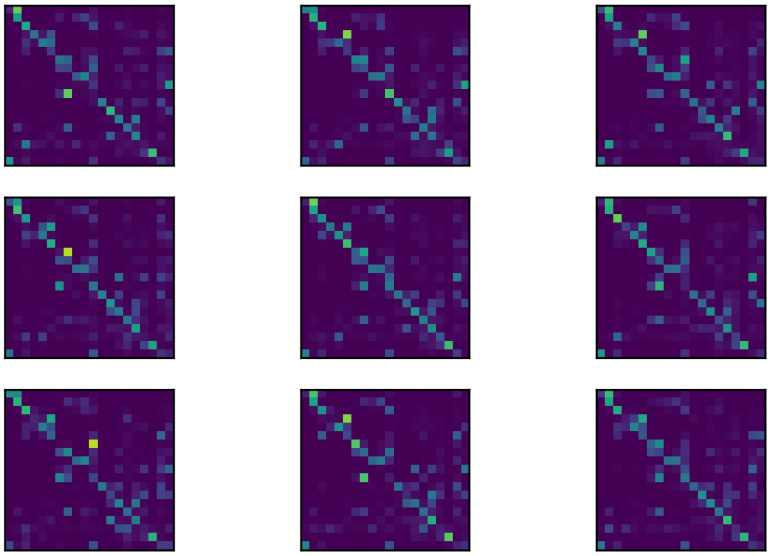

**Figure 4:** Nonparametric bootstrap.

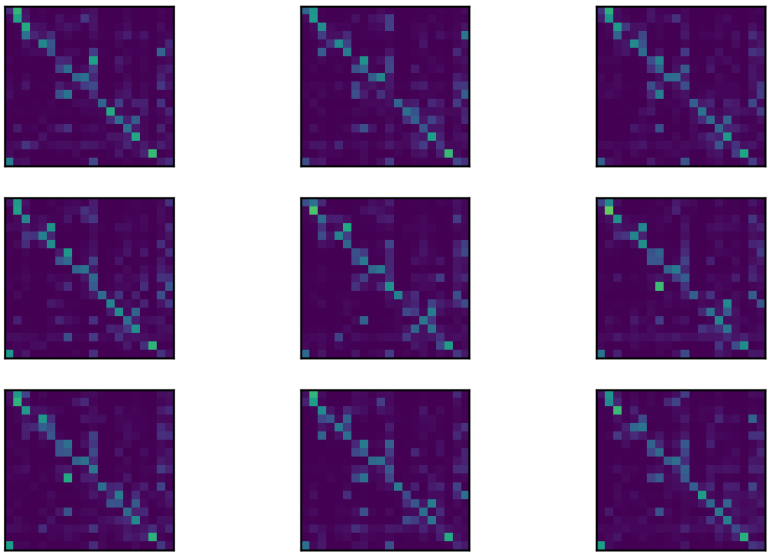

**Figure 5:** Parametric bootstrap.

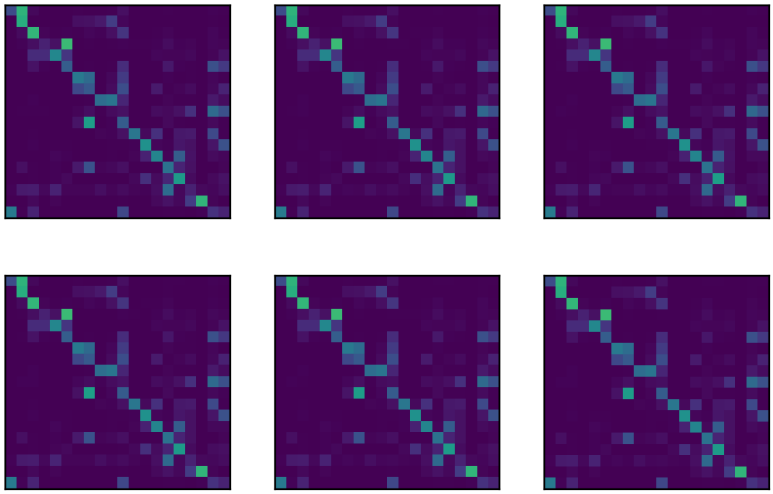

**Figure 6:** RUX samples.

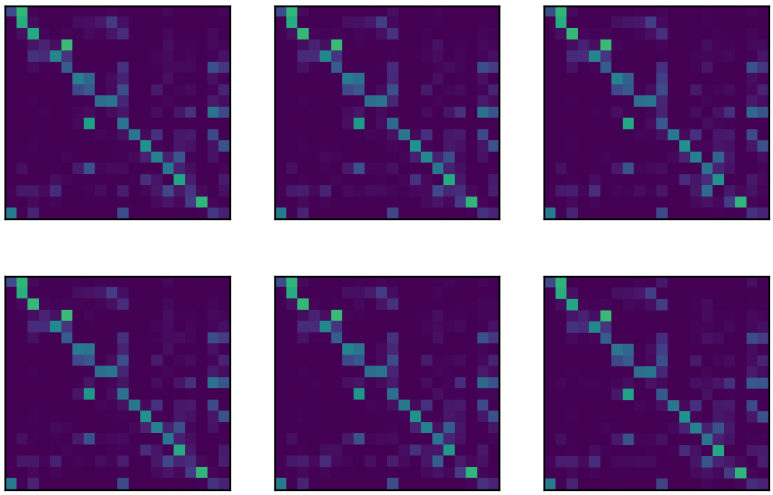

**Figure 7:** Uniform samples.

# 5 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] 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 [18]. 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

$$\hat{q} = \arg\max_q \left( \sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log \left( \sum_x \hat{p}(x|\ell) q(y|x) \right) + \kappa \sum_{xy} \log q(y|x) \right) \tag{1}$$

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 $\ell'$, 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 $a^\dagger \triangleq a^T (a a^T)^{-1}$. 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 $\delta$ small enough so that if $|p - p^*|_\infty < \delta$ then $\left| p^\dagger \right| < 2M$. Taking a further smaller $\delta$ 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

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

- $\hat{p}\tilde{q} = \hat{p} q^* + \hat{p}\hat{q} - \hat{p} q^* = \hat{p}\hat{q}$
- 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 $|\tilde{q} - q^*|_\infty \leq \epsilon$. 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 $N_{Y,\ell'} / \sum_\ell N_{Y,\ell} \geq \rho > 0$ for each $\ell'$, 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 $q^*_{xy} \geq c > 0$ for every $x, y$ – and this also means $(pq^*)_{\ell y} \geq c$ for every $\ell, y$, 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 $\epsilon, \pi$. Let $\delta$ 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 $\ell$ and so that with probability at least $\pi$ we have that $\hat{h}, \hat{p}$ so that $\left|\hat{h} - p^*q^*\right|_\infty, |\hat{p} - p^*|_\infty \leq \delta, \left|\hat{h} - p^*q^*\right|_\infty < c/2$. Then, with probability $\pi$, we must have

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

Now, since $\hat{q}$ is defined as the maximizer of a certain quantity (Equation 1), 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 $\ell'$, applying the uniformity condition $\rho$, 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 $\epsilon, \pi$ we can find $N$ high enough so that $D\left(\hat{h}^*_{\ell'}||(\hat{p}\hat{q})_{\ell'}\right) < \epsilon$ with at least probability $\pi$. This, in turn yields convergence in probability in the Euclidean or uniform metrics by Pinsker's inequality. □