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

Jackson Loper, Osnat Penn, Trygve Bakken, David Blei, Liam Paninski

June 4, 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 a calibration between the tools.

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." 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 [1]. 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 result x obtained from one tool on a particular specimen. 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 significantly 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 Relation to prior work

Our main goal is to achieve calibration without joint measurement. The main obstacle is what is known in the statistics world as an 'identifiability problem.' Due to this problem, we simply *cannot* directly estimate the calibration that we are interested in. The calibration is simply not "identifiable." Nonetheless, all is not lost. Our main contribution is to show that one can bound the extent of this identifiability problem. In short, the fundamental problem cannot generally be vanquished, but we can put it in its place.

Our identifiability analysis stands on the shoulders of a long history of turning probabilistic assumptions into bounds on unidentifiable parameters. The core idea of the MLM is to take an assumption (the 'Markov Link assumption,' which we will define shortly) and use it to place bounds on an unidentifiable quantity (namely the calibration q^*). When these bounds are fairly tight on all sides, we see that much can be learned even though what we want isn't identifiable. Much of the prior literature in this kind of direction comes from research into causality. For example, in [2] 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 [3]. Indeed the physics literature has contributed many key inequalities (cf. [4], [5], and the references therein). Perhaps the closest work to this one would be [6], 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.

It is perhaps worth mentioning that the actual theorem presented here has probably been derived before (although we did not find it in our review of the literature). Other results in the same flavor as the one presented here might also find a practical use for modern problems. There is a treasure-trove of ideas in the causality literature; this wealth has not yet been brought fully to bear on the challenging and important problems of calibration for modern experimental modalities. Our primary intention with this article is to bring attention to the practical utility of these kinds of result.

2 The Markov Link Method assumption

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

1. ℓ_i, the subpopulation. 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 ℓ_i, X_i or ℓ_i, Y_i. We can never observe ℓ_i, X_i, Y_i for any specimen i. Despite this obstacle, we would

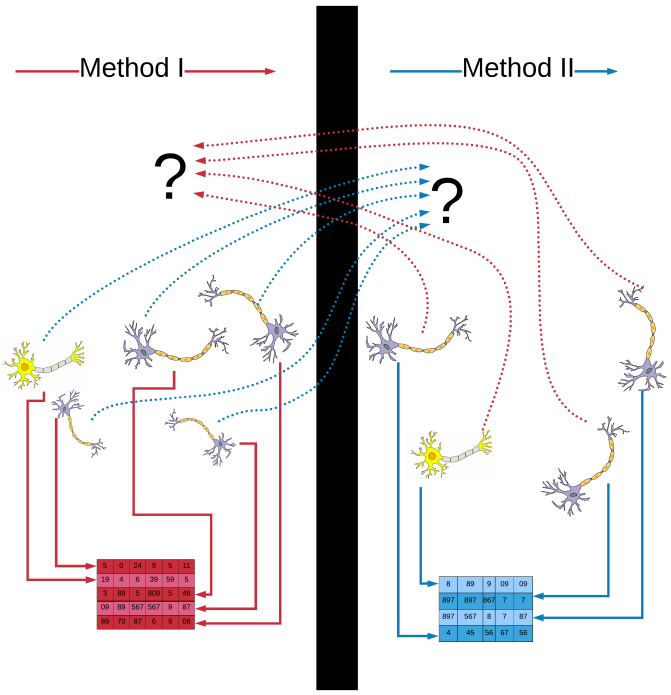

Figure 1: We consider the case that for each specimen there are two ways we might measure it. We divide the specimens into two groups. The specimens on the left are measured with technique I and the specimens on the right are measured with technique II. In the end, we are left wondering: what would have 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 subpopulation analysis.

like to estimate the conditional distribution $q^*(y|x) \triangleq \mathbb{P}(Y_i = y|X_i = x)$. To make this estimation at all possible, we need some kind of assumption.

The key assumption of the Markov Link Method is a conditional independence: that $\mathbb{P}(Y|X,\ell)$ is independent of ℓ, i.e.

The Markov Link Method Assumption
$\mathbb{P}(Y = y|X = x,\ell) = q^*(y|x)$ for every value of ℓ.

Intuitively, this signifies that the manner in which X predicts Y would be 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. [7]). 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, ℓ 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 ℓ, 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 if that reliability is adequately measured by X.

- 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. [8, 9]). A more modern approach is to assay the cell's transcriptome (cf. [10]). Unfortunately, modern high-resolution cell photography and single-cell sequencing technologies are both destructive. As a result, we can't always get both kinds of data for the same specimens. For cells native to regions full of diverse cell-types, it is thus quite hard to grasp the correspondence between these different kinds of classification systems. 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, ℓ might indicate some side information such as where in the body the cell was found, X would indicate a detailed classification of the cell according to its transcriptomics, and Y would indicate a coarser 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 ℓ 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 works on a particular cultured cell-line, what can we say about whether a treatment will work on an actual in-vivo cancer inside a patient? 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. [11]). However, we know that cultured cell-lines look quite different from in-vivo cells (cf. [12, 13]). These cell cultures are subject to quite different pressures, due to the fact that they survive on a plate instead of inside a human being. The Markov Link method can leverage the common side-information together with separate transcriptome information to understand the correspondence between in-vivo and cultured cells. If a particular drug is effective on a particular cultured cell-line, we can then look at the corresponding in-vivo transcriptomic profile. If we find human cancers that match this profile, they might be good candidates for further research using this particular drug. Here ℓ 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 ℓ and Y. Thus the MLM assumption may hold.

- Text/image correspondence. Automatic image captioning is an ongoing effort in machine learning (cf. [14]). 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 ℓ 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. X would indicate the image and Y would indicate a caption. If caption is largely 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 [15]. 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. [16, 17]), MLM makes zero assumptions about what calibrations we might expect. In this case, ℓ 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 rest of the paper proceeds as follows:

- We describe the identifiability problem and describe how our analysis puts this problem in our place. We give several examples attempting to give intuition as to why the identifiability problem may not be as severe as it first appears.
- Applying these ideas, we develop the Markov Link Method for calibration without joint measurements.
- We apply the MLM to two single-cell transcriptomic methods. The result uncovers some ways in which the methods appear to be well-calibrated and other areas where the calibration may not be as good.

3 The identifiability problem

Let us 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 ℓ_i, X_i for $i \in \{1 \cdots n\}$ and ℓ_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 our observed data in two matrices:

$$N^X_{\ell x} = |\{i \le n : \ell_i = \ell, X_i = x\}| \qquad N^Y_{\ell y} = |\{i > n : \ell_i = \ell, Y_i = x\}| \tag{1}$$

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:

1. Not-enough-data problems. We don't have infinite data, so we can't know the exact values of p^*, h^*.
2. 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 $\Theta(p^*, h^*)$.

This point of view suggests a practical procedure for estimating q^*:

1. Use traditional methods to develop estimators $\hat{h}, \hat{p}$ so that with more and more data we can guarantee that $\hat{h} \approx h^*$ and $\hat{p} \approx p^*$ with high probability. In particular, we will use a robust pseudocount estimator for p^* and use the MLM assumption to get a high-quality estimator for h^* that leverages information from N^X and N^Y.
2. Then we may suppose that q^* is probably somewhere *near* to $\Theta(\hat{p}, \hat{h})$.

It is not completely obvious that this will work. Certainly $\hat{p} \approx p^*$ and $\hat{h} \approx h^*$. Certainly $q^* \in \Theta(p^*, h^*)$. But does that guarantee that q^* is nearby to the set $\Theta(\hat{p}, \hat{h})$? Something could go quite wrong in this line of reasoning. Fortunately, subject to mild assumptions, one can show that nothing goes wrong. This is the content of our main technical contribution. It can be found in Appendix A.

Let us now take a moment to understand the set Θ. This set encodes exactly how it is that the Markov Link assumption enables us to understand something about the calibration q^*. That is, 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 Θ can be written more concisely as a matrix equality constraint on q^*: if $q \in \Theta(p^*, h^*)$, then $p^* q = h^*$. This equation is in fact a literal mathematical rendering of the Markov Link assumption of conditional independence. The consequences of this assumption and its corresponding 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; $\Theta(p^*, h^*)$ is nontrivial and so it is *impossible* to ever determine the true value of q^*, regardless of how much data we have. All we have is the equality and inequality bounds that arise from our knowledge that $q^* \in \Theta(p^*, h^*)$.

To gain intuition about the equality and inequality bounds that define the set Θ, 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

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

We thereby obtain an important constraint on what the calibration q^* might be.

This equation also takes on additional significance in light of the inequality bounds we have: $0 \leq q^*_{xy} \leq 1$ for every x, y. In other words, probabilities must be between 0 and 1, by definition. This may seem trivial, but it has important consequences in light of the constraints imposed by the equation $p^*q = h^*$. In particular, taking another look at that same equation, we see that

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

It follows immediately that $q^*_{3,1} \geq .25$. Each equation in the MLM system yields insights of this kind; together these insights form the constraints that define Θ. 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 $q^*_{1,1} \geq 94.\bar{4}\%$.

Example 3. Finally, let us consider one class of examples in which the entire situation can be visualized: $\ell = 1$, $X \in \{1,2\}$, $Y \in \{1,2\}$. Notice that q is completely defined by $\theta_1 = \mathbb{P}(Y = 1|x = 1)$ and $\theta_2 = \mathbb{P}(Y = 1|x = 2)$. The matrix equation $p^*q^* = h^*$ corresponds to a linear constraint on θ_1, θ_2, namely $p^*_{11}\theta_1 + p^*_{12}\theta_2 = h^*_{11}$. On the other hand, the fact that q^* must be a valid probability distribution tells us that θ_1, θ_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 Θ. It can be identified with a line segment, and the 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 Θ to be a very small region indeed. See Figure 2 to visualize this. In real-world cases this Θ is much higher-dimensional and harder to visualize, but in this simple case we can see it clearly.

4 The Markov Link Method

Let us now exactly describe our approach for calibrating in the absence of joint measurement. As in Equation (1) above, we assume that X and Y are discrete random variables and N^X, N^Y

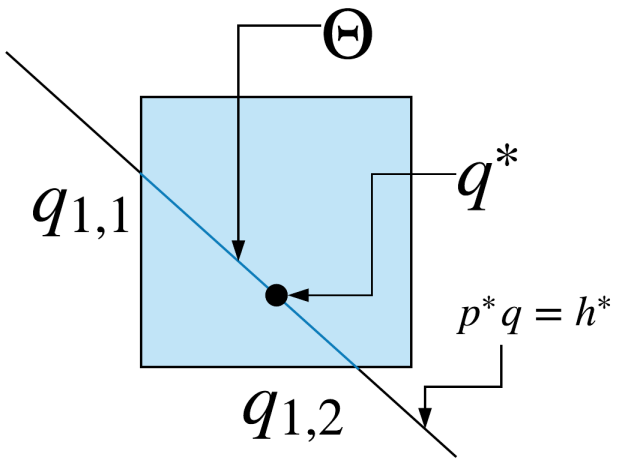

Figure 2: We consider the simple case $\ell = 1$, $X \in \{1,2\}$, $Y \in \{1,2\}$. In this case the space of values of the calibration q which are valid probability distributions can be understood as a box. The space of values of q which satisfy the matrix equation $p^*q = h^*$ can be understood as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

are the matrices describing the 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 has three main components:

Estimation Estimate p^* with robust pseudocount estimator (cf. [18]):

$$\hat{p}(x|\ell) \triangleq \frac{1 + N^X_{\ell x}}{\sum_{x'} N^X_{\ell x'}} \tag{2}$$

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

$$\hat{q} = \arg\max_q \sum_{\ell,y} N_{\ell y} \log\left(\sum_x \hat{p}(x|\ell) q(y|x)\right) + \kappa \sum_{xy} \log q(y|x) \tag{3}$$

The Dirichlet prior regularizer κ ensures this optimization problem has a unique solution; indeed, one may easily verify that the objective is strictly convex. In practice, we took κ to be any small constant. We use a majorization-minimization approach to solve this optimization problem. That is, we first take an initial guess for $\hat{q}$. We then define $\tilde{\pi}_{\ell,x,y} \leftarrow \hat{p}(x|\ell)q(y|x) / \sum_{x'} \hat{p}(x'|\ell)q(y|x')$. We then update $\hat{q}$ by $\hat{q} \leftarrow \arg\max_q \sum_{\ell,y,x} N_{\ell y} \tilde{\pi}_{\ell,x,y} \log(\hat{p}(x|\ell)q(y|x)) + \kappa \sum_{xy} \log q(y|x)$, which can be solved in closed form. We then iterate this procedure, redefining $\tilde{\pi}$ and then redefining $\hat{q}$ in each iteration. It is well-known that since the overall problem is convex we can guarantee that this procedure converges to the global optima (cf. [19] for a detailed exposition). In practice we found 5000 iterations to be sufficient for convergence.

This yields a high-quality, consistent estimator for $h^*(y|\ell) \triangleq \mathbb{P}(Y = y|\ell)$, namely $\sum_x \hat{p}(x|\ell)\hat{q}(y|x)$ (cf. the lemma in Appendix A for a simple consistency proof). It also gives us a unique estimate for q^*, although this estimate may suffer from errors due to incorrect assumptions, lack of data, or identifiability. Therefore, a vital part of the MLM is a criticism step.

Model criticism The MLM assumption cannot be checked directly. However, we can check certain aspects of it. In particular, the MLM assumption suggests a natural estimator for the distribution . On the other hand, a more traditional approach would etimate h^* from N^Y using a simple pseudocount method. We can compare both estimators for h^* using held-out log-likelihood of Y data. If the MLM estimate is considerably worse than the pseudocount method, it may indicate that the MLM assumption doesn't fit with the data.

Uncertainty characterization Measure our uncertainy about q^* using *Bootstrapped Rotationally Uniform eXtremal Paired (BRUXP)* samples and *Rotationally Uniform eXtremal Paired (BRUXP)*. The variability in these samples capture error both due to insufficient data and due to identifiability issues.

We now describe this very last step in more detail:

Definition 1. Let d denote a random direction in calibration space. That is, let d denote a matrix of the same dimensions as the calibration q, each entry of which is sampled from a random normal distribution. Sample with replacement to produce a surrogate dataset $N^{X,(1)}, N^{Y,(1)}$, estimate $\hat{p}^{(1)}, \hat{q}^{(1)}$ using the surrogate dataset and the method outlined above, and define $\hat{\Theta}^{(1)} \triangleq \Theta(\hat{p}^{(1)}, \hat{p}^{(1)}\hat{q}^{(1)})$. Repeat this process to obtain $\hat{\Theta}^{(2)} \triangleq \Theta(\hat{p}^{(2)}, \hat{p}^{(2)}\hat{q}^{(2)})$. Now, let us take $\tilde{q}^{(1)}$ to be the most extremal vertex of $\hat{\Theta}^{(1)}$ in the direction d and $\tilde{q}^{(2)}$ to be the most extremal vertex of $\hat{\Theta}^{(2)}$ in the direction $-d$. The tuple $d, \tilde{q}^{(1)}, \tilde{q}^{(2)}$ is said to be a **Bootstrapped Rotationally Uniform eXtremal Paired (BRUXP)** sample. If we adopt this same procedure without the resampling-with-replacement aspect, we will refer to it as **Rotationally Uniform eXtremal Paired (RUXP)**.

For each BRUXP sample, the pair $\tilde{q}^{(1)}, \tilde{q}^{(2)}$ gives for the range of what q^* might be along the direction d. The variability amongst many such samples indicates our uncertainty about the true $\hat{q}$.

We cannot make general claims about the bootstrap in this case, due to the usual difficulties with bootstrap and extremal statistics (cf. [20]). However, we do at least know that this procedure correctly handles the identifiability issue in the asymptotic limit. In particular, subject to mild reguarity conditions, we can show that $\inf_{q \in \Theta(\hat{p}, \hat{p}\hat{q})} |q - q^*| \to 0$ as we gather sufficient data. That is, with sufficient data BRUXP samples are guaranteed to represent the uncertainty we face from identifiability concerns. An exact statement may be found in Appendix A.

The the variation amongst the BRUXP samples is necessarily due to a combination of effects from insufficient data and effects from the identifiability issues. To separate issues, we recommend that one also look at Rotationally Uniform eXtremal Paired (RUXP) samples (i.e. BRUXP samples taken without using the sampling-with-replacement procedure). The variability among these samples corresponds primarily to uncertainty due to identifiability concerns, and ignores variability due to insufficient data.

Code for this procedure is available at https://github.com/jacksonloper/markov-link-method. All calculations in this paper can be replicated by going through a tutorial-style ipython notebook that can be found at the same address.

5 Empirical results

5.1 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, both methods have a notion of a "Lamp5 Egln3_1" cell-type. 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 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. [10]) 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. Indeed, it was found necessary in method II to designate some cells with certain "nxx" types that indicated different kinds of uncertainty about the cell-type. Thus even the set of cell-types was not the same between the two groups. Clearly the methods were not perfectly calibrated – but

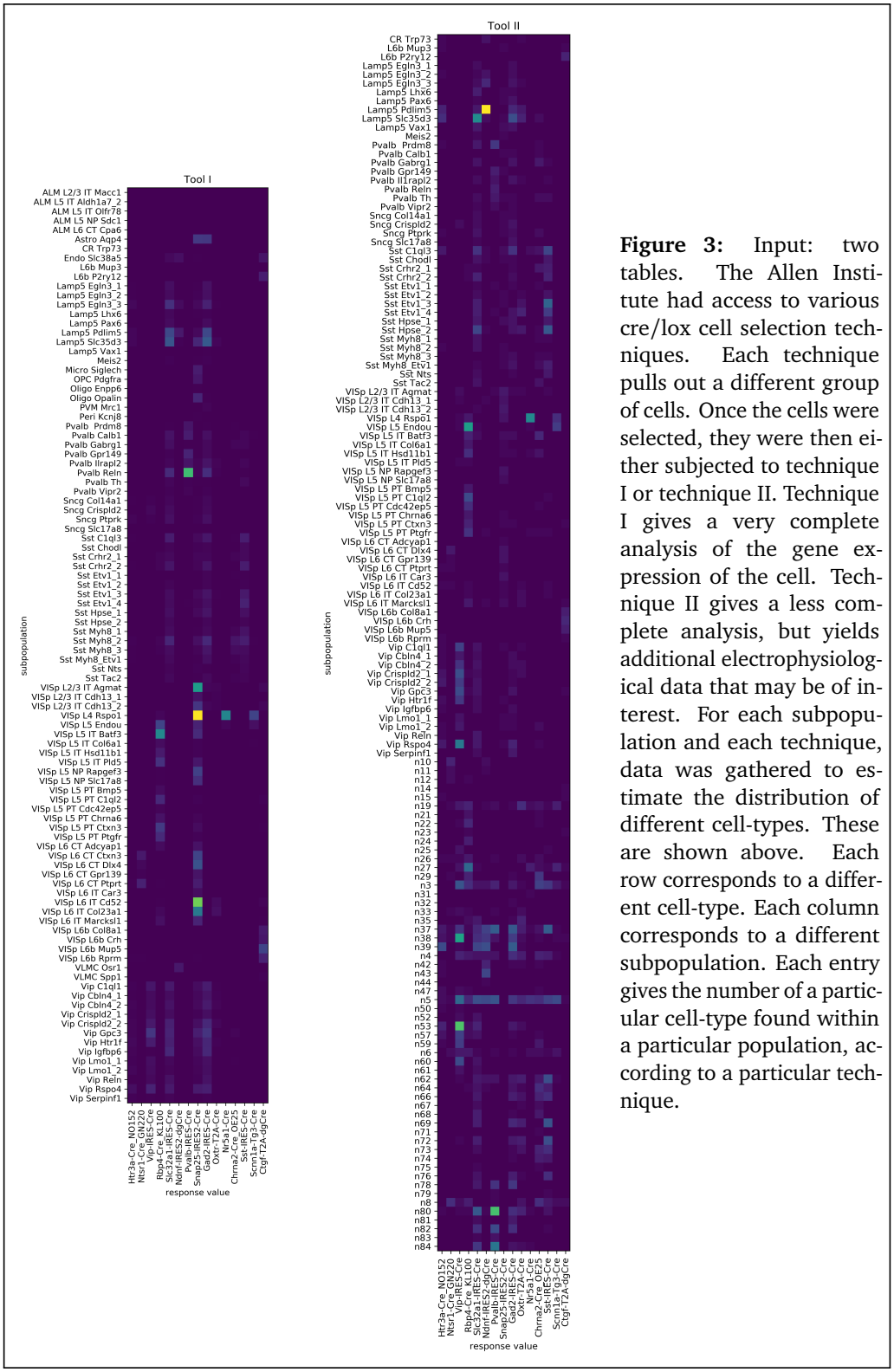

Figure 3: Input: two tables. The Allen Institute had access to various cre/lox cell selection techniques. Each technique pulls out a different group of cells. Once the cells were selected, they were then either subjected to technique I or technique II. Technique I gives a very complete analysis of the 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 row corresponds to a different cell-type. Each column corresponds to a different subpopulation. Each entry gives the number of a particular cell-type found within a particular population, according to a particular technique.

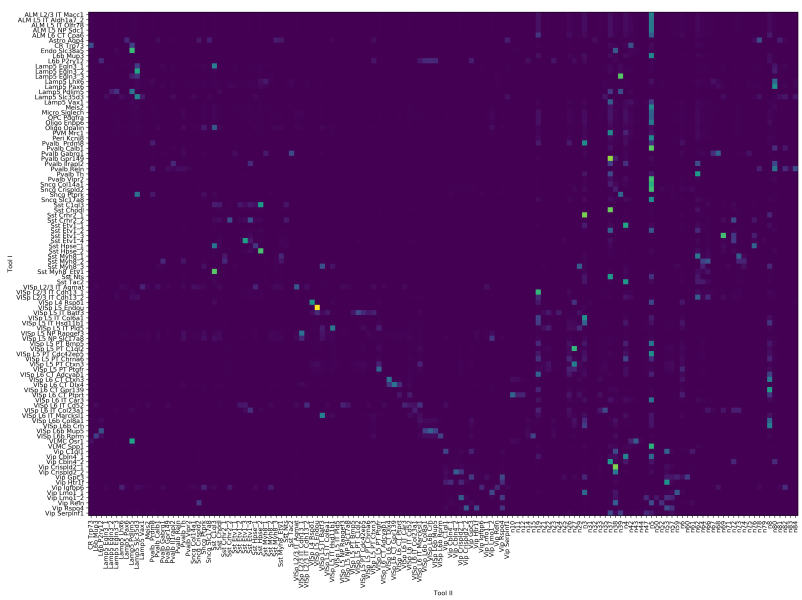

Figure 4: The $\hat{q}$ estimated by the MLM method with $\kappa = .01$. Notice the mass along a somewhat diagonally-oriented slice of the calibration. This corresponds to cell-types which appear to be designated with the same cell-type by both methods. The large region on the right corresponds to the "nxx" types in Tool II; these indicate different forms of uncertainty about the type of the cell.

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.

5.2 Estimation

The first step of the Markov Link Method is to estimate the calibration with the estimator $\hat{q}$ according to equations (2) and (3). Taking $\kappa = .01$, we obtained a calibration found in Figure 4. From this, it appears that there are many cells which would be identified with the same cell type in either method. For example, according to $\hat{q}$, if a cell is classified as type "VISp L5 Endou" by tool I, it appears there is a 91% chance it will be classified the same way by tool II. However, there are other cell-types which seem to have more ambiguity. Lest we make too many conclusions without considering the uncertainty in our estimates of q^*, let us proceed to the criticism phase of the MLM.

5.3 Model criticism

The Markov Link Method assumption suggests a natural estimator for the distribution of tool II results (Y) under each subpopulation (ℓ). We have denoted this distribution h^*, and the MLM estimates this distribution with $h^*(y|\ell) \approx \sum_x \hat{p}(x|\ell)\hat{q}(y|\ell)$. A more traditional method would be to estimate h^* using a pseudocount method (cf. [18]). We can compare these methods using held-out log likelihood. That is, we train both methods using a subset of our data. Then we compute the negative log likelihood of the held out data under both models. This indicates how concisely the model can represent the held out data. If the MLM method is much worse, it may be a sign that the MLM assumption is poor. Smaller values are better:

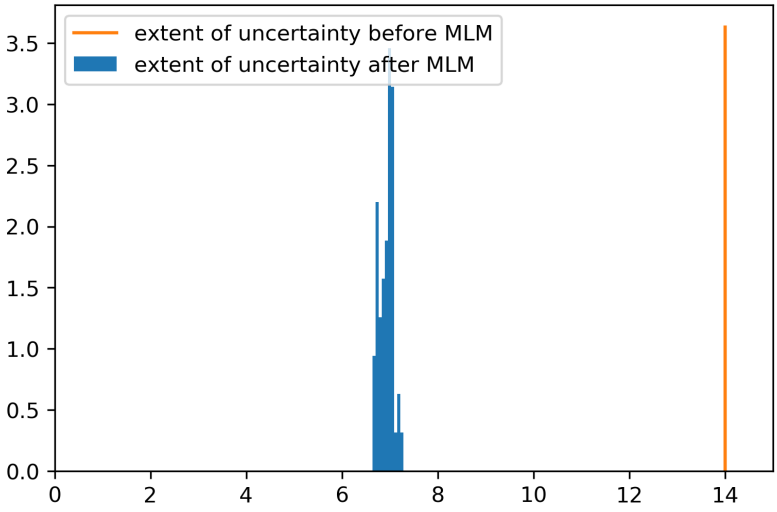

Figure 5: Here we plot a histogram of $|\tilde{q}_i^{(1)} - \tilde{q}_i^{(2)}|$ among fifty BRUXP samples $\{d_i, \tilde{q}_i^{(1)}, \tilde{q}_i^{(2)}\}_{i \in 1 \cdots 50}$. Each distance estimates the extent of our uncertainty about the calibration q along the direction d. We can compare this with the extent of the uncertainty we had before we applied our method, i.e. the diameter of the set of probability matrices with these dimensions. In this case the diameter of that set is exactly 14. In this sense, the MLM has enabled us to roughly half our uncertainty about the calibration.

- Traditional estimator: 3.96 nats per cell
- MLM estimator: 3.70 nats per cell

This is encouraging. Our model is as good at predicting held-out data as a more traditional pseudocount estimator (which makes no assumptions on the conditional independence structure of the data). In fact, the MLM estimate is able to outperform the traditional estimate, presumably because it can leverage the data in N^X to help us learn about the distributions on Y. At the very least, it is not a clear sign that the MLM assumption is inappropriate for this data.

5.4 Uncertainty characterization

We obtained BRUXP samples and RUXP samples for this dataset. These can be used to understand our uncertainty about the calibration q^* in an number of ways.

The distances between the paired vector in a BRUXP sample give us one kind of insight into the extent of our uncertainty. Recall that for each direction d the BRUXP sample comprises the extremal vertex in the direction d as well as the extremal vertex in the direction $-d$. The distance between these two vertices gives a kind of measure for the extent of our uncertainty. We plot a histogram of these distances in Figure 5. This figure suggest that this extent is around 7 (although we caution that general results on polytope diameter estimation are not encouraging, cf. [21]). By contrast, the diameter of the entire space of valid probability matrices is 14. In this sense, the Markov Link Method enables us to half our uncertainty about the calibration.

Qualitatively, we can inspect the BRUXP samples by focusing on a single value of X at a time. For example, in Figure 6 we focus on the distribution of Y given that method I measures a specimen as having the type 'Vip Rspo4.' In this figure we look at various BRUXP samples,

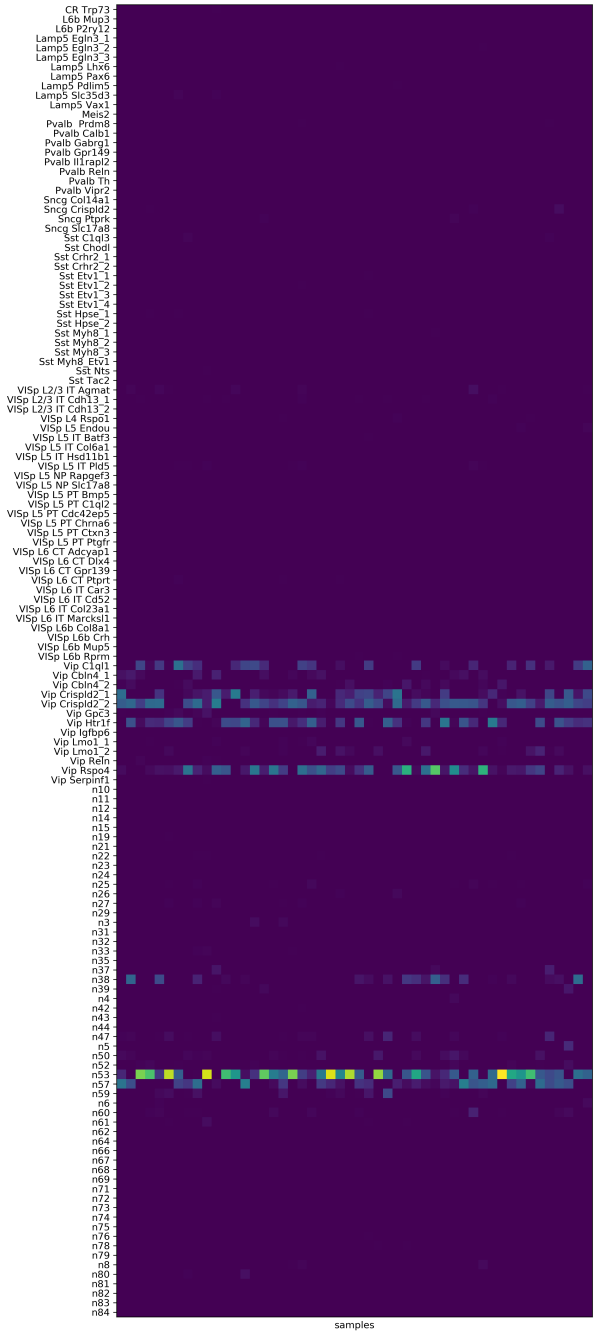

Figure 6: Each column represents a possible manner in which method I and method II are calibrated for cells which are designated as type 'Vip Rspo4' by method I. Together, the columns form a notion of our uncertainty about this aspect of the calibration. This plot shows that cells designated as 'Vip Rspo4' by method I may be designated by method II as taking on various 'Vip'-related celltypes and or some of the 'nxx' celltypes. We cannot be sure which of these columns more faithfully represents the true calibration between the methods.

lining up all the pairs into a single heatmap. We see that cells designated as type 'Vip Rspo4' by method I may be associated with one of the 'Vip'-related celltypes in method II or an 'n53' or 'n57' type (recall that these types indicate a group of possible cell types, of which 'Vip Rspo4' is one). For example, it may be that method II would identify such cells as 'Vip Htr1f' cells. This is potentially concerning. On the other hand, the plot above suggests that these cells are *not* ever identified as 'Vip Lmol_1' cells. That is, method II is able to accurately discern between cells designated as 'Vip Rspo4' by method I and cells designated as 'Vip Lmol_1' by method I. This is encouraging. To further resolve the various remaining ambiguities in the calibration, we would need to perform additional experiments involving a subpopulation that might be likely to include 'Vip Rspo4' cells but no 'Vip Htr1f' cells.

Let us next look at the BRUXP samples for cells designated by tool I as having the 'VISp L5 Endou' type; these samples are fairly unequivocal. Figure 7 gives strong support for the notion that the methods are very well-calibrated with respect to this cell-type. In all of the BRUXP samples we see that a cell designated as 'VISp L5 Endou' by method I would be likely to be designated with the same type with method II.

In the interests of trying to disentangle our uncertainty due to insufficient data from our uncertainty due to identifiability, we repeat the histogram-of-distances analysis on RUXP samples. Recall that RUXP samples are just like BRUXP samples, only we do not perform the resampling with replacement. This RUXP histogram gives us a lower bound on the diameter of the set $\Theta(\hat{p}, \hat{q})$. We expect this diameter to be smaller than the extent we measured in the first analysis: the RUXP analysis only captures uncertainty due to identifiability whereas the BRUXP analysis also captures our uncertainty due to insufficient data. This is indeed what we find. Figure 8 shows the result of the histogram-of-distances analysis on the RUXP samples. We see distances are generally closer to 6 and never as high as 7. Recall that 7 was the typical distance found in the original BRUXP analysis.

With more samples, we expect that we could get our original BRUXP analysis in Figure 5 to look more like the RUXP analysis in Figure 8. However, it appears that more samples might not be sufficient to bring the diameter of our uncertainty below 6. For that, we would need different *kinds* of experiments, using different subpopulations. For example, as we saw above, we could reduce some of the identifiability concerns with experiments using a new subpopulation that included 'Vip Rspo4' cells but not 'Vip Htr1f.'

Finally, in this special case there is an additional analysis that is appropriate. The team producing method I and method II designed these methods to have corresponding types. The hope was that a cell designated as 'Vip Rspo4' in method I would also be designated as 'Vip Rspo4' in method II. As we have seen above, we cannot be exactly sure what the calibration is. Is it perhaps possible that the calibration is nearly perfect? We can place a kind of upper bound on the precision of the calibration by looking for the *best possible* calibration that is consistent with the data we have observed. In particular, let d denote the direction of "perfect calibration" in which method I and method II agree perfectly. We can find the value of $q \in \Theta(\hat{p}, \hat{p}\hat{q})$ which is most extreme in this direction.

We show this 'best possible q' in Figure 9. The result is very encouraging: this best q suggests a strong calibration between the two methods. We emphasize that our method does not prove this calibration by any means; as we have seen, the data only allow us to know that the calibration up to a considerable degree of uncertainty. However, we can conclude that the observed data does not *rule out* the possibility that the methods are actually quite well calibrated on the whole. Of course, even for this best q we do see certain miscalibrations that might be worth further investigation. For example, it appears that cells designated as 'Vip Serpinf1' by method I might be designated as type 'Lamp5 Slc35d3' by method II.

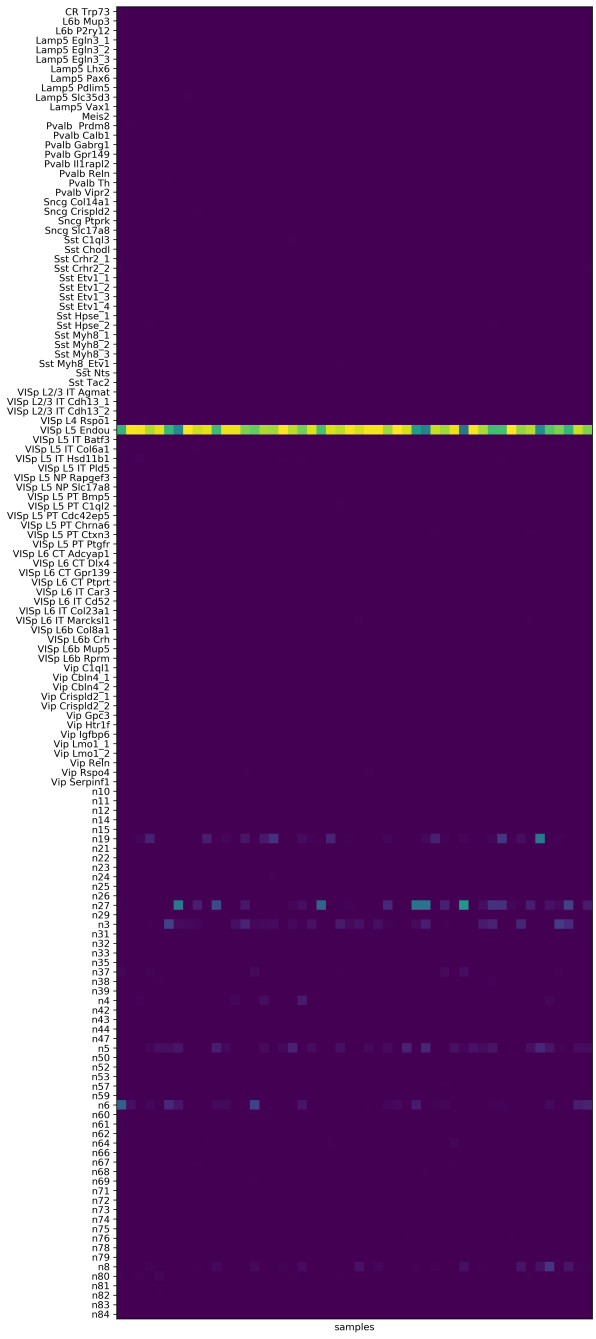

Figure 7: Each column represents a possible manner in which the methods are calibrated for cells designated by method I as being type 'VISp L5 Endou.' Notice that in every column there is strong signal at the corresponding type in method II. This suggests a good calibration.

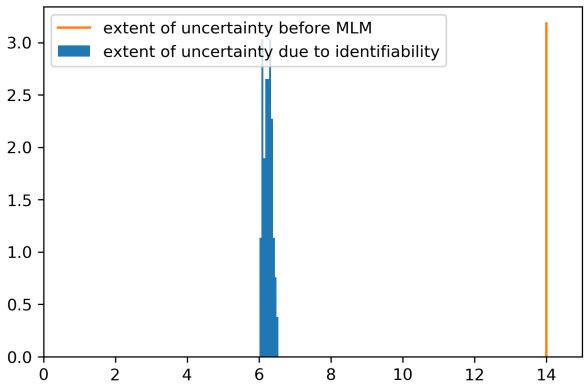

Figure 8: Here we plot a histogram of $|\tilde{q}_i^{(1)} - \tilde{q}_i^{(2)}|$ among fifty RUXP samples $\{d_i, \tilde{q}_i^{(1)}, \tilde{q}_i^{(2)}\}_{i \in 1 \cdots 50}$. Each distance estimates the extent of our uncertainty about the calibration q along the direction d that is due to identifiability issues alone. These distances do not measure the extent of our uncertainty due to insufficient data. We can compare this with the kinds of distances we saw in Figure , which were generally larger because they measured uncertainty due to both insufficient data and identifiability issues.

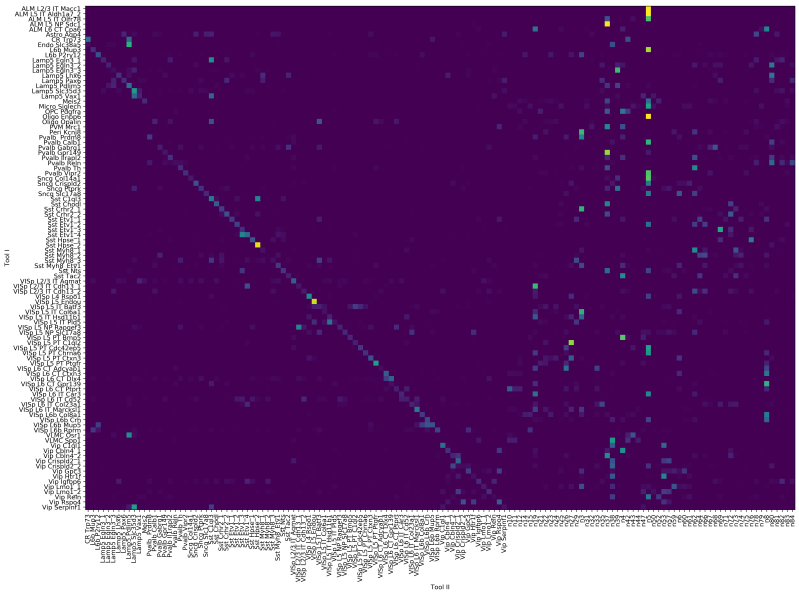

Figure 9: The value of q that yields the most precise correspondence between the two cell types while remaining consistent with the observed data. This gives a kind of upper bound on how well the methods may be calibrated. As we have seen throughout, there is an extent to which we simply cannot be sure what the calibration is. Within all this uncertainty, this figure shows something like the rosiest possible picture of how well the methods might be calibrated. The strong diagonal components in this calibration show a good correspondence, i.e. cells designated as a particular type under method I are likely to be associated with the same type in method II. We caution that there is no reason to suppose that this *is* the true calibration, but this figure shows that the data does not rule out that possibility.

6 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, we can rigorously bound our uncertainty. We have proposed the Markov Link Method as procedure to estimate the calibration and understand our uncertainty regarding that estimate. We investigated a real-world calibration problem; the MLM gave bounds on the accuracy of the calibration and also the suggested directions for future experiments to further refine our uncertainty about this calibration. Code is published at https://github.com/jacksonloper/markov-link-method, including a tutorial-style ipython notebook detailing every calculation used in this paper.

The Markov assumption is of course not the only one that we could have used, and may not be valid in every case; future work may be to investigate others. 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 simply excised from the data without comment. This would violate 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. Indeed, if the MLM method gives insensible results, this could actually serve as a useful clue that this disproportionate cell death is happening.

Another potential direction for future work is suggested by the 'nxx' types found in method II of our empirical investigation. These types are meant to indicate some uncertainty about the exact cell type. Each one is associated with a family of cell types. Our analysis did not need to rely on this information to proceed. However, making better use of it could considerably improve our results. Developing a rigorous framework to incorporate these kinds of structured data would be a worthy endeavor.

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

[3] 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.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] Peter Flach. *Machine learning: the art and science of algorithms that make sense of data*, page 279. Cambridge University Press, 2012.

[19] CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[20] Donald WK Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68(2):399–405, 2000.

[21] 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.

A 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

$$\hat{q} = \arg\max_q \left(\sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log \left(\sum_x \hat{p}(x|\ell) q(y|x) \right) + \kappa \sum_{xy} \log q(y|x) \right) \tag{4}$$

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 (albeit technical) to see that the same goes for $\hat{p}\hat{q} \to h^*$ (see Lemma 1). Thus, intuitively, the difficulty is this: by ensuring $|\hat{p} - p^*|_\infty, |\hat{p} - p^*|, |\hat{p}\hat{q} - h^*|_\infty, |\hat{p}\hat{q} - h^*|$ 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

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

Now, since $\hat{q}$ is defined as the maximizer of a certain quantity (Equation 4), we may be sure that it is greater than the same quantity evaluated at $q = q^*$. That is,

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

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. □