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

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

July 19, 2018

Abstract

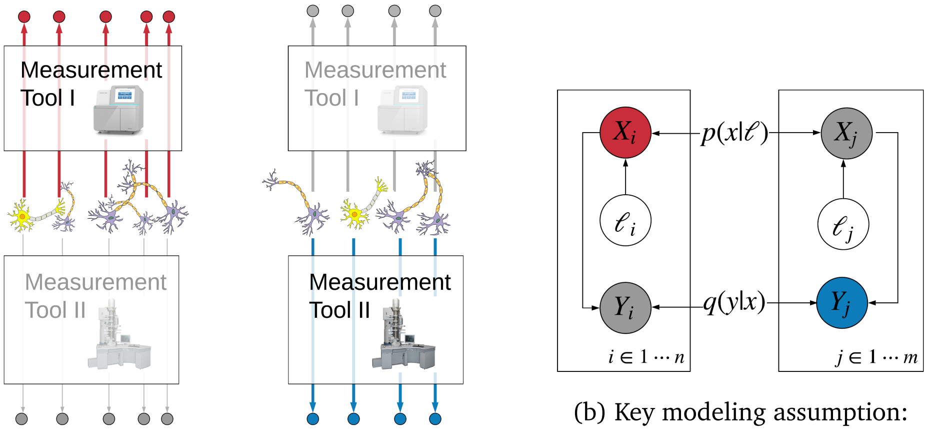
A proliferation of new experimental tools has left a serious gap: calibration. Two thermometers can be calibrated against each other by simply measuring the same bodies of water with both thermometers, but the problem is much harder for many modern tools. One common problem is that we do not have measurements 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 can be applied even if we cannot make any assumptions about what calibrations we might expect to see; instead, it uses 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 $q^*(y|x)$. Let us say we obtain measurement result x from one tool on a particular specimen. The number $q^*(y|x)$ indicates 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 with 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 applied 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 the subpopulations are highly heterogeneous and overlapping. 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.

This paper will proceed in the following sections:



(a) Setup: each specimen can be measured with either tool I or tool II, but can't be measured with both. How do we calibrate the tools?

(b) Key modeling assumption: the calibration q is the same for every subpopulation.

Figure 1: The Markov Link Method assumption makes calibration possible, even without joint measurement. In subfigure (a) we measure some neurons with tool I and others with tool II. How can we calibrate the tools? That is, how can we know what we might have seen if we had measured the first group with tool II or the second group with tool I? The Markov Link Method (MLM) gives a way to get a partial answer, using a subpopulation analysis. In subfigure (b) we present a plate diagram describing our beliefs about the conditional independencies in this situation; cf. [2] for a thorough explanation of this type of diagram. For each specimen we are interested in ℓ (subpopulation), X (tool-I measurement), and Y (tool-II measurement). For n samples we can see (ℓ, X) and for m samples we can see (ℓ, Y). However, for the first group of n samples we cannot observe Y, and for the second group of m samples we canot observe X. This is indicated in the diagram by the grayed-out circles. This lack of joint measurement makes it hard to learn the variable distributions. The MLM has a key assumption that makes it possible: we assume X is sufficiently detailed to *statistically isolate* Y from ℓ. This assumption is reflected in the arrows of the diagram; all paths from ℓ variables to Y variables pass through X variables. The assumption allows us to learn much about the important probability distributions, as parametrized by the tool-I measurement distribution $p(x|\ell)$ and the calibration $q(y|x)$.

- Description of how to apply the Markov Link Method, using an example to explain the process
- A variety of examples where the method may apply
- The method's performance in simulation
- The method's results on real-world data
- Relation to prior work
- Discussion

1 The Markov Link Method

We begin with an example.

Imagine that we have a large collection of human cells. We will call each cell a 'specimen.' For each specimen we have:

1. ℓ, the sampling strategy used to obtain the specimen. We will also call ℓ the specimen "subpopulation," for each strategy should capture a different subpopulation of the overall set of specimens. For example, we could have a few strategies based on selecting cells of particular sizes, a few strategies based on selecting cells from particular places in the body, and a few strategies based on based on selecting cells with particular proteomic markers. Each additional sampling strategy shows us a different subpopulation of the cells and strengthens what kinds of inferences are possible.
2. X, the cell's 'transcriptomic type,' as measured by looking at the transcriptomic activity in the cell.
3. Y, the cell's 'morphological type,' as measured by looking at an image of the cell.

Our task is to reconcile the two different notions of 'cell-type' indicated by X and Y, by constructing what we call a 'calibration.' This calibration should tell us what transcriptomic types are associated with what morphological types. For example, we might like to be able to say "This cell has transcriptomic type 3, therefore it probably has morphological type F." We can make math of this idea by defining a calibration as a conditional probability $q^*(y|x,\ell)$. This object indicate the likelihood that the morphological type is y given that the transcriptomic type is x and the specimen was sampled with strategy ℓ. In these terms, our task is to learn the calibration q^*.

Learning q^* is hard when "joint measurement" is impossible or impractical. By this we mean that for any given specimen we can either observe ℓ, X or ℓ, Y. We can never observe ℓ, X, Y for any specimen. This makes it much more difficult to estimate the calibration; the fundamental difficulty is sketched in Figure 1. To make calibration possible without joint measurement, we need an assumption additional assumption. The Markov Link Method supplies this assumption in the form of a conditional independence assumption. This makes it possible to learn about the calibration.

To make it possible to learn q^* even without joint measurement, the MLM makes the assumption that the calibration is the same for every sampling strategy:

The Markov Link Method Assumption
$q^*(y|x,\ell) = q^*(y|x)$ for each ℓ

This assumption can be understood through the following thought experiment. Let us say we are told that a cell has transcriptomic type x. We could then make predictions about the morphological type y. Now we are told new information: the cell was obtained via sampling strategy ℓ. Does this change our predictions about the morphological type? If so, the MLM assumption is violated. However, if X is sufficiently informative then learning the sampling strategy will not change our predictions about the morphological type. In this case, the MLM assumption holds.

If the MLM assumption holds, a frequentist point of view suggests the following procedure for estimating the calibration q^* from data:

1. Summarize our observed data in two matrices:

$$N^X_{\ell x} = |\{i \leq n : \ \ell_i = \ell, X_i = x\}|$$
$$N^Y_{\ell y} = |\{j \leq m : \ \ell_j = \ell, Y_j = x\}|$$

   That is, $N^X_{\ell x}$ indicates the number of specimens from subpopulation ℓ with transcriptomic type x. Likewise $N^Y_{\ell y}$ indicates the number of specimens from subpopulation ℓ which with morphological type y.

2. Define a confidence region for calibration q^*. To do this we incorporate a nuisance parameter: let $p^*(x|\ell)$ denote the probability that a cell sampled with strategy ℓ will have transcriptomic type x. We form a joint simultaneous confidence region for all of p^*, q^* using the following likelihood ratio statistic:

$$L(p,q) = \frac{1}{\Omega_X} \sum_{\ell,x} N^X_{\ell x} \log \frac{\hat{p}(x|\ell)}{p(x|\ell)} + \frac{1}{\Omega_Y} \sum_{\ell,y} N^Y_{\ell y} \log \sum_x \frac{\hat{h}(y|\ell)}{p(x|\ell) q(y|x)}$$

   where $\hat{p}(y|x) = N^X_{\ell x} / \sum_{x'} N^X_{\ell x'}$ and $\hat{h}(y|x) = N^Y_{\ell y} / \sum_{y'} N^Y_{\ell y'}$ denote empirical conditional distributions. We then define our confidence region as

$$R = \{(p,q) : \ L(p,q) < k\}$$

   and choose k to ensure any desired coverage. For example, to produce a 95% confidence region, we would like to choose k so that the probability that $p^*, q^* \in R$ is at least 95%. We refer the reader to Appendix A for details on choosing k.

3. Determine whether the confidence region is empty. If the region is empty then we can reject the MLM assumption hypothesis with a significance level of one minus the region coverage. Assuming we fail to reject, we can continue:

4. Sketch out the the confidence region, using examples. In most nontrivial cases the region R is high-dimensional and difficult to inspect directly. We therefore produce examples from this region to try to help the user understand what values of q are plausible. In particular, we sample $Z_{x,y} \sim$ Exponential(1) for each x, y and solve the problem

$$\max_{p,q \in R} \quad \sum_{x,y} Z_{x,y} \log q(y|x)$$

   This problem does not have a closed-form solution, but can generally be solved using an iterative procedure; we describe our method in detail in Appendix C. The result is a particular value of q which lies within R. In practice, we find these examples are almost always on the boundary of R. We can then repeat this process by resampling Z. By repeating this process many times, we can get a large collection of plausible values of q. For any fixed value of x we can visualize these samples by using a heatmap to plot $q(\cdot|x)$ for all the samples. See Figure 2 for an example.

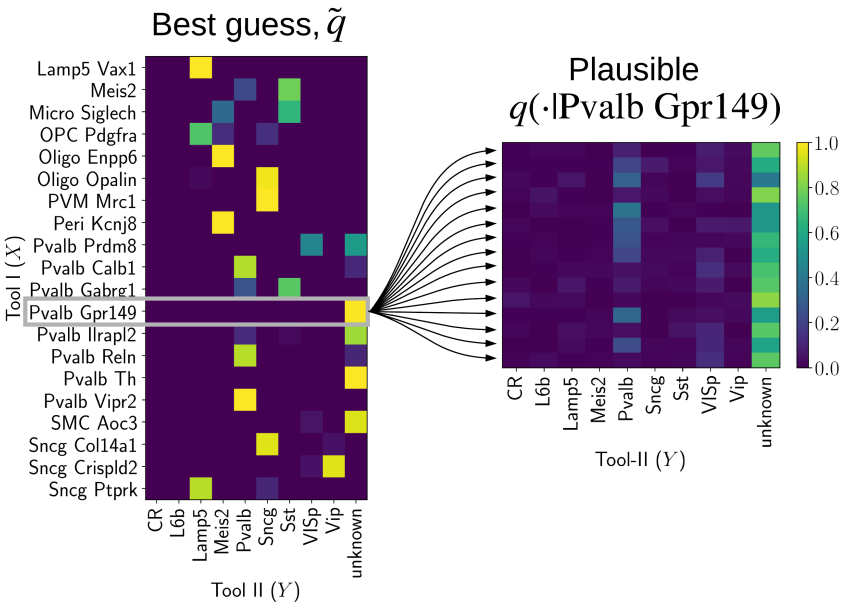


Figure 2: Estimating the calibration and measuring our uncertainty. On the left we visualize our best guess for true the calibration between two different tools for measuring cell-type. Tool I uses one approach to obtain a fine-grained classification of cells, whereas tool II uses a different approach and only gives coarse type indication. We can visualize our estimated calibration with a heatmap, formed by drawing a rectangle for each x, y with a color indicating the magnitude of $\tilde{q}(y|x)$. For example, examine the bright yellow square on the 'Pvalb Gpr149' row and the 'unknown' column. This suggests that if a cell has the 'Pvalb Gpr149' type according to tool I, it is extremely likely to be assigned the 'unknown' type by tool II. However, our best guess might not be correct. On the right we show many plausible alternatives for the 'Pvalb Gpr149' row of the calibration. Each possibility is drawn from a 95% confidence region for calibration. Some of the alternatives suggest that some 'Pvalb Gpr149' cells might not be classified as 'unknown,' but might instead be correctly identified as being of the 'Pvalb' type. The data we have cannot tell us which of these alternatives is closest to the truth. More data and more subpopulations would be necessary to reduce the variability within this confidence region. However, we can have some confidence that 'Pvalb Gpr149' cells are not being classified as 'Meis2' or 'Vip', insofar as none of the plausible calibrations give high probability to these tool II types.

5. Produce a point estimate. It is not generally possible to produce a consistent point estimate. Indeed, there are many situations in which the confidence region R does not concentrate even in the limit of infinite data. Nonetheless, if a point estimate is desired, one may pick point estimates $\tilde{p}, \tilde{q}$ by solving the minimization problem $\min L(p,q)$. We caution that this problem does not always have a unique solution, so the result may depend upon the minimization strategy used. Precise details for our algorithm may be found in Appendix C.

A Bayesian point of view would suggest a different procedure; this point of view may provide a useful complementary perspective and is worthy of future research. However, we caution that it is nontrivial to choose reasonable prior distributions for p^*, q^*. Priors which may seem uninformative (such as the Dirichlet distribution) can actually carry significant and unanticipated weight, due to the high-dimensional nature of the problem. We give an example in B.

The frequentist confidence region allows us to avoid making any modeling assumptions. However, we are then faced with the task of understanding a high-dimensional confidence region. As described above, our solution is to look at examples from inside the region. We emphasize that these examples are not drawn uniformly within the confidence region; instead the examples illustrate something more like the boundary of the region, sketching out what might be possible. We concede that this method is somewhat ad-hoc and hope that future research may discover a superior method to visualize this high-dimensional confidence region. For now, it appears to be a practical solution.

2 Examples where the Markov Link Method may apply

The validity of the Markov Link Method assumption 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. One 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. [3]). 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. [4, 5]). A more modern approach is to assay the cell's transcriptome (cf. [6]). 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 holistic 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.

- Radiometric calibration. Different cameras measure light differently. For example, each camera has different lens distortions. Different cameras also have different ways of transforming photon counts into digital information. Fortunately, joint measurement is generally possible with cameras; simply take a picture of the same object with both cameras. Unfortunately, an adequate amount of joint measurement is sometimes hard to come by. For example, with expensive astronomy-grade settings, it can be difficult to balance the need for calibration with the total amount of the sky one wants to cover (cf. [7]). For example, instead of requiring different cameras to take pictures of exactly the same portion of sky at exactly the same time, the subpopulations ℓ could represent portions of the sky. Various conditions may cause these portions of the sky to appear differently over time, but if we assume this variability is independent of the calibration itself, the MLM assumption may apply.

- 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. [8]). However, we know that cultured cell-lines look quite different from in-vivo cells (cf. [9, 10]). 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. [11]). 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 the 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 [12]. However, some of the problem may simply be a matter of calibration. 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 processes used by each lab. Once the labs are properly calibrated, we can meaningfully combine both datasets. Unlike other tools to deal with lab or batch effects (e.g. [13, 14]), 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.

3 Simulation results

We use simulations to test the capabilities of the Markov Link Method to estimate a calibration. We will:

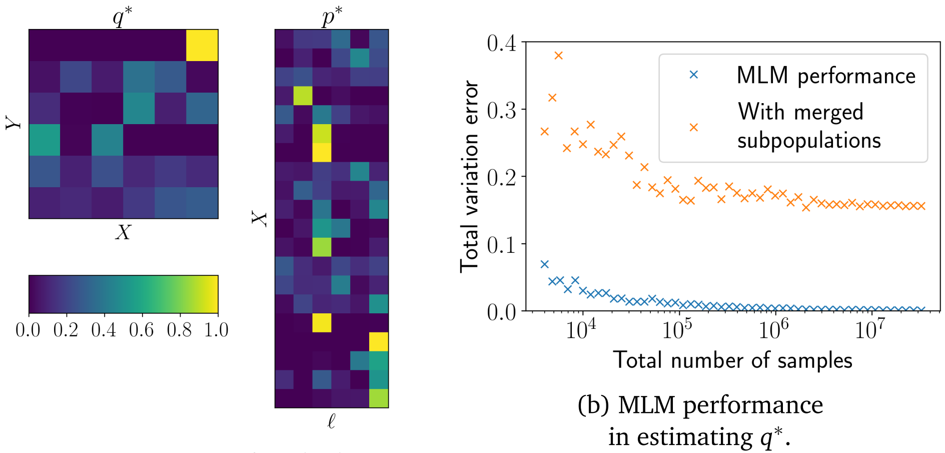
- Pick parameters to be ‘ground truth’:
  - A calibration, $q^*(y|x)$.
  - A tool-I measurement distribution, $p^*(x|\ell)$. Recall that this indicates the probability that tool I measures x for a specimen gathered with sampling strategy ℓ.
  - A tool-II measurement distribution, $h^*(y|\ell)$. This indicates the probability that tool II measures y for a specimen gathered with sampling strategy ℓ. We assume the the MLM assumption holds, and so this can be calculated from q^*, p^* by the formula $h^*(y|\ell) = \sum_x p^*(x|\ell)q(y|x)$.
- Select a number of samples to take per subpopulation, s, and produce a simulated dataset for each subpopulation ℓ:

$$(N^X_{\ell 1}, N^X_{\ell 2} \cdots N^X_{\ell \Omega_X}) \sim \text{Multinomial}(s, p^*(\cdot|\ell))$$
$$(N^Y_{\ell 1}, N^Y_{\ell 2} \cdots N^Y_{\ell \Omega_Y}) \sim \text{Multinomial}(s, h^*(\cdot|\ell))$$

- Apply the Markov Link Method to the generated samples to obtain a calibration estimate $\tilde{q}$.
- Compute the error of the estimate $\tilde{q}$, as measured by a kind of averaged total variation distance, namely

$$\frac{1}{2\Omega_\ell} \sum_{\ell x} |\tilde{q}(y|x) - q^*(y|x)|$$

  This distance ranges between zero and one. A distance of zero indicates that $\tilde{q} = q^*$, and a distance of one indicates that the two calibrations place mass on completely disjoint sets, i.e. $q(y|x) = 0$ whenever $\tilde{q}(y|x) > 0$ and vice-versa.



(a) 'Ground truth' for p^*, q^*.

(b) MLM performance in estimating q^*.

Figure 3: The Markov Link Method accurately estimates the calibration, as long as there are enough subpopulations. We pick parameters as 'ground truth' for p^*, q^*; these are as shown in subfigure (a) as heatmaps. We then create simulated datasets using these parameter values, use the Markov Link Method to produce its estimator, and see and measure the total variation error of the estimator. This error ranges from zero to one, where zero indicates no error and one indicates complete disagreement. In subfigure (b) we see that the MLM is able to estimate the ground truth with negligible error, as long as it has enough samples. However, we also consider another simulation: we merge some of the subpopulations in the ground truth parameters so that they become indistingiushable. In this degraded setup, the MLM is unable to learn the true calibration no matter how many samples we have.

For the first simulation, we will look at a case in which the MLM performs well. We will assume there are twenty different samping strategies, and tools I and II both measure specimens as being in one of six categories. Our particular choices of ground truth can be found in Figure 3. This figure also shows that the MLM is able to accurately estimate the ground truth calibration q^*, as long as there are enough samples.

We also used this simulation to test the coverage of the confidence region. Recall that the 'actual coverage' of a confidence interval indicates the probability that a confidence region contains the ground truth parameters. This coverage depends upon the particular choice of ground truth. The 'nominal' coverage indicates the intended coverage of the interval, and may be more or less than the actual coverage. We estimated the actual coverage for the ground truth we have chosen. Taking 200 sample datasets with $s = 80$ samples per subpopulation, we found that an MLM confidence region with 95% nominal coverage has an actual coverage of about 99%. In this sense our region is 'conservative,' which was somewhat by design. We refer the reader to Appendix A for details. We also investigated other values of s, finding that the model became less conservative with more samples. For example, for datasets $s = 800$ with the actual coverage appears to be approximately 97.5%.

In our second simulation, we consider a case where the MLM does not succeed in accurately estimating the calibration. We use the same ground truth parameters as in first simulation, but we will merge groups of sampling strategies. In particular, we will divide the sampling strategies into four groups, each with five members. We will pretend that we cannot distinguish subpopulations within any given group. In this case, Figure 3 shows that the Markov Link Method cannot accurately estimate the calibration, no matter how many samples we have. This is due to a fundamental identifiability issue which is discussed in detail in Appendix D.

Despite the fact that we could not get an accurate point estimate for the calibration in the second simulation, the confidence regions appear to remain on solid ground. We took 200 simulated datasets with $s = 400$ samples per subpopulation (1600 samples total) and computed MLM confidence regions with 95% nominal coverage. We found an actual coverage of about 96.5%.

For our final simulation, we will dig deeper into the question of when the calibration can be perfectly estimated with enough data. To this end, we will consider cases in which the number of subpopulations is four, and there are six possible outcomes for both measurement tools. For many choices of p^*, q^*, this will make it *impossible* to correctly determine the calibration q^* without joint measurement. However, for certain values of q^* the truth can be recovered. To see this, we will look at a variety of calibrations. On the one hand, we will consider the case that X, Y are independent, i.e. $q^*(y|x) = q^*(y|x')$ for every x, x'. On the other hand, we will consider the case that X, Y are deterministically related, in particular $X = Y$. We will also consider every calibration "in-between" these two extremes. These in-between calibrations are found by linear interpolation. For each calibration we will generate datasets of various sizes and try to estimate the calibration from the datasets using the MLM. In Figure 4, we see that in the two extreme cases it may be possible to recover the calibration. However, for in-between cases it is indeed not possible. This is true even in the limit of infinite data. We leave a complete mathematical understanding of this result to future work. For the purposes of this paper, we content ourselves by noting a special case of particular interest: that X and Y can take on one of 2^k values and are related to each other by an invertible deterministic function. In this case, with enough enough samples of only $k+1$ different sampling strategies, we can determine both that the relationship is deterministic and the exact specification of the invertible function. This result is proven in Appendix D, Theorem 1.

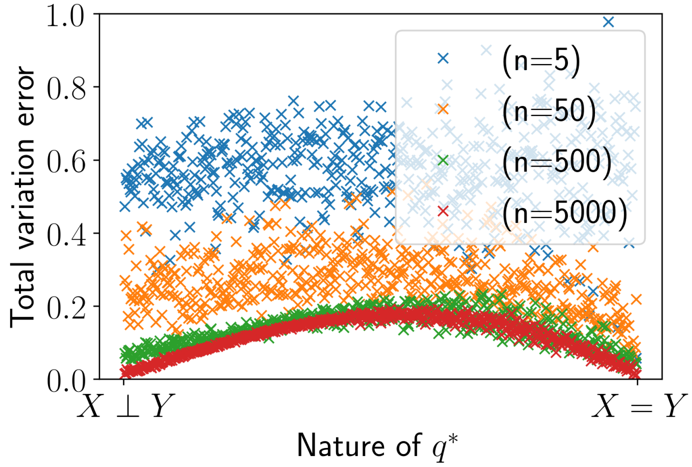


Figure 4: The error of the MLM estimator depends on the nature of q^*. In our third simulation we fix a particular p^* and consider a range of q^*s. For each possible q^* we generate a certain number of samples per subpopulation and then try to use the resulting dataset to estimate q^*. On the one hand, we consider the case that q^* causes X, Y to be independent; in this case we can get a good estimates of q with enough data. On the other hand, we consider the case that q causes X, Y to be deterministically bound, i.e. $q^*(y|x) = 1$ if $x = y$ and $q(y|x) = 0$ otherwise. In this case it is also possible to estimate q with enough data. In-between, it is not possible to exactly determine q^*, even in the asymptotic limit of infinite data.

4 Empirical results for cell-types

We now apply the MLM to real data.

4.1 Background

Our motivation for this problem arose from looking at Allen Institute cell-type assignment of cells, performed using two different experimental tools. Each tool takes a neuron and determines what type of neuron it is. However, both tools destroy the cell in the process of measuring it. The first tool uses a standard single-cell RNA sequencing pipeline to determine transcriptomic expression in the cell, and determine the cell-type based on these results. The second tool also uses transcriptomic information, but additionally obtains electrophysiological and morphological properties of the neuron. This comes at the cost of a degraded transcriptomic signal, requiring new methods to estimate the cell-type. We refer the reader to [6] for details on the two methods.

Best efforts were made to use biological intuition to calibrate the methods. For example, the tool II protocol has a notion of a "Lamp5" cell type. The tool I method refines this idea into many sub-types, such as "Lamp5 Pdlim5" and "Lamp5 Slc35d3." If a cell was designated as "Lamp5 Pdlim5" under the tool I, the hope was that it be given the "Lamp5" type by tool II. 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 sometimes a cell labelled one way with one method would get labelled quite differently 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

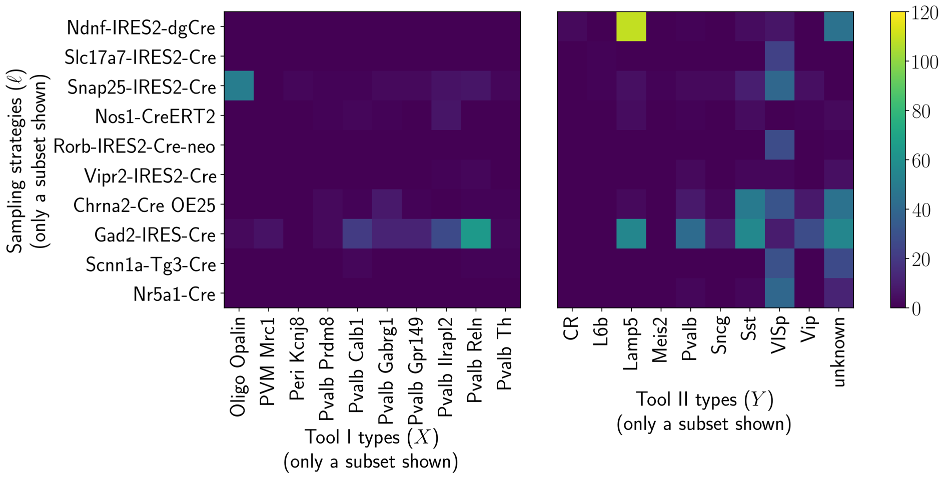


Figure 5: The input to the Markov Link Method: data about each tool with many sampling strategies but without any joint measurement. Here we show a portion of the real-world data to which we applied the MLM. The Allen Institute gathered neurons from the visual cortex of mice, using a variety of cre/lox-based sampling strategies (cf. [6]). Each strategy was designed to bring out different subpopulations of cells. For each sampling strategy, one classification protocol (tool I) was used to estimate the some of the neurons' cell-types, and another protocol (tool II) was used to estimate the other neurons' cell-types. Tool I recognizes 104 different types of cells. Tool II has a coarser notion of cell-type, distinguishing only 10 types. For each sampling strategy and for each tool, we tabulated the number of cells assigned to each type. Above we show a subset of these results through heatmaps; the color of each square indicates the number of specimens sampled with a particular technique that were found to have a particular type according to a particular tool. Using this kind of data, our task is to calibrate the two classification protocols. That is, we want to be able to ask question of the following form: "if a neuron is classified as being of type 'Peri Kcnj8' by tool I, how might it have been classified by tool II?"

whether the two methods were properly calibrated: a variety of sampling strategies. Using a cre/lox system (cf. [6]) 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. The result of this process was two tables, parts of which are shown in Figure 5. While it seemed clear that these tables should say something about the calibration, it was not obvious how to best use this information. It was for this purpose that the MLM was developed.

4.2 Results

As described in Section 1, to apply the MLM we first produced an approximate 95% confidence region for p^*, q^*. We found that this confidence region was non-empty, signifying that we could not reject the hypothesis that the MLM assumption holds. We then used the Markov Link Method of producing a variety of example calibrations inside the confidence region. We then visualized these examples one cell-type at a time. That is, we first select a tool-I cell-type x (such as $x =$'Lamp5 Egln3 1'). Then for any calibration q, we can consider the vector q_x formed by looking at $q(y|x)$ for every value of y. We can then visualize many calibrations by plotting many of stacking these vectors together into a heatmap. The result for various choices of X may be found in Figure 6.

Inspecting these visualizations we can see different aspects of our knowledge about how well-calibrated the two tools are:

- For some tool-I types, such as 'Lamp5 Pdlim5,' the two methods appear very well-calibrated. All the examples from the confidence region suggest that cells measured with type 'Lamp5 Pdlim5' are likely to be assigned type 'Lamp5' by tool II.
- For some types, such as 'Lamp5 Egln 1,' we simply have no idea. There is tremendous variability within the confidence region; each example looks different.
- For some types, such as 'Vip Crispld2 2,' it appears that cells of this type are difficult for tool II to understand. In particular, for most example calibration they have a reasonably high probability of being assigned the 'unknown' type. However, we cannot be sure of the the exact depth of this issue, because in some plausible example calibrations it appears that these cells are in fact correctly labelled as being of the 'Vip' type fairly often. More data is necessary.

The variability in the confidence regions suggests what we need to do in order to more closely determine the value of the calibration. For example, if we could develop more unique sampling strategies which will include 'Lamp5 Egln 1' cells, this would help us resolve our ambiguity about this aspect of the calibration. Indeed, going back to the original data, it is easy to see why this ambiguity appeared in the first place. Cells of the 'Lamp5 Egln 1' type only appear in any number when using the sampling strategies 'Gad2-IRES-Cre' and 'Slc32a1-IRES-Cre.' Both of those sampling strategies yield a fairly similar mix of types when measured with tool II. To get a better resolution of the calibration, we would need a sampling strategy that included 'Lamp5 Egln 1' cells but represents a significantly different slice of the overal population. For a particular proposed experiment, simulations such as those found in Section 3 can be used to determine how many samples might be required to get an accurate estimate of the calibration.

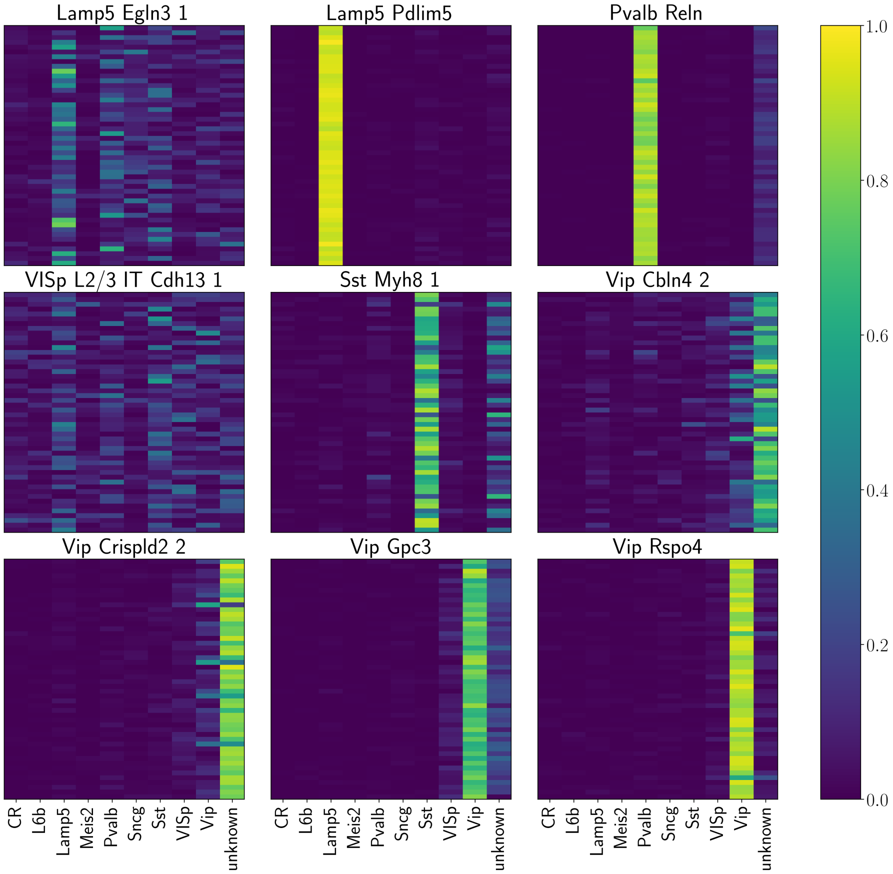


Figure 6: Estimating the calibration and measuring our uncertainty on real data. Here we present visualizations of the confidence region found by the Markov Link Method applied to Allen Institute data. The colors indicates the value of $q(y|x)$ for a particular cell-types x, y and a particular calibration q. All of the calibrations are examples from a 95% confidence interval. For each individual plot, x is fixed, each column indicates a particular value of y, and each row indicates a different plausible q. Each plot helps us answer the question: how might specimens labelled with type x by tool I be labelled by tool II? For example, consider the 'Lamp5 Pdlim5' plot. For every calibration (i.e. every row of the plot), we see a bright yellow in the 'Lamp5' column. This gives some confidence that specimens assigned type 'Lamp5 Pdlim5' by tool I will be assigned type 'Lamp5' by tool II. By contrast, consider the 'Lamp5 Egln3 1' plot. This plot shows we have very little knowledge about the calibration for cells identified as this type by tool I. Each calibration in the confidence region gives a completely different picture. The data we have simply isn't sufficient to help us know this aspect of the calibration.

5 Relation to prior work

In this paper we sought to achieve calibration without joint measurement. The main technical obstacle is what is known in the statistics world as an 'identifiability problem.' Due to this problem, we often simply *cannot* directly estimate the calibration that we are interested in. The calibration is simply not "identifiable." Our main contribution is to produce a an approximate confidence region for the parameter of interest.

The necessity to deal with such identifiability issues has become of increasing concern as nature of model parameters and data has grown more intricate. There are many examples in the literature. How much can we hope to learn about the parameters of complex biological reaction networks, if we can only observe part of the process (cf. [15])? Given what we can observe in practice, how much can we ever hope to learn about how cancers divide and spread through the body (cf. [16, 17])? Do the tests we have enable us to actually learn the parameters of our models for how materials bend and break (cf. [18])? At their heart, these are all questions of estimation the face of potential unidentifiability.

This work is inspired by a large literature of examples where assumptions are used to bound potentially unidentifiable parameters. Some of this literature comes from the field of causality. For example, in [19] Bonet produces regions 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 [20]. More generally, the physics literature has contributed many key assumptions that bound unidentifiable parameters (cf. [21], [22], and the references therein). Perhaps the closest work to this one would be [23], 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 closely refine our estimates and by considering the entire space of possible joint distributions instead of simply a particular property of the joint.

6 Discussion

When joint measurement is unavailable, it can be difficult to calibrate two measurement tools.

In this paper we see that a conditional independence assumption can help overcome this difficulty. We proposed a procedure, the Markov Link Method (MLM), to use this assumption to estimate the calibration and measure uncertainty. Code is published at https://github.com/jacksonloper/markov-link-method, including a tutorial-style ipython notebook detailing every computation made in this paper.

To understand our uncertainty, we use to a confidence region; this comes with all the usual attendent caveats. For example, we here take the view that calibrations inside the confidence region are somehow 'plausible' or 'likely.' As has been pointed out numerous time, such a statement cannot be philosophically justified without some at least a hint of Bayesian reasoning (cf. [24]). We leave a more rigorous Bayesian treatment of this problem for future research. The key problem will be to find a suitable prior; this search comes with certain dangers which we are not sure how to circumvent. We give an example of these dangers in Appendix B.

Theory aside, the MLM appears to provide reasonable estimates in simulation and give useful insight to real data. Applied to neuronal cell-type data, the MLM clarified how two different cell-type classification systems are related. We saw that some aspects of the two systems seem well-calibrated, but others we are less sure about. The nature of the variability in the confidence region suggested directions for experiments which would further refine our understanding.

In future work, it may be interesting to consider additional assumptions that might help us estimate calibrations. For example, as it stands the Markov Link Method only works if each tool can only return one of a relatively small number of measurement values. In some cases, we may believe that similar measurement values should have similar probabilities. Such smoothness assumptions would make it possible to apply the MLM to measurement tools which can return many different values.

More generally, it is clear that good assumptions can help us get real insight for tough problems. Even if these assumptions are not sufficient to allow us to perfectly identify our object of interest, confidence regions can often be constructed. By probing these regions carefully, we can learn what the data actually has to say and what experiments will help us 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] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

[4] 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.

[5] 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.

[6] 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.

[7] Nikhil Padmanabhan, David J Schlegel, Douglas P Finkbeiner, JC Barentine, Michael R Blanton, Howard J Brewington, James E Gunn, Michael Harvanek, David W Hogg, Željko Ivezić, et al. An improved photometric calibration of the sloan digital sky survey imaging data. *The Astrophysical Journal*, 674(2):1217, 2008.

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

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[16] Andrew F Brouwer, Rafael Meza, and Marisa C Eisenberg. Parameter estimation for multistage clonal expansion models from cancer incidence data: A practical identifiability analysis. *PLoS computational biology*, 13(3):e1005431, 2017.

[17] E Georg Luebeck and Suresh H Moolgavkar. Multistage carcinogenesis and the incidence of colorectal cancer. *Proceedings of the National Academy of Sciences*, 99(23):15095–15100, 2002.

[18] Stefan Hartmann and Rose Rogin Gilbert. Identifiability of material parameters in solid mechanics. *Archive of Applied Mechanics*, 88(1-2):3–26, 2018.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] Richard D Morey, Rink Hoekstra, Jeffrey N Rouder, Michael D Lee, and Eric-Jan Wagenmakers. The fallacy of placing confidence in confidence intervals. *Psychonomic bulletin & review*, 23(1):103–123, 2016.

[25] Alessandra R Brazzale, Anthony Christopher Davison, Nancy Reid, et al. *Applied asymptotics: case studies in small-sample statistics*, volume 23. Cambridge University Press, 2007.

A Confidence interval details

Recall that we have an $\Omega_\ell \times \Omega_X$ matrix N^X and an $\Omega_\ell \times \Omega_Y$ matrix N^Y, sampled according to

$$(N^X_{\ell 1}, N^X_{\ell 2} \cdots N^X_{\ell \Omega_X}) \sim \text{Multinomial}(n_\ell, p^*(\cdot|\ell))$$
$$(N^Y_{\ell 1}, N^Y_{\ell 2} \cdots N^Y_{\ell \Omega_Y}) \sim \text{Multinomial}(m_\ell, h^*(\cdot|\ell))$$

where $p^*(x|\ell)$ is some conditional distribution and there is some $q^*(y|x)$ such that $h^*(y|\ell) = \sum_x p^*(x|\ell)q^*(y|x)$. We have then defined

$$L(p,q) = \frac{1}{\Omega_X} \sum_{\ell,x} N^X_{\ell x} \log \frac{N^X_{\ell x}}{n_\ell p(x|\ell)} + \frac{1}{\Omega_Y} \sum_{\ell,y} N^Y_{\ell y} \log \frac{N^Y_{\ell y}}{m_\ell \sum_x p(x|\ell) q(y|x)}$$

We then constructed the confidence region

$$R = \{p,q : \ L(p,q) \le k\}$$

This is sometimes called so-called likelihood ratio confidence interval, and its properties are fairly well-understood. We refer the reader to [25] for a useful exposition. The essential point is that in the asymptotic limit the distribution of $L(p^*, q^*)$ depends only upon the number of zero probability values in the parameters p^*, h^*. For any number of zero probability values, the asymptotic distributions are known exactly; they are scaled sums of χ^2 distributions, one for each subpopulation and each tool. The difficulty is of course that we do not know how many probability value are zero. This can can be circumvented if one is only interested in a conservative confidence region, i.e. one whose actual coverage is at least as great as its nominal coverage. In particular, if we assume that *no* values are zero we can a confidence interval which is guaranteed to be conservative; this is an immediate result of the fact that χ^2 distributions with more degrees of freedom stochastically dominate those with fewer. Unfortunately, using these χ^2 distributions can be quite inadvisble in low-sample regimes. For instance, one may readily see that if some values of p^*, h^* are small (but nonzero), then the distribution of $L(p^*, q^*)$ will be much more similar to sums of χ^2 distributions with fewer degrees of freedom – at least until the number of samples grows very very large.

We find it practical and accurate to choose k by using the data itself. In particular, we choose a pseudocount c and take $\bar{p}(y|x) = N^X_{\ell x} + c/n_\ell + c$ and $\bar{h}(y|\ell) = N^Y_{\ell y} + c/m_\ell + c$ to give estimates of p^*, h^*. For a nominal coverage of $1 - \alpha$, we then choose k so that

$$\mathbb{P}\left(\frac{1}{\Omega_X} \sum_{\ell,x} N^X_{\ell x} \log \frac{N^X_{\ell x}}{n_\ell \bar{p}(x|\ell)} + \frac{1}{\Omega_Y} \sum_{\ell,y} N^Y_{\ell y} \log \frac{N^Y_{\ell y}}{m_\ell \bar{h}(y|\ell)} < k \right) = \alpha$$

when N^X, N^Y are drawn from $\bar{p}, \bar{h}$ respectively. Asymptotically $\bar{p} \to p^*$ and $\bar{h} \to h^*$, so it seems reasonable to hope that this will give the correct coverage. Simulations appear to bear this out, though a further mathematical inquiry is certainly warranted.

The pseudocount bears some discussion. We introduce this pseudocount in order to err on the side of a slightly conservative confidence region. As we said before, when the parameters p^*, h^* have many small values it is generally possible to produce a smaller confidence region for any given level of coverage. We therefore apply a pseudocount to give a slight boost to events that appear to have small probability. In practice we find that this encourages the confidence region to err on the side of being conservative, giving us greater confidence that the true parameter lies within the confience region.

B The Bayesian Alternative

Todo

C Convex optimization problems

Todo

D Identifiability case studies

Our analysis of the identifiability issues associated with the Markov Link Method is not completely satisfactory. We are able to make rigorous guarantees about consistency for any given true value of p^*, q^*, but we are not content with our understanding of the situations in which identifiability is and is not a serious problem.

Here we present two suggestive case studies which we hope may inspire future research. In both cases we are able to prove something of interest – but not quite as much as we might hope.

D.1 Permutation matrices

Let us consider the case that $\Omega_\ell = \{1, 2, \cdots k\}$ and $\Omega_X = \Omega_Y = \{1, 2 \cdots, 2^{k-1}\}$. That is, there are k separate subpopulations, and both tool I and tool II can return one of 2^{k-1} possible values. Let us furthermore assume that

$$q^*(y|x) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{else} \end{cases}$$

and $p^*(x|\ell) = A_{\ell,x}$, where this matrix is given by

$$A = 2^{2-k} \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & \cdots & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & \cdots & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & \cdots & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \\ & & & & & & \vdots & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & \cdots & 1 & 0 & 1 & 0 \end{pmatrix}$$

That is, the xth column of the first $k-1$ rows is simply the binary expansion of the number $x-1$, and the last row alternates 1s and 0s. Now let us say we have perfect knowledge of $p^*(x|\ell)$ and $h^*(y|\ell) = \sum_x p^*(y|\ell)q^*(y|x)$. Notice that due to the simple structure of q^* we obtain $h^*(y|\ell) = A^{\ell,y}$. However, let us imagine we know nothing about the true value of q^*.

How much can we say about q^*, if we only had knowledge of p^* and h^*? On the one hand, we observe that in the absence of any other constraints, the object q^* has 2^{2k-3} degrees of freedom. This is because there are 2^{k-1} values of ℓ and for each subpopulation $q^*(\cdot|\ell)$ must lie in the 2^{k-2}-dimensional simplex on 2^{k-1} atoms. On the other hand, we see that the Markov Link Assumption gives us $k \times (2^{k-1} - 1)$ linear constraints on the value of q. Indeed, for each subpopulation in $1\cdots k$ and each value of $y \in 1\cdots 2^{k-1}$ we have an equation of the form

$$\sum_x p(y|\ell)q(y|x) = h(y|\ell)$$

Of these $k \times 2^{k-1}$ constraints, k of them are redundant with the fact that $\sum_y q(y|x) = 1$. Thus, altogether, the Markov Link Assumption together with approximate knowledge of p,h gives us $k \times (2^{k-1}-1)$ linear constraints. It would follow that q would have $2^{2k-3} - k \times (2^{k-1}-1)$ degrees of freedom yet remaining.

In conclusion, a simple degrees-of-freedom counting argument would suggest that there will be substantial ambiguity about what value q might take on, if our only knowledge about q is that it must satisfy $\sum_x p^*(y|\ell)q(y|x) = h^*(y|\ell)$. Indeed, we have *exponentially many* more degrees of freedom than we have constraints.

However, the reality is that q is exactly determined by p^*, h^*. This is possible because there are inequality constraints which also govern q, namely $q(y|x) \geq 0$. Thus, while a simple degrees-of-freedom counting argument might suggest that we would have substantial identifiability issues in this problem, the reality is quite the opposite. This idea is made rigorous in the following theorem.

Theorem 1. *Let p^*, h^* be as they are defined above. Then there is exactly one q that is consistent with p^*, h^* and the Markov Link assumption. That is, q^* is the only possible q satisfying*

$$\sum_y q(y|x) = 1$$

$$\sum_x A_{\ell,x} q(y|x) = A_{\ell,y}$$

$$q(y|x) \geq 0$$

Proof. We prove by recursion. First take the case $k = 2$. In this case the result holds trivially, since $X, Y \in \{1\}$.

Now consider a general case $k > 2$. Let us focus on the constraints implied by the second-to-last row population. It is straightforward to see that these constraints imply

$$0 = q(y|x) \qquad \forall y \leq 2^{k-2}, x > 2^{k-2}$$

Indeed, for each $y \leq 2^{k-2}$ we obtain a constraint showing that $\sum_{x > 2^{k-2}} q(y|x) = 0$, which yields that in fact $q(y|x) = 0$ for every $x > 2^{k-2}$ and every $y \leq 2^{k-2}$.

It follows that for $y \leq 2^{k-2}$ our original constraints may be rewritten as

$$\sum_{x \leq 2^{k-2}} A_{\ell x} q(x|y) = A_{\ell y} \qquad \forall y \leq 2^{k-2}$$

This is an example of our problem with k one smaller. Applying our inductive hypothesis, we may thus obtain that $q(y|x) = q^*(y|x)$ for the first 2^{k-2} values of x, y. Moreover, since $\sum_{y \leq 2^{k-2}} q^*(y|x) = 1$, we see that q must also satisfy $q(y|x) = 0$ for $y > 2^{k-2}$ and $x \leq 2^{k-2}$. Thus we have seen that $q = q^*$ for all entries except those in which $x, y \geq 2^{k-2}$.

For $x, y \geq 2^{k-2}$ we linearly combine equations concerning the first, last, and second to last rows of A with factors of $1, 1, -1$ respectively. We obtain constraints showing that $\sum_{x \leq 2^{k-2}} q(y|x) = 0$ for each $y > 2^{k-2}$. We can then use the same reasoning to obtain that $q = q^*$ for the remaining values of x, y. □

This result is somewhat robust to slight perturbations in p^*, h^*. In particular, if we have some $\hat{p} \approx p^*$ and $\hat{h} \approx h^*$ then at each stage of the argument we can replace statements of the form $q(y|x) = 0$ with statements of the form $q(y|x) \leq \epsilon$. Applying this with the kinds of arguments above will show that we can be sure that $\hat{q}$ is arbitrarily close to q^* if we know that $\hat{p}, \hat{h}$ are sufficiently close to p^*, h^*.

However, it turns out that the relationship between q and p^*, h^* is not robust in every situation. In the next section we will see that it can in fact be quite discontinuous:

D.2 A discontinuity

Let us consider the case that $\Omega_\ell = \{1\}$ and $\Omega_X = \Omega_Y = \{1, 2\}$. That is, there is only one population (no subpopulations) and both tool I and tool II can return one of 2 possible values. We will now consider two possiblities:

1. First let us take the case

   - $\mathbb{P}(X = 1) = p^*(1) = 0$
   - $\mathbb{P}(X = 2) = p^*(2) = 1$
   - $\mathbb{P}(Y = 1) = h^*(1) = 0$
   - $\mathbb{P}(Y = 2) = h^*(2) = 1$

   In this case the MLM assumption $\sum_x p^*(x)q(y|x) = h^*(y)$ can be used to prove that $q(1|2) = 0, q(2|2) = 1$, but we now have *absolutely no* knowledge of $q(1|1), q(2|1)$. This is because we simply never observed the case $X = 1$ (it occurs with probability zero), and so we cannot possibly have any knowledge about $q(y|x)$ for $x = 1$.

2. Now let us take a slight variation:

   - $\mathbb{P}(X = 1) = p^*(1) = 0.01$
   - $\mathbb{P}(X = 2) = p^*(2) = 0.99$
   - $\mathbb{P}(Y = 1) = h^*(1) = 0$
   - $\mathbb{P}(Y = 2) = h^*(2) = 1$

   In this case we can again prove that $q(1|2) = 0, q(2|2) = 1$, but we can also prove that $q(1|1) = 0, q(2|1) = 1$.

3. Now we take yet another slight variation:

   - $\mathbb{P}(X = 1) = p^*(1) = 0.01$
   - $\mathbb{P}(X = 2) = p^*(2) = 0.99$
   - $\mathbb{P}(Y = 1) = h^*(1) = 0.01$
   - $\mathbb{P}(Y = 2) = h^*(2) = 0.99$

   In this case we can prove that $q(1|2) \leq 1/99$ and $q(2|1) \geq 1 - 1/99$, but we again cannot prove almost anything about $q(2|1)$. In particular, it is easy to produce cases in which $q(2|1) = 0$ and other cases in which $q(2|1) = 1$.

The disturbing thing about this example is that by making infinitesimal perturbations to p^* we can pass from uncertainty to complete certainty back to uncertainty. It is for this reason that in this paper we refuse to ever treat $\hat{p}, \hat{h}$ as fixed and given, always considering the space of perturbations around any such values.

It is worth noting that these kinds of problems essentially vanish if the true q^* is bounded away from zero i.e. $q^*(y|x) > c$ for every x, y for some $c > 0$. If this holds, together with a certain linear independence assumption, we can guarantee that the true q^* is close to the set $\Theta(\hat{p},\hat{h})$ when $\hat{p},\hat{h}$ are good approximations to p^*,h^* (recall that we defined Θ in Equation (**??**)). In particular:

Theorem 2. *Fix any q^* satisfying $q^*(y|x) > c$ for some $c > 0$. Let us further assume that the matrix $B^*_{\ell,x} = p^*(x|\ell)$ has linearly independent rows. Fix any p^* and define $h^*(y|\ell) = \sum_x p^*(x|\ell)q^*(y|x)$. Then by taking any $\hat{p},\hat{h}$ sufficiently close to p^*,h^* we can ensure that q^* is arbitrarily close to some point in the set*

$$\hat{\Theta} \triangleq \left\{ q \in T : \sum_x \hat{p}(x|\ell)q(y|x) = \hat{h}(y|\ell) \quad \forall \ell, y \right\} \tag{1}$$

where by T we mean the transition matrix polytope, $T = \{q : q(y|x) \geq 0, \sum_y q(y|x) = 1\}$.

Proof. Let $\hat{A}$ denote the affine plane $\hat{A} = \{q : \sum_x \hat{p}(x|\ell)q(y|x) = \hat{h}(y|\ell)\}$. Thus $\hat{\Theta} = T \cap \hat{A}$.

Now fix any $\hat{p},\hat{h}$. Now let $\tilde{q}$ be the orthogonal Euclidean projection of q^* to the affine space A. That is, $\tilde{q}$ is minimizes a sum-of-squares difference to q^* among all the points in A. The linearly independent rows of B^* allow us to bound the spectral norm of the right-pseudoinverse of $\hat{B}_{\ell,x} = \hat{p}(x|\ell)$, by taking $\hat{p}$ sufficiently close to p^*. If we furthermore require $\hat{h}$ sufficiently close to h^*, we can use this to ensure the projection distance is small. That is, we can force $\tilde{q}$ to be arbitrarily close to q^*. Using the fact that $q^*(y|x) > c$ we can thereby furthermore insure that $\tilde{q}(y|x) \geq 0$. Finally, it is straightforward to see that the projection leaves the constraint $\sum_y q(y|x) = 1$ unchanged. Thus $\tilde{q} \in \hat{\Theta}$ and $\tilde{q}$ is arbitrarily close to q^*. □

This theorem is certainly a step in the right direction, but we emphasize that it does require some assumptions. It is our opinion that the linear independence requirement is fairly mild (if it is not met then subpopulations can simply be merged together). However, the positivity requirement is quite troubling. In many cases of interest the true calibration q^* has genuine zeros: pairs of measurements between the two tools which are fundamentally incompatible. In this case such a theorem cannot be applied.

However, it may be that the above theorem's requirement ($q(x|y) > c > 0$) is much stronger than is necessary. A precise understanding of the discontinuity example above has remained elusive. Better understanding could lead to a much more accurate and practical estimates. We leave it for future work.