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 2, 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 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.

1 The Markov Link Method assumption

The MLM is only applicable if a certain assumption is met. In this section we will articulate this assumption as clearly as possible. 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 we can identify a variety of these subpopulations. They are not required to be disjoint subpopulations; it is only necessary that each subpopulation captures a different slice of the overall population. 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 measuring 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. We would like to estimate as much as possible about the distribution of $X, Y|\ell$. In this paper, we will consider the case that this estimation is facilitated by an additional assumption:

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. In the language of statistics, the assumption is that the variable X is 'sufficient' for ℓ. If this assumption is not met, then the method presented here is not applicable. If it *is* met, it opens the door for new methods to estimate three important objects:

- The calibration $q^*(y|x)$
- The distribution of tool-I measurement values found in each subpopulation, $p^*(x|\ell)$
- The distribution of tool-II measurement values found in each subpopulation, $h^*(y|\ell)$.

Under the MLM assumption, it is straightforward to see that these three objects are all bound together by the formula $h^*(y|\ell) = \sum_x p^*(x|\ell)q^*(y|x)$ for each ℓ, y. Using this equation, we will see that it is possible to estimate properties of the calibration q^* even without access to joint measurement. It also makes it possible to perform more efficient estimation of h^*. Theoretically we could also get more efficient estimation of p^*, but we did not find plausible cases where this actually occured.

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

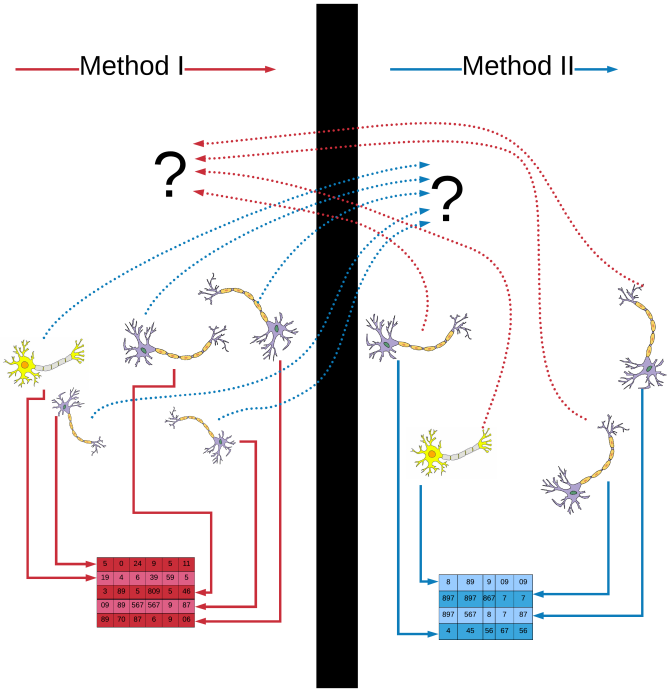


Figure 1: We consider the case that for each specimen there are two ways we might measure it. We divide the specimens into two groups. The specimens on the left are measured with technique I and the specimens on the right are measured with technique II. In the end, we are left wondering: what would have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

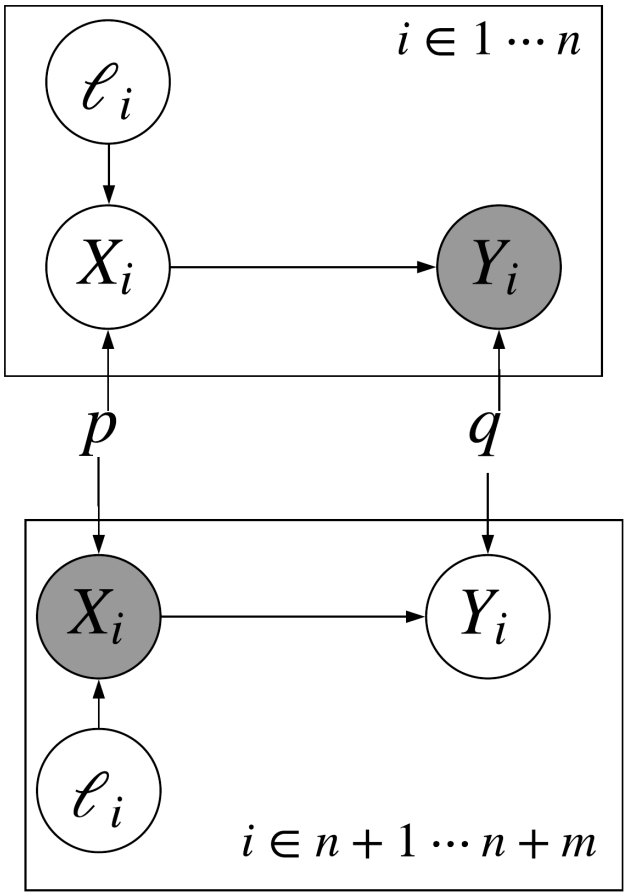


Figure 2: Another way to articulate the Markov Link Method is to look at the graphical model associated with the MLM assumption. This reveals that we have two kinds of data. In one case (above), we can observe ℓ, X, but Y is hidden. For other kinds of data (below) we can observe ℓ, Y, but X is hidden. All of this data is joined by the unknown parameters p, q. Here p indicates the distribution of tool-I measurement values from each subpopulation, and q indicates the calibration. Our task is to use these different kinds of data to determine the parameters p, q as best as we can. Once these parameters are known, notice that distribution of tool-II measurement values for each subpopulation, which we denote h, can be determined immediately using the Markov Link Assumption: $h^*(y|\ell) = \sum_x p^*(x|\ell)q^*(y|x)$.

the reliability of the machine that performs the test? In each test run there will be some variability induced by the machine itself, which induces a measurement error. In practice, some kind of assumptions about part homogeneity are used to approximate this error (cf. [2]). However, if we have two testing machines we can use the MLM to obtain a calibration between the machines, even though we can never test the same part with both machines. This enables us to bound the overall measurement error. In this case, 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. [3, 4]). A more modern approach is to assay the cell's transcriptome (cf. [5]). Unfortunately, modern high-resolution cell photography and single-cell sequencing technologies are both destructive. As a result, we can't 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.

- 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. [6]). 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. [7]). However, we know that

cultured cell-lines look quite different from in-vivo cells (cf. [8, 9]). 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. [10]). 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 [11]. 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. [12, 13]), 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 The Markov Link Method: input and output

The Markov Link Method is a procedure from estimating the distribution $X, Y|\ell$ from data, under the Markov Link Method assumption. Details on the exact algorithm and convergence results are given in Section 7. For now, let us content ourselves with looking at the input and the output of the MLM:

Input We assume we are given n samples using tool I and m samples using tool II.

- Samples $(\ell_1, X_1) \cdots (\ell_n, X_n)$ sampled such that $\mathbb{P}(X_i = x|\ell_i = \ell) = p^*(x|\ell)$

- Samples $(\ell_{n+1}, Y_{n+1}) \cdots (\ell_{n+m}, Y_{n+m})$ sampled such that $\mathbb{P}(Y_i = x | \ell_i = \ell) = h^*(x|\ell)$

We assume that ℓ, X, Y are discrete random variables with finite support. We also assume that the MLM assumption holds, i.e. $h^*(x|\ell) = \sum_y p^*(x|\ell) q^*(x|y)$ for some unknown q^*. It would be interesting to generalize the method to non-discrete random variables, but we leave this for future work. Note that if the random variables are low-dimensional and continuous, the MLM can still be applied by discretizing the continuous space into a finite number of bins.

Output Our goal is to estimate p^*, q^*, h^* from the samples. In this effort, we return:

- Point estimates, $\hat{p}, \hat{q}, \hat{h}$.
- Quantitative estimates of the variability of these point estimates due to insufficient data, $\Delta_p, \Delta_q, \Delta_h$.
- An quantitative estimate of the variability of $\hat{q}$ due to so-called 'identifiability' issues, $\bar{d}$.
- Variability visualizations indicating a set of "plausible" values for each row of q; this gives a qualitative counterpart to the quantitative measures of variability.

Precise definitions and mathematical results are given in section 7.

4 Simulation results

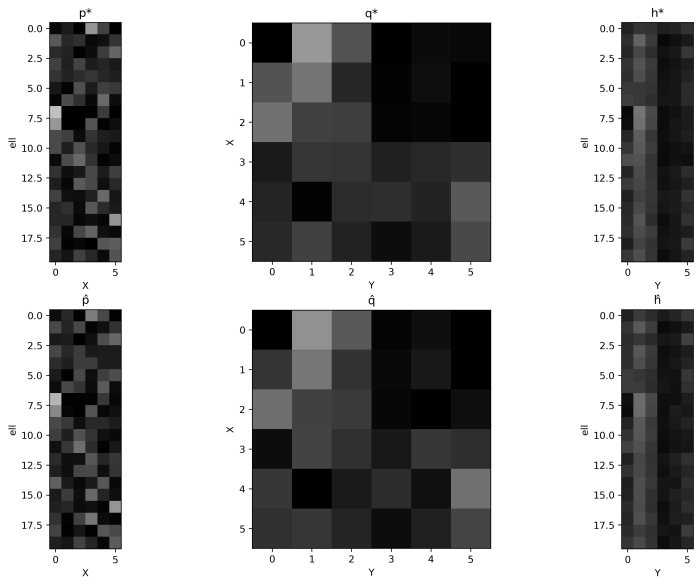
Before we show the results of the MLM on real data, let us look at some simulation results to see what the MLM can and cannot achieve. In particular, we will:

- Pick a set of possible parameter values for p^*, q^*, h^*.
- Generate data according to these distributions, namely n samples according to p^* and m samples according to h^*.
- Pretend to forget the true values of p^*, q^*, h^*.
- Apply the Markov Link Method to the generated samples.
- Compare the results of the MLM with the original true values.

Our purpose is to see how well the method performs if all of its assumptions are met. We will measure performance two ways:

1. First, we will measure the error of the estimators $\hat{p}, \hat{q}, \hat{h}$, i.e. how much they differ from the true generating distributions p^*, q^*, h^*. For h we will compare this measurement error against a standard pseudocount estimator (we will not do the same for p, because the MLM estimator for p *is* a pseudocount estimator). We refer the reader to Section 7 for details on this estimator.
2. Second, we will look at how well the MLM has estimated its uncertainty about the estimates. We know that there will always be some amount of error. For this reason, the MLM offers two tools to estimate the amount of error that may be present. In this section we will particularly focus on the quantities $\Delta_p, \Delta_q, \Delta_h$ (which approximately indicate variability due to insufficient data) and $\bar{d}$ (which approximately indicates variability due to so-called identifiability concerns). For a precise explanation of these two sources of variability, we refer the reader to Section 7.

For our first simulation, we will look at a case in which the MLM performs well. We consider twenty different subpopulations (which we will label $\{1, 2, 3, \cdots 20\}$). We will assume



Variable	MLM error	Pseudocount error	Δ
p	1.5	N/A	1.6
q	0.6	N/A	0.6
h	0.9	1.4	1.0

Figure 3: For our first simulation, we imagine that the true values of p^*, q^*, h^* are indicated by the first row of heatmaps above. The color of each entry indicates a probability between 0 and 1, with 0 being represented by black and 1 being represented by white. For example, the entry at the third row and the fifth column of the q plot indicates $\mathbb{P}(Y = 5|X = 3) = q^*(5|3)$. The second row of heatmaps indicates the estimates for these quantities returned by the MLM. We can also look at this more quantitatively, as shown in the table. For each variable, p, q, h, we can compute the total variation distance between our estimates and the truth. For h we can also compare the error we get when using a more standard pseudocount estimator. Finally, in the right-most column of the table, we show the Δ variability measures for each variable. These are intended to roughly indicate the level of error we would expect due to insufficient data. They are caclulated using a bootstrap method. We can see that the MLM variability measured do seem to give a good indication of the actual error level.

that tool I always returns an integer measurement in $\{1, 2, 3, 4, 5, 6\}$ and tool II returns a measurement in $\{1, 2, 3, 4, 5, 6\}$. We will assume we have one hundred samples for each subpopulation for each method. Some results of this simulation can be seen in Figure 3. This figure shows that the MLM gives us an edge in estimating h, performing better than a traditional pseudocount estimator. It also allows us to estimate q with decent accuracy (which is only possible because of the Markov Link Assumption). The figure also shows that the Δ variability measures seem to accurately predict the estimation error. The $\bar{d}$ for this simulation was identically zero, because there were no identifiability issues in this case; this is related to the fact that there were more subpopulations than there were possible values that tool I could return.

For our second simulation, we will look at a case where the MLM does a good job of estimating p, h, but does not succeed in accurately estimating q. We use the same data as in the previous simulation, but we will merge groups of subpopulations into four larger subpopulations. This will make it easier to estimate $h(y|\ell), p(x|\ell)$, since we have fewer

500 samples per population:

Variable	MLM error	Δ
p	0.12	0.16
q	1.33	1.19
h	0.17	0.15

10^{10} samples per population:

Variable	MLM error	Δ
p	10^{-7}	10^{-3}
q	0.79	0.01
h	10^{-3}	10^{-3}

Figure 4: Our second simulation is similar to our first. However, instead of twenty subpopulations, we will assume we only have four subpopulations with five times the amount of data for each subpopulation. This makes estimating p, h easier, but it doesn't make estimating q any easier. Even when we give ourselves a virtually unlimited number of samples, perfect estimation of q is just not possible. The Δ measures of the MLM do *not* accurately reflect this fact. It is for this reason that the MLM additionally provide the quantity $\bar{d}$. For this simulation $\bar{d} = 3.5$, which allows us to know that there are identifiability concerns, and also to bound the magnitude of those concerns.

subpopulations and for each one we will have five times as much data. The results are shown in Figure 4. As we would expect, estimation error of p, h improves. However, estimation of q does not improve at all. In fact, it is considerably worse. To understand this, we run a further simulation with the same true distributions but a virtually infinite number of samples. Even in this case, we see that accurate estimation of q is still not achieved; we end up with a final error of 0.79. This significant error is due to a fundamental identifiability issue which is discussed in detail in Section 7. Accurate estimation of q is simply impossible in this case, regardless of the number of samples. This is related to the fact that the number of subpopulations is smaller than the number of different values that tool I can report. What is worse, the Δ_q reports only .01 for the predicted error of $\hat{q}$. It is for this reason we must also look at $\bar{d}$, which approximates the error due to identifiability. For this simulation we calculate that $\bar{d} = 3.5$, which captures the fact that our estimator $\hat{q}$ may suffer from nontrivial identifiability errors. Indeed, it appears to substantially overestimate the total error. This problem is somewhat unavoidable without additional assumptions. A precise statement of the theoretical guarantees regarding the ability of the MLM to estimate its own error due to identifiability issues can be found in Section 7. I

For our next simulation, we will dig deeper into the question of when q can be perfectly estimated. 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. In the general case, this will make it *impossible* to correctly determine 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 values of q. 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 q "in-between" these two extremes. These in-between qs are found by linear interpolation. For each such q we will generate datasets of various sizes and try to estimate q from the datasets using the MLM. The results are to be found in Figure 5. These results show how in the two extreme cases it may be possible to recover q, but 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 with noting that if X and Y can take on one of 2^k values and are related to each other by an invertible deterministic function, then both the fact that the relationship is deterministic and the exact specification of the invertible function can be determined with only $k + 1$ properly chosen

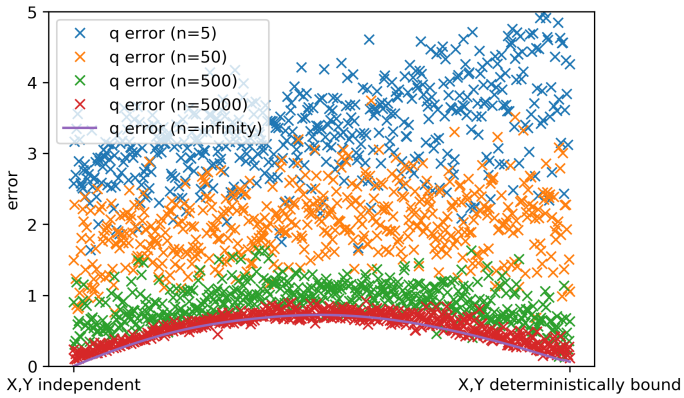


Figure 5: In our third simulation we fix a particular p and consider a range of qs. For each possible q we generate a certain number of samples per subpopulation and then try to use the resulting dataset to estimate q. On the one hand, we consider the case that q causes X, Y to be independent; in thie case we can get a good estimate of q. In some sense this is because $q(y|x) = q(y|x')$ for every x, x', and so there are simply fewer values to estimate. 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 (given enough data). In-between, it is not possible to exactly determine q, even in the asymptotic limit of infinite data.

subpopulations (at least in the asymptotic limit of infinite data). This result is proven in Appendix D.

5 Empirical results for cell-types

We now turn to an application of the MLM to real data.

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, each technique destroyed the cell in the process of measuring it.

Best efforts were made to use biological intuition to calibrate the methods. For example, the Y method has a notion of a "Lamp5" cell type. The X method has refined this type into many sub-types, such as "Lamp5 Pdlim5" and "Lamp5 Slc35d3." If a cell was designated as "Lamp5 Pdlim5" under the X method, the hope was that it be given the "Lamp5" type under the Y 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 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 whether the two methods were properly calibrated: sub-populations. Using a cre/lox system (cf. [5]) they were able to pick out specific, overlapping subpopulations of neurons.

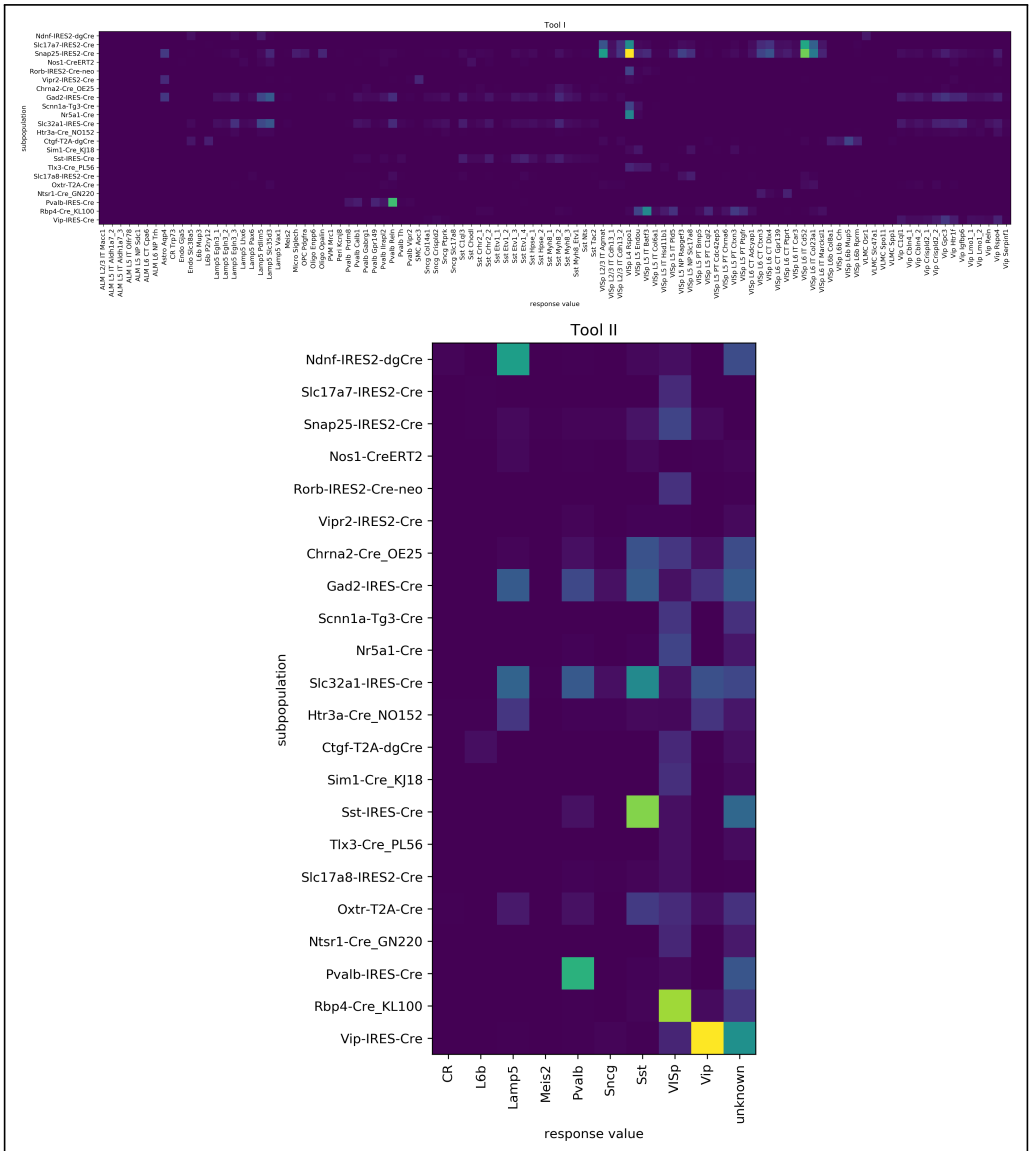


Figure 6: Input: two tables. The Allen Institute had access to various cre/lox cell selection techniques. Each technique pulls out a different group of cells. Once the cells were selected, they were then either subjected to technique I or technique II. Technique I gives a very complete analysis of the 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 results. These are shown above. Each row corresponds to a different subpopulation. Each column corresponds to a different measurement value. Each entry of the heatmaps indicate the number of specimens measured to have a particular measurement value within a particular subpopulation (as measured by the corresponding technique). Notice that tool I has many more different measurement results that it can return; this is why there are many more columns in the first heatmap.

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, shown in Figure 6. 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.

5.2 Results

Using the Markov Link Method, we obtained an estimate $\hat{q}$ for the true calibration. This calibration is shown in Figure 7. From this figure, it appears that the two methods may be fairly well-calibrated. For example, according to $\hat{q}$, if a cell is classified as type "Lamp5 Pdlim5" by tool I, it appears there is a 99.9% chance it will be classified with the "Lamp5" type under tool II. However, there are other cell-types which seem to have more ambiguity. Yet others seem to perhaps be classified incorrectly *most* of the time; we see that "Vip Igfbp6" type is most often classified as a "Pvalb" type instead of a "Vip" type.

Everything in the preceeding paragraph should be taken with a significant grain of salt; it is quite treacherous to directly interpret the estimator $\hat{q}$, because we know that this estimator for the calibration may suffer from errors. This is both due to insufficient data and to identifiability concerns (these are discussed in detail in Section 7). For a qualitative assessment, we instead advocate looking at the variability visualizations which the MLM produces.

For example, in the right side of Figure 8 we show the variability visualization for specimens which are measured as being "Vip Igfbp6" by tool I. Whereas in the original point estimate $\hat{q}$ we found that "Vip Igfbp6" seemed to be strongly associated with the "Pvalb" type, this figure shows that in reality we are not quite sure what types these specimens may be associated with. Each row corresponds to a different plausible distribution on y. As we can see, they truly run the gamut. In the left side of the same figure we also show the same visualization applied to the row of $\hat{q}$ corresponding to specimens measured as "Lamp5 Pdlim5" by tool I. We see very little variability in this visualization, indicating significant confidence that the "Lamp5 Pdlim5" type of tool I is truly well-calibrated to match the "Lamp5" type of tool II.

In cases where the visualization suggests significant variability, the MLM then suggests what subpopulations we should look at in order to resolve the ambiguities. For example, to resolve the ambiguities about "Vip Igfbp6" specimens, it appears it would be most important to obtain samples of subpopulations of cells which have many cells measured as "Pvalb" by technique II and very few cells measured as "Sncg."

The MLM also yields estimates for h^*. As in our simulation section, these quantities could also be estimated using more traditional pseudocount methods. We can compare the MLM estimates with the traditional methods, using the method of log-likelihood on held-out data. We find that both methods achieve a rate of 1.3 nats per entry on held out data. This gives us confidence that the MLM assumption might be right; if the MLM assumption was wrong our estimation process could force h to take on an incorrect value. This is no proof that the MLM assumption is correct for this data, but it is a good sanity-check.

Finally, the MLM provides quantitative values to give a sense for the the error in our estimators, namely $\Delta_p, \Delta_q, \Delta_h, \bar{d}$. These can be difficult to interpret because they are single numerical quantities that summarize error in the whole of p, q, h. As we have seen from our visualizations, certain aspects of q may be fairly tightly identified with very little error whereas others may not be clear at all. Numerical figures must somehow "average" over all these complications. For this reason, the qualitative visualizations may be more helpful.

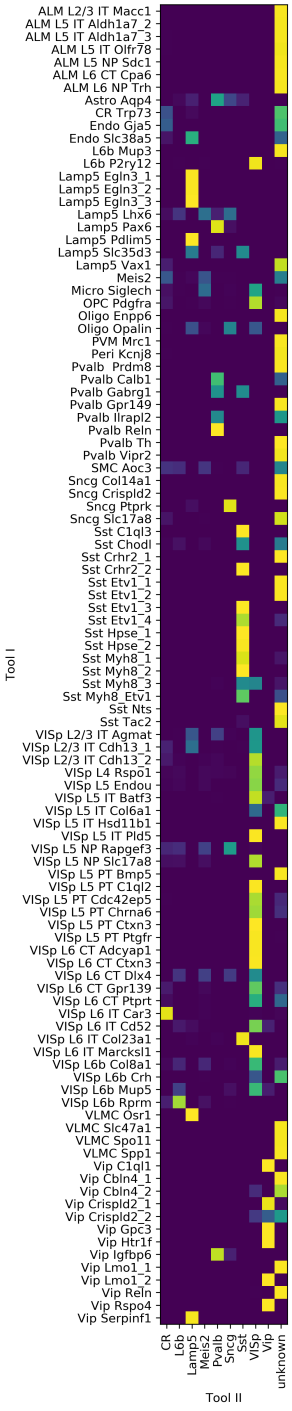


Figure 7: The $\hat{q}$ estimated by the MLM.

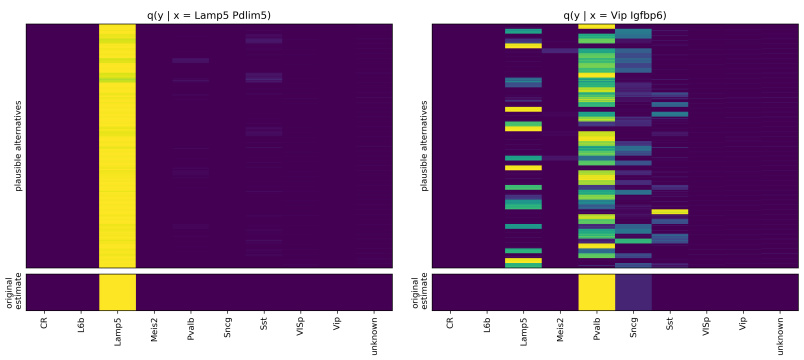


Figure 8: The MLM provides *variability visualizations* for each row of the calibration q. In this figure we show two such examples. To produce the visualization we must first select a row of q. This corresponds to particular measurement value that can be obtained from tool I. In this case we select "Lamp5 Pdlim5" to produce the left visualization and "Vip Igfbp6" to produce the right visualization. The bottom of each visualization shows the value of $\hat{q}(y|x)$ for the selected type x (it shows this through a heatmap with a single row). Here $\hat{q}$ is the Markov Link Method estimator we have described before. Above this original $\hat{q}$ value, we show a heatmap with many other possible examples for what else this row might look like. The variability among these possibilities gives a qualitative measure of our uncertainty about that particular row of the calibration q.

However, if a quantitative measurement is needed, they are readily available: we find that $\Delta_p = 6.7, \Delta_q = 50.3, \Delta_h = 3.7, \bar{d} = 18.0$.

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 gives insensible results, this could actually serve as a useful clue that this disproportionate cell death is happening.

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.

7 The Markov Link Method

Here we describe in detail the Markov Link Method, the identifiability problems which it must overcome, and some theoretical guarantees for the method. For the benefit of the reader, we repeat the basic inputs to the method:

- Samples $(\ell_1, X_1) \cdots (\ell_n, X_n)$ sampled such that $\mathbb{P}(X_i = x | \ell_i = \ell) = p^*(x|\ell)$
- Samples $(\ell_{n+1}, Y_{n+1}) \cdots (\ell_{n+m}, Y_{n+m})$ sampled such that $\mathbb{P}(Y_i = x | \ell_i = \ell) = h^*(x|\ell)$

We assume that ℓ, X, Y are discrete random variables with finite support. We also assume that the MLM assumption holds, i.e. $h^*(x|\ell) = \sum_y p^*(x|\ell) q^*(x|y)$ for some unknown q^*.

The method proceeds as follows:

1. Summarize the input data, with two matrices:

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

   That is, $N^X_{\ell x}$ indicates the number of specimens from subpopulation ℓ which yielded measurement x under tool I. Likewise $N^Y_{\ell y}$ indicates the number of specimens from subpopulation ℓ which yielded measurement y under tool II.

2. Produce point estimates.

   - For p, we use a simple pseudocount estimator, namely

$$\hat{p} = \arg\max_p \sum_{\ell, x} (N^X_{\ell x} + 1) \log p(x|\ell)$$

   - For h, we take a similar idea. However, in this case, we make use of the Markov Link Assumption to increase the effficiency of our estimator. Let S denote the set of values of h which are consistent with p and the Markov Link Assumption, i.e.

$$S = \left\{ h : \ \exists q : \ \sum_x \hat{p}(x|\ell) q(y|x) = h(y|\ell) \quad \forall \ell, y \right\}$$

     And then we define

$$\hat{h} = \arg\max_{h \in S} \sum_{\ell, y} (N^Y_{\ell y} + 1) \log h(x|\ell)$$

   - Finally, we estimate q. Towards this end, we define Θ as the set of values of q which are consistent with p, h and the Markov Link Assumption. That is, let

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

     Any q inside Θ should be considered equally plausible, since it yields the same distribution on the observable data. However, for the purposes of determining

variability due to lack of data, we take a very particular estimator of q, namely the so-called "analytic center" of Θ.

$$\hat{q} = \arg\max_{q \in \Theta(\hat{p},\hat{h})} \sum_{x,y} \log q(y|x)$$

This gives us a single estimator for q, which we will use below.

Several remarks are in order about these point estimates.

- The pseudocount. An obvious alternative to the pseudocount estimator would be to simply use the maximum likelihood estimator, i.e. $\arg\max_p \sum_{\ell,x} N^X_{\ell x} \log p(x|\ell)$. However, in the datasets we examined, there was simply not enough data to support this kind of estimator. Some of the subpopulations had very few samples, requiring some degree of smoothing. Pseudocounts are just one among a host of ways this smoothing could have been achieved. However, we chose pseudocounts because they are easy to interpret both mathematically and by scientists. We simply find the maximum likelihood estimator *as if* the data had been that given by the pseudocount-augmented values. We here chose a pseudocount value of 1, but in practice an expert can meaningfully evaluate what kinds of pseudocounts each sample category should take.
- We do use the Markov Link Assumption to estimate h. We note that an alternative way to estimate h would be to simply use a pseudocount estimator just as we did to estimate p. However, if the Markov Link Assumption holds, we should be able to use this fact to get a superior estimators by enforcing this constraint. We saw in simulations that this does indeed occur at least in some cases.
- We don't use the Markov Link Assumption to estimate p. The process outlined above follows three distinct stages: estimate p, estimate h, then estimate q. This process has the advantage that each individual problem is convex, with a unique solution which can be found with classic methods from convex optimization (see Appendix C for details). However, there is a fundamentally different way we could seek our point estimates. We could try to *simultaneously* optimize all of these objects to maximize the likelihood of the observables (subject to the Markov Link Assumption). Unfortunately, we were unable to get rigorous guarantees for the asymptotic properties of such a method. It is for this reason that we took this multi-stage approach.

3. Produce quantitative estimates for the variance of our estimators, using the bootstrap. We a series of K surrogate datasets $(N^{X,(1)}, N^{Y,(1)}) \cdots (N^{X,(K)}, N^{Y,(K)})$ by sampling with replacement within each subpopulation from the original data. For each surrogate $(N^{X,(i)}, N^{Y,(i)})$ we use a procedure similar to the one described above to produce point estimates $\hat{p}^{(i)}, \hat{q}^{(i)}, \hat{h}^{(i)}$. The only difference between the bootstrap point estimators and the original point estimators is that we use no pseudocount for the bootstrap estimators. This is because the pseudocount can in some cases make the estimator variance look smaller than it is. Finally, for each variable, p, q, h, we compute the average total variation distance between our original point estimates and these bootstrap estimates. Here by total variation distance we mean for example that $\left\|\hat{q} - \hat{q}^{(i)}\right\|_{TV} \triangleq \frac{1}{2}\sum_{x,y} |\hat{q}(y|x) - \hat{q}^{(i)}(y|x)|$. We denote these variation distance averages by $\Delta_p, \Delta_q, \Delta_h$.
4. Produce quantitative estimates for our uncertainty about q^* due to identifiability issues. In some cases it may be that q^* is not identifiable from the data. That is, even in the limit of an infinite number of specimens, it may be impossible to determine the exact value of q^*. This is because we can only observe q^* through the distribution on $Y|\ell$ (i.e. h^*). Thus, if there are two values q_1, q_2 such that $h^*(y|\ell) = \sum_x p^*(x|\ell) q_1(y|x) = \sum_x p^*(x|\ell) q_2(y|x)$, then without joint measurement on (X, Y) we would never be able to tell if $q^* = q_1$ or $q^* = q_2$. Nonetheless, even if q^* cannot be determined exactly,

it is possible to put bounds on what values of q^* are consistent with the data: the Markov Link Method assumption does say that $q^* \in \Theta(p^*, h^*)$, where Θ is given by Equation (1). This can tell us something about what the calibration q^* looks like. What is more, subject to mild conditions, we can show that in the limit of infinite data, q^* is very close to some point in $\Theta(\hat{p}, \hat{q})$. This is the subject of Theorem 1 in Appendix A. To understand our uncertainty about q^* due to identifiability, we must therefore understand the size of this set. To do so, we use the simple total variation diameter approximation procedure (see Appendix B for details) to estimate the diameter for the set associated with each of our bootstrap samples, $d^{(i)} \approx \text{diam}(\Theta(\hat{p}^{(i)}, \hat{h}^{(i)}))$. To estimate the error due to identifiability in our particular case, we use the median of these diameter estimates, which we denote by $\bar{d}$.

5. Produce qualitative variability visualizations. Fix any value of x. Now consider any bootstrap estimate $\hat{p}^{(i)}, \hat{q}^{(i)}, \hat{h}^{(i)}$. For each y we can compute the value of $q \in \Theta(\hat{p}^{(i)}, \hat{q}^{(i)})$ with the maximal value of $q(y|x)$. We call this an "extremal bootstrap estimate,"

$$q^{(i,x,y)} = \sup_{q \in \Theta(\hat{p}^{(i)}, \hat{q}^{(i)})} q(y|x)$$

We can look at all of these estimates simulatenously by forming a square matrix $M_{y_1,y_2} = q^{(i,x,y_1)}(y_2|x)$ and plotting this matrix as a heatmap. If there are no identifiability issues concerning the estimation of q^*, every row of the matrix will be the same. If there are significant identifiability issues, different rows may be quite different. We can repeat this process for each bootstrap estimate; this yields a very tall and thin matrix formed by stacking all the matrices together. If there is sufficient data, we expect each block of this matrix to be approximately the same (since the bootstrap resampling will have little effect). Thus, if there is sufficient data and there are no identifiability issues every row of the entire tall matrix will be similar. For visualization, we can squish the matrix so that each row is much thinner than each column. The result is a single image that captures the variability for one particular row of q.

8 Relation to prior work

Our main goal is 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." 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 [14] 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 [15]. Indeed the physics literature has contributed many key inequalities (cf. [16], [17], and the references therein). Perhaps the closest work to this one would be [18], 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.

References

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

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

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

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

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

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] Andreas Brieden, Peter Gritzmann, Ravindran Kannan, Victor Klee, László Lovász, and Miklós Simonovits. Deterministic and randomized polynomial-time approximation of radii. *Mathematika*, 48(1-2):63–105, 2001.

A Mathematical results

- 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 $|a|_\infty \leq |a| \leq \sqrt{n}\,|a|_\infty$ 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 $q^* \in T_{|\Omega_X|,|\Omega_Y|}$.
- Let $h^* = p^*q^*$. Here by p^*q^* we intend the matrix multiplication of p^* and q^*. We will go back and forth between thinking of p^*, q^*, h^* as matrices and thinking of them as distributions.
- We require the matrix q^* has strictly positive entries, $q^*_{xy} \geq c > 0$.
- For each ℓ draw n_ℓ samples from p^*_ℓ and m_ℓ samples from h^*_ℓ. Let $N^X_{\ell x}, N^Y_{\ell y}$ summarize the resulting number of each kind of sample we obtained.
- We require that there are fewer rows of p^* than columns, and each row is linearly independent. If there are more rows than columns, then there is generally no identifiability issue so the theorem below is not relevant. If there are fewer rows than columns but the rows are not independent, then the result applies after enough rows are combined.
- Let $\hat{p}_{\ell x} = (N^X_{\ell,x} + 1)/(\sum_{x'} N^X_{\ell,x'} + 1)$

- Let
$$L(q,\kappa) = \sum_{y,\ell}(N^Y_{\ell y}+1)\log\left(\sum_x \hat{p}(x|\ell)q(y|x)\right) + \kappa\sum_{x,y}\log q(y|x)$$
and take $\hat{q} = \lim_{\kappa\to 0}\arg\max_q L(q,\kappa)$.
- Let $\hat{\Theta} = \{q:\ \hat{p}\hat{q} = \hat{p}q\} \cap T_{|\Omega_X|,|\Omega_Y|}$.

Theorem 1. *If $n_\ell, m_\ell \to \infty$ in such a way that $m_\ell/m_{\ell'} \geq \rho > 0$ for each ℓ,ℓ', then $\inf_{q\in\hat{\Theta}} |q^* - q|_\infty \to 0$ in probability.*

Proof. It is well-known that $\hat{p} \to p^*$ in probability (in both the uniform or the Euclidean norm, which are of course equivalent in this case). It is easy to see that the condition on m_ℓ together with the strict positivity of q^* guarantee that the same goes for $\hat{p}\hat{q} \to h^*$. Thus, intuitively, the difficulty is this: by ensuring $|\hat{p}-p^*|, |\hat{p}\hat{q}-h^*|$ sufficiently small, can we find some $\tilde{q}\in\hat{\Theta}$ so that $|\tilde{q}-q^*|$ is arbitrarily small? It turns out we can.

There is a simple method to find such a $\tilde{q}$: we take $\tilde{q}$ to be the Euclidean projection of q^* to the plane defined by $\{q:\ \hat{p}\hat{q} = \hat{p}q\}$. If $p^* \approx \hat{p}$ and $\hat{p}\hat{q} \approx p^*q^*$, it is easy to see that we can can ensure this projection will not carry us a very large distance, and so $\tilde{q}$ and q^* will be close, as desired. However, it might carry us to a point outside of the transition matrix polytope; in particular, after projection we might see some negative values. Fortunately, because we have insisted that the truth is strictly positive, we are guaranteed that this is not a problem in the asymptotic limit.

Let us make this argument formal. We haved defined $c > 0$ as 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
$$\begin{aligned}|\tilde{q}-q^*| = \left|\hat{p}^\dagger\hat{p}(\hat{q}-q^*)\right| &\leq 2M\,|\hat{p}\hat{q}-\hat{p}q^*| \\ &\leq 2M\,|\hat{p}\hat{q}-p^*q^*| + 2M\,|(p^*-\hat{p})q^*| \\ &\leq 2M\frac{\epsilon}{4M} + \frac{2M\epsilon}{4M\sqrt{|\Omega_X||\Omega_Y|}}\sqrt{|\Omega_X||\Omega_Y|}\,|q^*|_\infty \leq \epsilon\end{aligned}$$
- $\hat{p}\tilde{q} = \hat{p}q^* + \hat{p}\hat{q} - \hat{p}q^* = \hat{p}\hat{q}$
- The rows of $\tilde{q}$ sum to 1. This is easy to see, because the rows of q^* sum to 1 and the rows of $\hat{q}$ sum to 1, and so $\tilde{q}\mathbf{1} = q^*\mathbf{1} + \hat{p}^\dagger\hat{p}(\hat{q}-q^*)\mathbf{1} = \mathbf{1}+0$ as desired.
- The entries of $\tilde{q}$ are positive. Indeed, the the smallest value of q^* is c, and we have already argued that $|\tilde{q}-q^*|_\infty \leq \epsilon$. Thus the smallest value of $\tilde{q}$ is at least $c-\epsilon$, and we have required $\epsilon < c$.

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

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

B Total Variation Diameter Estimation

We follow the algorithm of [19] in estimating of the diameter of a convex polytope C. We note that diameter estimation is a difficult problem, and if the polytope is particularly pathological and high-dimensional it may be quite difficult to estimate the diameter accurately. Nonetheless, it is a practical solution to the difficulty at hand.

Let $C \in \mathbb{R}^n$ denote a convex polytope (note that the n here does not correspond to the number of samples n in the main body of this paper). We will be interested in a certain kind of diameter of C, namely

$$\begin{aligned}\operatorname{diam}(C) &= \frac{1}{2} \sup_{x,y\in\Theta} \sum_{i=1}^{n} |x_i - y_i| \\ &= \frac{1}{2} \max_{z\in\{-1,1\}^n} \left(\sup_{x,y\in\Theta} \sum_{i=1}^{n} z_i(x_i - y_i) \right) \\ &= \frac{1}{2} \max_{z\in\{-1,1\}^n} \left((\sup_{x\in\Theta} z \cdot x) - (\inf_{x\in\Theta} z \cdot x) \right)\end{aligned}$$

For any fixed z, the inner maximization and minimization problems can be solved using linear programming. We can therefore estimate the diameter by taking every possible z and solving this problem. In practice, this may be computationally prohibitive. If so, we can randomly subsample values of z to obtain a lower bound on the diameter.

We can apply this to estimate the diameter of our set of interest, namely

$$\Theta(p,h) = \left\{ q : \sum_x p(x|\ell) q(y|x) = h(y|\ell) \right\}$$

In this case Θ is a space of conditional probability distributions. However, the constraints that define Θ are all linear, and so the same method can be applied.

C Convex optimization problems

To be determined

D Permutation matrices are perfectly identified

To be written