The Markov Link Method: a nonparametric approach to combine observations from multiple experiments

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

August 7, 2018

Abstract

Using observations from multiple experiments is essential to building larger scientific theories. For example, we might analyze different aspects of human neurons, such as morphology, transcriptomics, or electrophysiology. A strong theory would account for each perspective as well as the relationship between the perspectives. Unfortunately, we rarely have a single experiment that simultaneously measures every viewpoint, uses a sufficient sample size, and gets sufficient measurement precision. In practice, we must therefore conduct different experiments and find a way to link the results. The Markov Link Method (MLM) offers an algorithm to reason about these 'links.' The method is nonparametric: it makes no assumptions about what the link should look like. We evaluate the MLM on a pair of single-cell RNA techniques and gain new insight about the link between the two methods.

The last century has seen a proliferation of experimental techniques. Different perspectives arise from different techniques (and even from different labs performing the 'same' protocol). Ideally, this yields theories that account for all perspectives. Instead, sometimes, this leaves us with many perspectives and no idea how they are connected.

Our goal is to connect different perspectives. To make this idea rigorous, we assume each experiment has two aspects:

- a *sampling procedure* determines what we study
- a *measurement procedure* defines the perspective we take in studying it.

If two experiments differ in both aspects, it may be difficult to gain insight by combining their data. For example, imagine one experiment performs a transcriptomic measurement of liver cells and another applies a morphological measurement of heart cells. Without additional experiments or knowledge, looking at these two experiments together may not be very helpful. On the other hand, if two experiments have some aspect in common, there is something to be learned. In this sense, a collection of experiments can be understood as a network: a pair of experiments are connected if they have some aspect in common. In some cases, we can also connect two experiments if they differ in both aspects but we have prior knowledge (e.g. if the sampling procedures are different, but a reweighting can be used to correct for the difference). When many experiments can be joined into a common network, much can be learned.

We here propose the The Markov Link Method (MLM), an algorithm that uses a network of experiments to learn the relationship between two different perspectives. We formalize this through what we call the 'measurement link' parameters, $q^*(y|x)$. The meaning of these parameters is defined as follows. Let us say we obtain measurement result x from one measurement procedure. The number $q^*(y|x)$ indicates the probability of obtaining result y from the second measurement procedure applied to measure the same specimen. The link tells us how the properties measured by each procedure are associated. For example,

the link lets us answer questions such as "I have observed the transcriptome of this cell – what morphologies might this cell have?"

If we have a single large experiment that measures each specimen with both measurement procedures, it is straightforward to estimate the link. We can examine specimens through both perspectives simultaneously. However, in many cases we have some experiments in which specimens are measured with one procedure, other experiments using the other procedure, and few or no experiments where specimens are measured with both procedures. This is particularly true if the measurement techniques destroy or alter the specimen in the process of measuring it.

In this paper we focus particularly on the case that we have a network of experiments based on several sampling procedures and exactly two different measurement procedures. We also suppose that the measurement procedures return categorical observations, i.e. a 'measurement' assigns a specimen to one of a finite set of categories. The MLM could also generalize to numerical observations, but in that case it would be appropriate to incorporate some knowledge of smoothness (e.g. through a Gaussian assumption). Categorical data requires no additional modeling assumptions, so for simplicity we will here focus on the categorical case. This setup, sketched in Figure 1, is perhaps the simplest example that highlights all the important assumptions and properties of the MLM.

1 The Markov Link Method

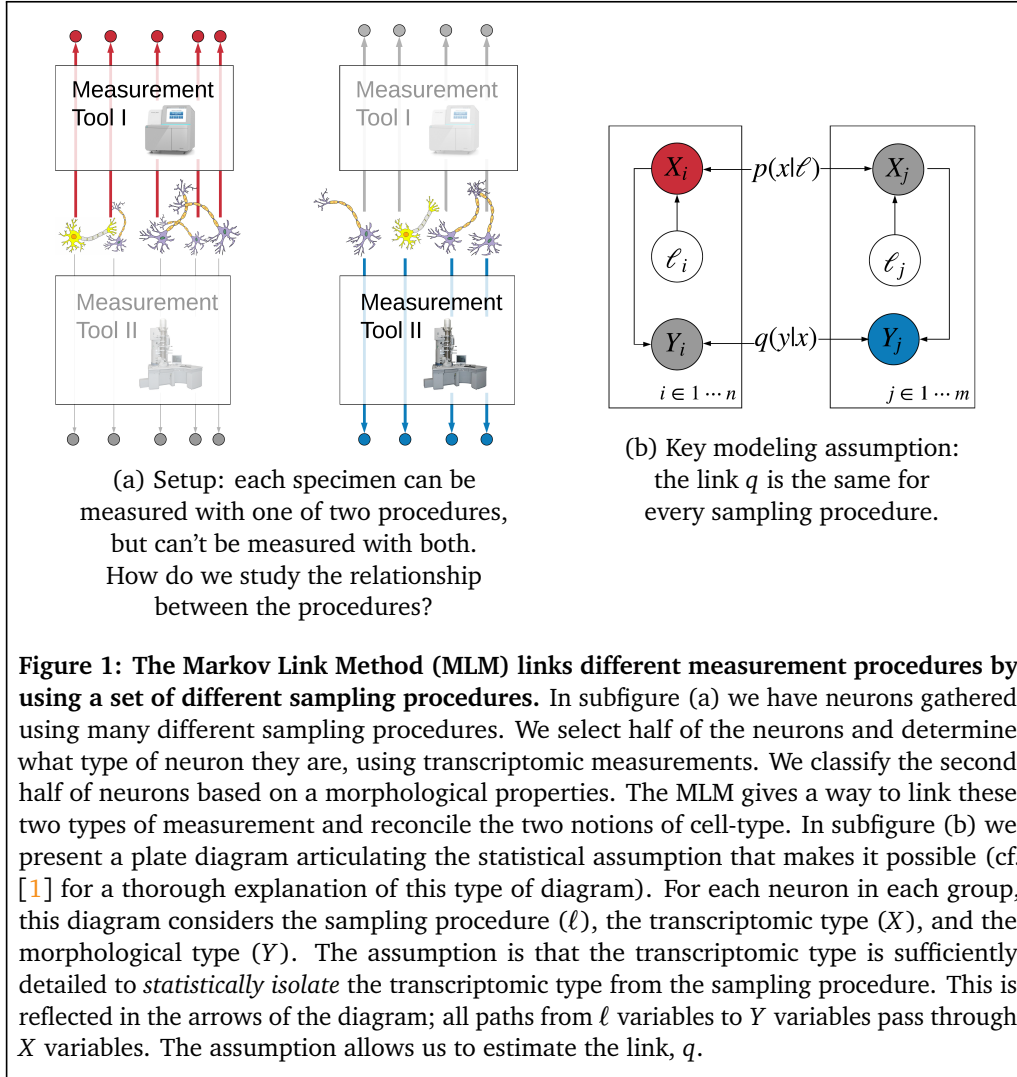Imagine that we have a large collection of human cells. For each cell we are interested in

1. ℓ, the sampling procedure used to obtain the specimen. For example, we can select different subpopulations of cells based on proteomic markers. For each marker we might have a different sampling procedure designed to select cells with that marker.

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. We would like to know what transcriptomic types are associated with what morphological types. For example, we might like to say "This cell has transcriptomic type 3, therefore it probably has morphological type F." Our main assumption is that these kinds of associations are unaffected by the choice of sampling procedure. In particular:

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

Here $\mathbb{P}(Y|X,\ell)$ indicates the probability that the morphological type is y given that the transcriptomic type is x and the specimen was sampled with strategy ℓ. The assumption is that this probability actually does not depend upon ℓ. This can be understood intuitively through a thought experiment. Let us imagine we have attained perfect understanding of cellular biology. Someone selects a cell using one of a set of sampling procedures and measures the transcriptomic activity of the cell. We are then told the cell's transcriptomic activity (but we are not told which of the sampling procedured was used to gather the specimen). Using our perfect knowledge of the physical systems, we could then make predictions about a measurement of the cell's morphology. Now we are told new information: we are told which sampling procedure was used to gather the specimen. Would this

2

(a) Setup: each specimen can be
measured with one of two procedures,
but can't be measured with both.
How do we study the relationship
between the procedures?

(b) Key modeling assumption:
the link q is the same for
every sampling procedure.

Figure 1: The Markov Link Method (MLM) links different measurement procedures by using a set of different sampling procedures. In subfigure (a) we have neurons gathered using many different sampling procedures. We select half of the neurons and determine what type of neuron they are, using transcriptomic measurements. We classify the second half of neurons based on a morphological properties. The MLM gives a way to link these two types of measurement and reconcile the two notions of cell-type. In subfigure (b) we present a plate diagram articulating the statistical assumption that makes it possible (cf. [1] for a thorough explanation of this type of diagram). For each neuron in each group, this diagram considers the sampling procedure (ℓ), the transcriptomic type (X), and the morphological type (Y). The assumption is that the transcriptomic type is sufficiently detailed to *statistically isolate* the transcriptomic type from the sampling procedure. This 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 estimate the link, q.

substantially change our predictions? If so, the MLM assumption is violated. However, if the transcriptomic measurement is sufficiently informative, learning the sampling strategy would not substantially change our predictions about the cell morphology. As long as we can find a measurement procedure that is sufficiently informative in this way, we can apply the MLM. A list of examples where the MLM might apply may be found in Section 2.

Our task is to estimate the link, q^*, from a collection of experiments. We suppose that for each sampling procedure we have two experiments: one that measured transcriptomics and another that measured morphology. We have no experiments in which both types were measured for the same specimens. This setup is sketched in Figure 1. Different setups could also be considered. For example, we might additionally have a small experiment in which both measurement procedures could be applied to the same specimen. We might have more than two measurement procedures. We might have measurement procedures that yield numerical observations instead of categorical ones. Generalizing the method to all these cases should be straightforward, but we leave it for future work.

In the specific case we will focus on, all of the experimental data can be summarized with two matrices, N^X and N^Y. Here $N^X_{\ell x}$ indicates the number of specimens gathered with procedure ℓ that have transcriptomic type x. Likewise $N^Y_{\ell y}$ indicates the number of specimens gathered with procedure ℓ that have morphological type y. For each transcriptomic type x and morphological type y, the Markov Link Method uses the data to produce two objects:

1. $\hat{q}(y|x)$ – a point estimate for the link parameter $q^*(y|x)$

2. $C_{x,y}$ – a credible interval which contains the true link parameter $q^*(y|x)$ with high probability

To produce these objects we proceeds as follows. Let $\tilde{P}(\cdot|\ell) \sim \text{Dirichlet}(N^X_\ell + 1)$ and $\tilde{H}(\cdot|\ell) \sim \text{Dirichlet}(N^Y_\ell + 1)$ for each ℓ. For any pair h, h' let

$$D(h|h') = \sum_{\ell,y} h(y|\ell) \log \frac{h(y|\ell)}{h'(y|\ell)}$$

denote the Kullback-Leibler divergence from h to h'. For each $\tilde{P}, \tilde{H}$, let $E(\tilde{P}, \tilde{H})$ denote the minimum value of $D(\tilde{H}|h')$ among all h' such that there exists q satisfying $h'(y|\ell) = \sum_x \tilde{P}(y|\ell)q(y|x)$. Let P, H denote variables carrying the distribution of $\tilde{P}, \tilde{H}$ after conditioned on $E(\tilde{P}, \tilde{H}) = 0$. Let $\Theta(p,h) = \{q : h(y|\ell) = \sum_x p(x|\ell)q(y|x)\}$. We then define the MLM objects as follows:

1. $\hat{q} \triangleq \mathbb{E}[\arg\min_{q \in \Theta(P,H)} \sum_{x,y} q(y|x)^2]$.

2. The lower bound of $C_{x,y}$ is defined as the largest number c such that $\min_{q \in \Theta(P,H)} q(y|x) \geq c$ with probability 97.5%. Likewise the upper bound is the smallest number c such that $\max_{q \in \Theta(P,H)} q(y|x) \leq c$ with probability 97.5%.

Code to compute these objects is published at https://github.com/jacksonloper/markov-link-method, including a tutorial-style ipython notebook detailing every computation made in this paper.

Our approach arises naturally from a Bayesian point of view. To understand them, first notice that there are several 'nuisance parameters' we must consider in order to estimate the link q^*. Let $p^*(x|\ell)$ denote the probability that a cell sampled with procedure ℓ will have transcriptomic type x and $h^*(y|\ell) \triangleq \sum_x p^*(x|\ell)q(y|x)$ denote the probability such a cell will have morphological type y. Note that the distribution of the data N^X, N^Y is entirely determined by p^*, h^*. In particular, if someone changed q^* without changing p^*, h^* there would be no way for us to know. If this can occur, q^* is said to be 'nonidentifiable.' Therefore, instead of trying to directly estimate q^* (which may be impossible to accurately

estimate) we focus on three quantities which can be estimated from data:

$$q_{\mathrm{mid}}^* \triangleq \underset{q \in \Theta(p^*, h^*)}{\arg \min} \sum_{x,y} q(y|x)^2$$

$$q_{\mathrm{lo},x,y}^* \triangleq \min_{q \in \Theta(p^*, h^*)} q(y|x)$$

$$q_{\mathrm{hi},x,y}^* \triangleq \max_{q \in \Theta(p^*, h^*)} q(y|x)$$

Note that all three of these optimization problems have unique solutions and would be straightforward to solve if p^*, h^* were known; optimization details can be found in Appendix A. Moreover, regardless of any identifiability issues for q^* itself, the objects $q_{\mathrm{lo},x,y}^*$ and $q_{\mathrm{hi},x,y}^*$ are always identifiable for every (x, y) and define bounds on the true value of $q^*(y|x)$.

The MLM algorithm therefore adopts a Bayesian point of view to estimate q_{mid}^* and create a credible interval for $q_{\mathrm{lo},x,y}^*, q_{\mathrm{hi},x,y}^*$. We begin with uniform prior beliefs for the parameters of p^*, h^*. That is, our prior is that $p^*(\cdot|\ell), h^*(\cdot|\ell)$ carry Dirichlet(1) distributions for each ℓ. Following the Bayesian philosophy, we would like to incorporate any additional knowledge through conditioning. We have two important pieces of knowledge we would like to incorporate: that the MLM assumption holds (i.e. there is some q^* so that $\sum_x p^*(x|\ell)q^*(y|x) = h^*(y|\ell)$) and we have observed the data N^X, N^Y. By conditioning on this knowledge, we obtain the posterior distribution carried by the variables P, H described above. In terms of these variables, the MLM estimator $\hat{q}$ is simply the posterior mean for q_{mid}^* and the MLM credible interval $C_{x,y}$ is an interval which contains $q_{\mathrm{lo},x,y}^*, q_{\mathrm{hi},x,y}^*$ with high probability.

In practice, we were not able to find a way to compute the credible interval and posterior means exactly. It seems difficult to even sample the conditional random variables P, H. Common approaches to this type of problem involve Markov-Chain Monte Carlo and Variational Inference, but we were unable to make these approaches work in practice. It seems nontrivial to work with the condition $E(p, h) = 0$ that formalizes the MLM assumption. We instead take a somewhat naïve approach. We start with $\tilde{P}, \tilde{H}$, which carry the distribution of p^*, h^* conditioned on the observed data (but not conditioned on the fact that the MLM assumption holds). We then simply project $\tilde{H}$ to the space of parameters which are consistent with the MLM assumption. That is, we take $\hat{H}$ to be the solution of the minimization problem $\min_{h'} D(\tilde{H}|h')$, subject to the constraint that $E(\tilde{P}, h') = 0$. Optimization details can be found in Appendix A. We hope that the distribution of $\tilde{P}, \hat{H}$ will approximate the desired posterior P, H. We then use samples from this distribution to compute what we need. The implications of this approximation are difficult to reason about; we hope further study will clarify when this is a reasonable procedure. In the mean time, we content ourselves that it is a practical solution to a difficult problem.

A frequentist point of view might provide a valuable counterpoint to our Bayesian analysis. However, we were unable to produce effective confidence intervals for the link parameters. There are at least two difficulties: the link is unidentifiable and the true parameter values may lie on the edge of the parameter space. These challenges make traditional appproaches based on asymptotic normality difficult to apply. Perhaps future work could uncover a way to overcome these challenges.

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. [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.

- 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. [3]). 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. [4]). However, we know that cultured cell-lines look quite different from in-vivo cells (cf. [5, 6]). 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. [7]). 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 [8]. 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. [9, 10]), 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.
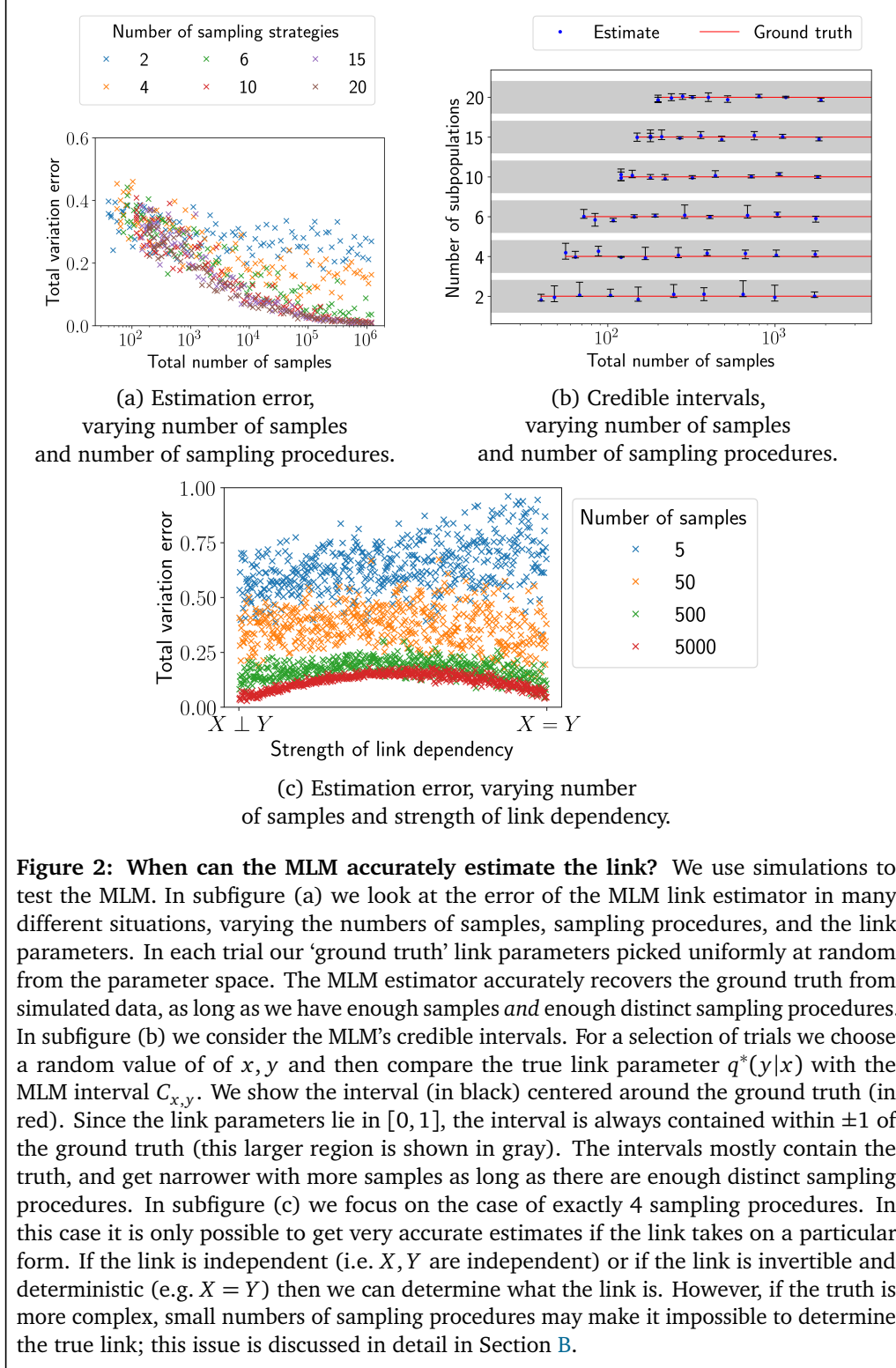
3   Simulations

We used simulations to test the two parts of the MLM: the point estimators $\hat{q}(y|x)$ and credible intervals $C_{x,y}$. We would ordinarily hope for two things:

Estimator convergence   As we obtain more samples, $\hat{q}(y|x)$ should converge to the true link $q^*(y|x)$ for each x, y.

Conservative coverage $q^*(y|x) \in C_{x,y}$ with high probability.

Our simulations yielded good news and bad news. The good news: the MLM generally achieves conservative coverage. In many cases, the MLM also achieves estimator convergence. The bad news: in many cases the MLM fails to achieve estimator convergence. Indeed, we can show that the identifiability issues block any possibility of success in some cases; we detail this problem in Appendix B. The following two simulations give a good overview of the relevant issues:

- How does MLM performance depend on the number of samples and the number of sampling procedures? We suppose we have two measurement procedures which return categorical measurements among six categories. Thus the first procedure yields $X \in \{1, 2, 3, 4, 5, 6\}$ and the second yields $Y \in \{1, 2, 3, 4, 5, 6\}$. To see how the method performs in different circumstances, we will run many trials. In each trial we fix the number of sampling procedures and pick 'ground truth' parameters uniformly at random from the parameter space. We then simulate a dataset from these ground truth parameters, fixing the total number of samples (spreading these samples equally among all combinations of sampling procedure and measurement procedure). Finally, we apply the MLM to the simulated data to get the point estimates $\tilde{q}(y|x)$ and the credible intervals C_{xy}. We measure the overall estimator convergence using a kind of

(a) Estimation error,
varying number of samples
and number of sampling procedures.

(b) Credible intervals,
varying number of samples
and number of sampling procedures.

(c) Estimation error, varying number
of samples and strength of link dependency.

Figure 2: When can the MLM accurately estimate the link? We use simulations to test the MLM. In subfigure (a) we look at the error of the MLM link estimator in many different situations, varying the numbers of samples, sampling procedures, and the link parameters. In each trial our 'ground truth' link parameters picked uniformly at random from the parameter space. The MLM estimator accurately recovers the ground truth from simulated data, as long as we have enough samples *and* enough distinct sampling procedures. In subfigure (b) we consider the MLM's credible intervals. For a selection of trials we choose a random value of of x, y and then compare the true link parameter $q^*(y|x)$ with the MLM interval $C_{x,y}$. We show the interval (in black) centered around the ground truth (in red). Since the link parameters lie in $[0, 1]$, the interval is always contained within ± 1 of the ground truth (this larger region is shown in gray). The intervals mostly contain the truth, and get narrower with more samples as long as there are enough distinct sampling procedures. In subfigure (c) we focus on the case of exactly 4 sampling procedures. In this case it is only possible to get very accurate estimates if the link takes on a particular form. If the link is independent (i.e. X, Y are independent) or if the link is invertible and deterministic (e.g. $X = Y$) then we can determine what the link is. However, if the truth is more complex, small numbers of sampling procedures may make it impossible to determine the true link; this issue is discussed in detail in Section B.

total variation distance:

$$\text{Error}(\tilde{q}, q^*) = \frac{1}{12} \sum_{x=1}^{6} \sum_{y=1}^{6} |\tilde{q}(y|x) - q^*(y|x)|$$

This error ranges between zero and one. Zero error indicates that $\tilde{q} = q^*$. An error of one indicates that the estimate has completely incorrect beliefs about the probability mass, i.e. $q^*(y|x) = 0$ whenever $\tilde{q}(y|x) > 0$ and vice-versa. We also look at the MLM credible intervals and see how often they cover the true parameters.

The results are summarized in Figure 2. In trials with more samples, the estimator generally has lower error. However, to make the error actually converge to zero we need at least six distinct sampling procedures. This figure also shows that the intervals work correctly regardless of of the number of samples or sampling procedures; they include the ground truth with high probability. With many samples and sampling procedures, the intervals are small and concentrated around the truth. With fewer samples or sampling procedures, the intervals are sometimes forced to be larger. However, our uncertainty is not the same for every parameter. In some cases we are able to obtain a very tight credible interval for one particular parameter even though the overall estimator error is high.

- How does estimator convergence depend upon the link parameters themselves? In each trial of this simulation we use four sampling procedures and our measurement procedures yield one of six categories. In the previous simulations we saw that estimator convergence was generally impossible in this situation, due to the small number of sampling procedures. However, in our previous simulations we picked the link parameters uniformly from parameter space. Now we will be more choosey. On one extreme, we will produce trials where the link makes X, Y independent, i.e. $q^*(y|x) = q^*(y|x')$ for every x, x'. On the other extreme, we will have trials where X, Y are deterministically related by the equation $X = Y$. We will also consider every link "in-between" these two extremes (found by convex combinations). Figure 2 shows that estimator convergence is actually possible in the two extreme cases. However, estimator convergence fails for the in-between cases. We leave a complete mathematical understanding of this result to future work. For the purposes of this paper, we content ourselves by noting one interesting case: let the measurements return one of 2^k categories and let us use only $k + 1$ carefully chosen sampling procedures. Now suppose the link defines any invertible deterministic function between X and Y. In this case, with enough enough samples, we can determine both that the relationship is deterministic and the exact specification of the invertible function. This result is proven in Appendix B, Theorem 1.

4   Empirical results for cell-types

4.1   Background

Our motivation for this problem arose from the Allen Institute's investigation of neuronal cell-types. The institute has a variety of methods for examining a neuron and determining what kind of neuron it is. However, many of these procedures destroy the neurons in the process of measuring them. We'll look at two procedures in particular, which we'll call 'Standard' and 'Patch.' Standard uses a single-cell RNA sequencing pipeline to determine transcriptomic expression in the cell and determines the cell-type based on these results. Patch also uses transcriptomic information from RNA sequencing. Patch additionally obtains electrophysiological and morphological properties of the neuron. This comes at the cost of
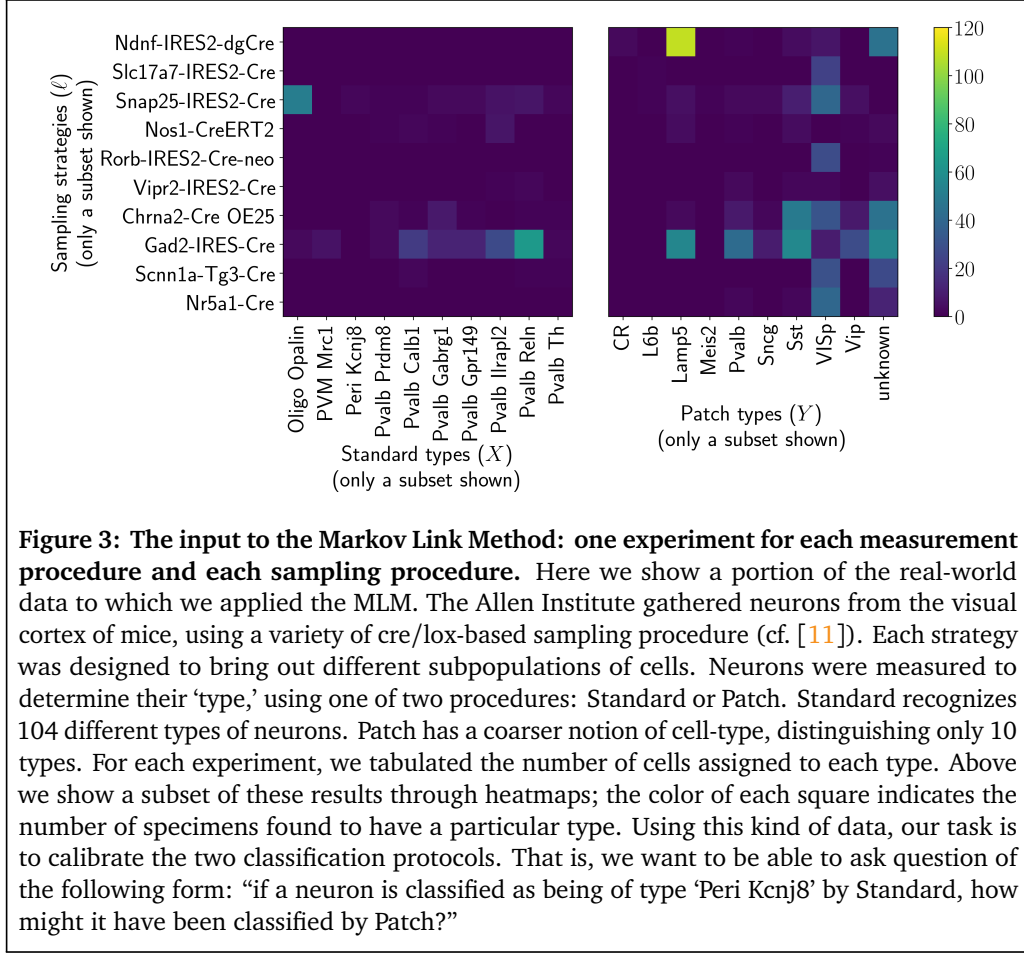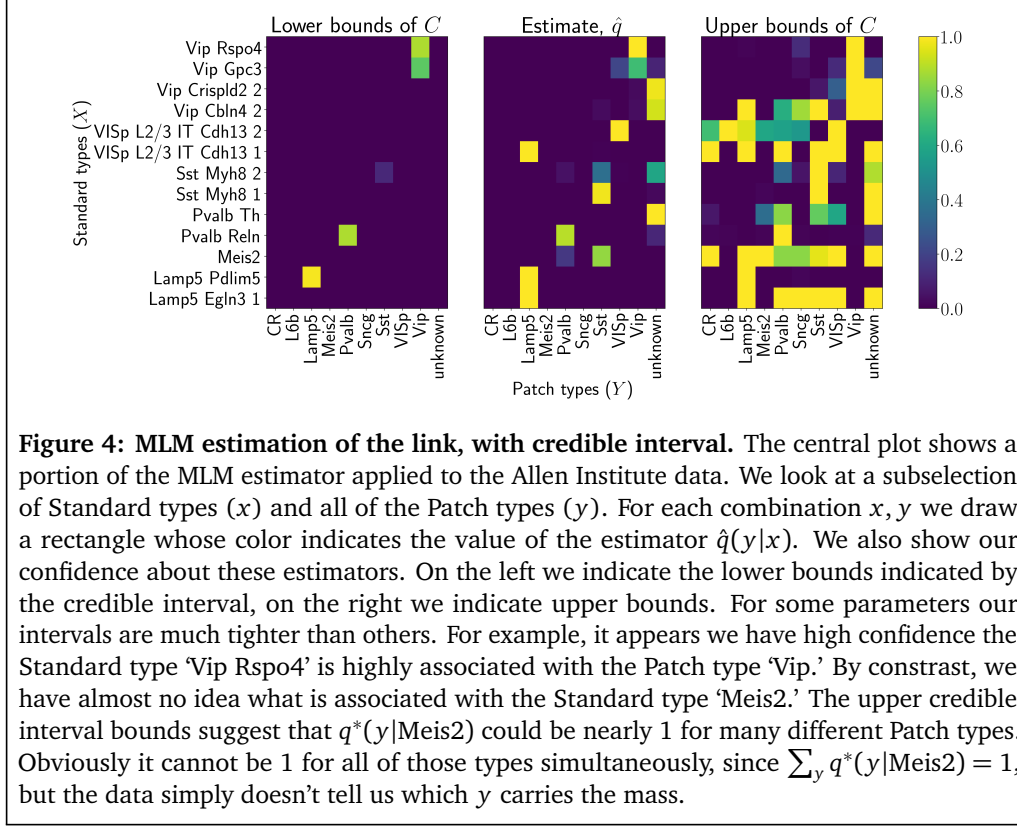
Figure 3: The input to the Markov Link Method: one experiment for each measurement procedure and each sampling procedure. Here we show a portion of the real-world data to which we applied the MLM. The Allen Institute gathered neurons from the visual cortex of mice, using a variety of cre/lox-based sampling procedure (cf. [11]). Each strategy was designed to bring out different subpopulations of cells. Neurons were measured to determine their 'type,' using one of two procedures: Standard or Patch. Standard recognizes 104 different types of neurons. Patch has a coarser notion of cell-type, distinguishing only 10 types. For each experiment, 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 found to have a particular type. 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 Standard, how might it have been classified by Patch?"

Figure 4: MLM estimation of the link, with credible interval. The central plot shows a portion of the MLM estimator applied to the Allen Institute data. We look at a subselection of Standard types (x) and all of the Patch types (y). For each combination x, y we draw a rectangle whose color indicates the value of the estimator $\hat{q}(y|x)$. We also show our confidence about these estimators. On the left we indicate the lower bounds indicated by the credible interval, on the right we indicate upper bounds. For some parameters our intervals are much tighter than others. For example, it appears we have high confidence the Standard type 'Vip Rspo4' is highly associated with the Patch type 'Vip.' By constrast, we have almost no idea what is associated with the Standard type 'Meis2.' The upper credible interval bounds suggest that $q^*(y|\text{Meis2})$ could be nearly 1 for many different Patch types. Obviously it cannot be 1 for all of those types simultaneously, since $\sum_y q^*(y|\text{Meis2}) = 1$, but the data simply doesn't tell us which y carries the mass.

a degraded transcriptomic signal, requiring new methods to estimate the cell-type. We refer the reader to [11] for details on the two methods.

Best efforts were made to use biological intuition to calibrate the methods. For example, Patch has a notion of a "Lamp5" cell type. Standard gives a more granular analysis, dividing this type into many sub-types, such as "Lamp5 Pdlim5" and "Lamp5 Slc35d3." If a cell was designated as "Lamp5 Pdlim5" using Standard, the hope was that it be given the "Lamp5" type by Patch. 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 whether the two methods were properly calibrated: a variety of sampling procedures. Using a cre/lox system (cf. [11]) they were able to pick out various overlapping subpopulations of neurons using different procedures. Each procedure was expected to yield different proportions of the different cell-types. For each sampling procedure and each measurement procedure, many specimens were sampled and their cell-types determined. The result of this process was two tables, parts of which are shown in Figure 3. 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

We applied the MLM to this data to obtain point estimates $\tilde{q}(y|x)$ and credible intervals C_{xy}. In Figure 4 we visualize these objects for selected values of x, y. Each parameter has a slightly different story. For example, for the Standard type $x =$'Vip Rspo4' and the Patch type $y =$'Vip,' we have that $C_{x,y} = [.92, 1.0]$. This supports the idea that the true link satisfies $q^*(y|x) \geq .92$: at least 92% of the cells classified as type 'Vip Rspo4' by the Standard method will be classified as 'Vip' by the Patch method. However, for other types there is more ambiguity. For the Standard type $x =$'Lamp5 Egln 1' and the Patch types $y =$'Lamp' we have $C_{xy} = [0, 1.0]$. The data we have does not give us a definitive answer as to whether cells with Standard type 'Lamp5 Egln 1' are being classified with Patch type 'Lamp.'

The variability in the credible 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 link. 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 Patch. 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 overall 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 link, q^*.

5   Relation to prior work

The task of this paper was infer the link between different measurement procedures, so that experimental data can be meaningfully combined. There is an enormous literature on this subject. For example, when experiments are performed in batches, the exact measurement procedures can vary slightly between batches. The entire field of 'batch effects' is devoted to handling these problems. The general approach is to use some knowledge of the procedures to make modeling assumptions about the links. These assumptions give us a way to estimate the link (cf. [10]). If different measurement procedures yield results in the same space, we can also implicitly articulate these kinds of assumptions by placing a metric on the space. We suppose that different measurement procedures should yield results that are 'close' in this metric. We can then use optimal transport techniques to produce a link based on these assumptions (cf. [12]). From the most general point of view, we are engaged in meta-analysis; we refer the reader to [13] for a general introduction to the field.

The main distinguishing characteristics of this paper are two-fold: we place no assumptions on the nature of the link and we take the identifiability issues seriously. We refer the reader to [14] for an introduction to what we mean by identifiability. We discuss the identifiability issues for our problem in particular in Appendix B. Our only assumption is that the sampling procedures can be statistically isolated from the link. This assumption says nothing about what the link itself is. We take a Bayesian approach, but we only use our prior beliefs to estimate quantities which actually can be estimated, carefully avoiding the identifiability issues. We use these quantities to produce intervals which account for uncertainty due both to small sample sizes and to identifiability issues.

The main technical contribution of this paper was figuring out how to use the MLM assumption to get credible intervals that worked in practice. In this we were 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 [15] 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 [16]. More generally, the physics literature has contributed many key assumptions that bound unidentifiable parameters (cf. [17], [18], and the references therein). Perhaps the closest work to this one would be [19], 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

It can be difficult to understand how two different measurement procedures are related to each other. This makes it difficult to understand how to combine knowledge from experiments which use different procedures. The problem is particularly tricky if we cannot look at the same specimen through the perpsective of both procedures simultaneously.

In this paper we suggest one way to overcome these difficulties. We formalize the 'relationship' between two measurement procedures through the notion of a 'measurement link.' Given that we have measured a specimen with one procedure and obtained the outcome x, the link parameter $q(y|x)$ tells us the probability that we would obtain outcome y if we measured the same specimen with the second procedure. We present the Markov Link Method assumption, which roughly states that the link can be statistically isolated from the sampling procedures. We provide the Markov Link Method (MLM) algorithm that uses this assumption to estimate the link and measure our 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.

The MLM appears to provide reasonable values in simulation and gives 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 credible intervals suggested directions for experiments which would further refine our understanding.

In future work, it would be interesting to use additional assumptions to help us estimate calibrations. For example, as it stands the Markov Link Method will only yield narrow credible intervals if each measurement tool returns one of a finite number of measurement outcomes. 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 a continuum of 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, credible regions can 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] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

[3] 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.

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

[5] 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.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] Esteban G Tabak and Giulio Trigila. Explanation of variability and removal of confounding factors from data through optimal transport. *Communications on Pure and Applied Mathematics*, 71(1):163–199, 2018.

[13] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. *Introduction to meta-analysis*. John Wiley & Sons, 2011.

[14] Eric Walter. *Identifiability of parametric models*. Elsevier, 2014.

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

[16] 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.

[17] 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.

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

[19] 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.

A Numerical issues

There are three numerical problems which the MLM must solve. Here we detail our method for solving each of them.

1. Projecting to the MLM assumption. Fix any parameters $P, \tilde{H}$. One step in the MLM involves projecting $\tilde{H}$ to the set of parameters which are consistent with P and the MLM assumption. In particular, we defined

$$D(h|h') = \sum_{\ell,y} h(y|\ell) \log \frac{h(y|\ell)}{h'(y|\ell)}$$

and we needed to solve the problem

$$\min_h D(\tilde{H}|h)$$

subject to the constraint that there exists some q such that $h(y|\ell) = \sum_x p(x|\ell)q(y|x)$. Parametrizing valid h through q, we obtain the problem

$$\max_q \sum_{\ell,y} \tilde{H}(y|\ell) \log \left(\sum_x P(x|\ell)q(y|x) \right)$$

Taking derivatives one can readily show that this problem is convex. We solve it using an iterative method. We initially guess that q is uniform. We then repeatedly make the updates

$$q(y|x) \propto q(y|x) \sum_\ell p(x|\ell) \frac{\tilde{H}(y|\ell)}{\sum_x p(x|\ell)q(y|x)}$$

until convergence. Our convergence criteria is that all parameters change less than 10^{-5} in a single iteration.

2. Linear programming. Fix any P, H. To deal with the identifiability issues, we defined $\Theta(p,h) = \{q : h(y|\ell) \triangleq \sum_x p(y|\ell)q(y|x)\}$. The MLM requires us to solve linear optimization problems within Θ, such as

$$\min_{q \in \Theta(P,H)} q(y|x)$$

We solve these problems using the `cvxopt` python package.

3. Quadratic programming. The MLM also requires us to solve quadratic optimization problems within Θ:

$$\min_{q \in \Theta(P,H)} \sum q(y|x)^2$$

We solve these problems using the `cvxopt` python package.

B Identifiability

The issue of identifiability comes up repeatedly throughout this paper. Here we give a brief overview of the fundamentals of this issue and how it effects us. We also 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.

Consider experiments yielding an $\Omega_\ell \times \Omega_X$ matrix N^X and an $\Omega_\ell \times \Omega_Y$ matrix N^Y, sampled according to

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

where $p^*(x|\ell)$ is some conditional distribution and there is some $q^*(y|x)$ such that $h^*(y|\ell) = \sum_x p^*(x|\ell)q^*(y|x)$. This is simply a restatement of the assumptions we have made throughout the data about how our parameters of interest (p^*, q^*, h^*) are related to the data we can observe (N^X, N^Y).

It is well-known that we can obtain arbitrarily good estimates of p^*, h^* by taking enough samples (i.e. taking n_ℓ, m_ℓ sufficiently high). Let us therefore imagine for a moment that we in fact have *perfect knowledge* of p^*, h^*. Even so, the data does not necessarily tell us the value of the link parameters q^*. There may be many possible links, q, which are all equally consistent with p^*, h^*. Each such link yields the exact same distribution on the data we can observe, so there can be no way to use data to distinguish among them. This is known as a 'nonidentifiability problem.' Even with infinite data, we simply cannot identify exactly what the value of q^* might be.

We will now look at some examples:

B.1 A simple failure case

Consider the case that $\Omega_\ell = \Omega_Y = 2$ and $\Omega_X = 3$. That is, there are 2 separate sampling procedures, tool I recognizes 3 categories and tool II recognizes 2 categories. In particular, let us imagine that $p^*(x|\ell) = A_{\ell x}$ and $h^*(y|\ell) = B_{\ell y}$ where A, B are matrices given by

$$A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{2}{6} & \frac{1}{6} & \frac{3}{2} \end{pmatrix}$$

$$B = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

Rows correspond to different sampling procedures and columns correspond to a different measurement outcome. Now let $q^*(y|x) = C_{xy}$, another matrix. The Markov Link Method assumption then tells us that $A \times C = B$, where $\times$ indicates matrix multiplication. This corresponds to $\Omega_\ell \times \Omega_Y = 4$ equations. We also have a normalizing constraint that $\sum_y q^*(y|x) = 1$, which creates $\Omega_X = 3$ additional equations. However, these normalizing constraints actually make two of the MLM assumption constraints redundant. In the end, we have 5 constraining equations on the matrix C. However, the matrix C contains six numbers. The result is a degree of freedom in C, corresponding to an aspect of q^* that we simply cannot resolve. For example, here are two choices of C which are both consistent with the equation $A \times C = B$:

$$C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \qquad C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{2}{3} & \frac{1}{3} \end{pmatrix}$$

B.2 Permutation matrices

Consider the case that $\Omega_\ell = k$ and $\Omega_X = \Omega_Y = 2^{k-1}$. 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 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 the same problem we started with – except 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. $\qquad\square$

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:

B.3   Discontinuity

Consider the case that $\Omega_\ell = 1$ and $\Omega_X = \Omega_Y = 2$. That is, there is only one sampling procedure (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^*, where

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

and by T we mean the transition matrix polytope, $T = \{q : q(y|x) \geq 0, \sum_y q(y|x) = 1\}$. 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 $\hat{\Theta} = \Theta(\hat{p}, \hat{h})$.*

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 link 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.