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

Jackson Loper, Osnat Penn, Trygve Bakken, David Blei, Liam Paninski,
Additional Authors To Be Determined

August 16, 2018

Abstract

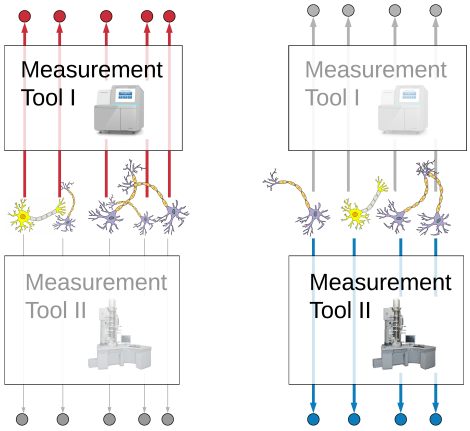
We propose the Markov Link Method (MLM), a way to synthesize experiments based on several measurement procedures. For example, we might have different experiments measuring different aspects of human neurons, such as morphology, transcriptomics, or electrophysiology. The MLM uses this data to link these different aspects. It answers the question "We have observed the transcriptome of this cell – what might its morphology have been?" The method is nonparametric: it makes no assumptions about what the link looks like. We evaluate the MLM on a pair of single-cell RNA techniques and gain new insight about the link between the two methods.

We here propose the Markov Link Method (MLM), a technique to understand how different measurement procedures are related. For our purposes, we assume each experiment has two aspects:

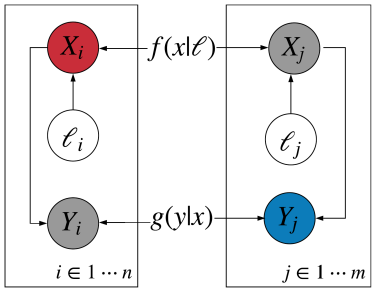
- A *selection procedure* defines how we select cells. For example, one experiment might select only liver cells, and another might select only cells which express the Vip protein.
- A *measurement procedure* defines what we measure about each cell. For example, one experiment might measure the transcriptomic profile of a cell, and another experiment might measure the cell morphology.

To synthesize experiments with different measurement procedures, we must understand how the measurement procedures are related. For example, we would like to be able to answer questions such as "We have measured the transcriptome of this cell – what might its morphology have been?" We formalize this idea through what we call the 'measurement link,' $g(y|x)$. Specifically, let us say we obtain measurement result x from one measurement procedure. The number $g(y|x)$ indicates the probability of obtaining result y from the second measurement procedure applied to measure the same specimen.

We here propose the The Markov Link Method (MLM), an algorithm that uses a collection of experiments to estimate the measurement link, g. In this paper we focus particularly on the case that we have experiments based on several selection 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. It embodies a common situation in biological problems, such as is found in (1).



(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 g is the same for every selection procedure.

Figure 1: The Markov Link Method (MLM) links different measurement procedures by using a set of different selection procedures. In subfigure (a) we have neurons gathered using many different selection 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. (2) for a thorough explanation of this type of diagram). For each neuron in each group, this diagram considers the selection 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 selection 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, g.

1 The Markov Link Method

The Markov Link Method is a general way to estimate measurement link from a collection of experiments. For concreteness, we will use experiments about human cells as a running example. Let us say that we have gathered many human cells. For each cell we are interested in

1. ℓ, the selection procedure used to obtain the specimen
2. X, the cell's 'transcriptomic type,' as measured through a transcriptomic analysis.
3. Y, the cell's 'morphological type,' as measured by looking at an image of the cell.

Our task is to link the transcriptomic and morphological perspectives on 'cell-type,' as manifested through the X and Y measurement procedures. 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 selection procedure. In particular:

The Markov Link Method Assumption
$\mathbb{P}(Y|X,\ell) = g(y|x)$ 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 states that that this probability actually does not depend upon ℓ. This assumption 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 selection procedures and measures the transcriptomic activity of the cell. We are then told the cell's transcriptomic type (but we are not told which of the selection procedures was used to gather the specimen). Using our perfect knowledge of the physical systems, we could then make predictions about the cell's morphological type. Now we are told new information: we are told which selection procedure was used to gather the specimen. Would this 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.

More formally, our task is to estimate the link g from a collection of experiments. We suppose that for each selection 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, $\mathscr{D}_X$ and $\mathscr{D}_Y$. The matrix $\mathscr{D}_X$ tabulates the transcriptomic types found in specimens gathered with each selection procedure; $\mathscr{D}_Y$ does the same for morphological types. For each transcriptomic type x and morphological type y, the Markov Link Method uses this data to produce two objects:

1. $\hat{g}(y|x)$ – a point estimate for the link $g(y|x)$
2. $C_{x,y}$ – a credible interval which contains the true link $g(y|x)$ with high probability

To estimate these, we must consider certain 'nuisance' quantities. The link is our primary target of study, but we cannot estimate the link without considering these other objects. Our observable data, $\mathscr{D}_X, \mathscr{D}_Y$, is governed by the following conditional distributions:

- $f(x|\ell)$ – the probability that a cell sampled with procedure ℓ will have transcriptomic type x
- $h(y|\ell)$ – the probability that a cell sampled with procedure ℓ will have morphological type y

It is straightforward to estimate f, h from $\mathscr{D}_X, \mathscr{D}_Y$ using standard methods. We would like to use these distributions to tell us something about our object of interest, the link g. Under the Markov Link Method assumption, this link is connected to f, h through the equations

$$\sum_x f(x|\ell) g(y|x) = h(y|\ell) \qquad \forall y, \ell \tag{1}$$

With these equations and knowledge of f, h, we can produce an estimate for the link: $\hat{g}_{f,h}$. If the equations are underdetermined, we can also produce bounds for the link. We define these objects precisely in Appendix A. However, in practice we do not actually have exact knowledge of f, h. We only have estimates from data. We take a Bayesian perspective to account for our uncertainty in this estimation. For our prior beliefs, we assume a noninformative uniform prior $\mathbb{P}$. Following the Bayesian philosophy, we then incorporate new knowledge by conditioning. We have two important pieces of knowledge about f, h. First, we have observed the data, $\mathscr{D}_X, \mathscr{D}_Y$. Second, we know from the MLM assumption that there exists *some* value g such that Equation (1) holds; let $\mathscr{A}$ denote the event that this holds. By conditioning on our knowledge, we obtain the posterior distribution $\mathbb{P}(df, dh|\mathscr{D}_X, \mathscr{D}_Y, \mathscr{A})$. Our point estimate $\hat{g}$ is then defined as a Bayes estimator for $\hat{g}_{f,h}$ with respect to this posterior:

$$\hat{g}(y|x) \triangleq \mathbb{E}\left[\hat{g}_{f,h}|\mathscr{D}_X, \mathscr{D}_Y, \mathscr{A}\right] = \int \hat{g}_{f,h} \mathbb{P}(df, dh|\mathscr{D}_X, \mathscr{D}_Y, \mathscr{A})$$

This estimate minimizes the expected squared risk error under the posterior distribution. We can use similar strategies to produce the credible intervals $C_{x,y}$. Exact details can be found in Appendix A. Code to compute $\hat{g}(y|x)$ and $C_{x,y}$ is published at https://github.com/jacksonloper/markov-link-method, including a tutorial-style ipython notebook detailing every computation made in this paper.

2 Examples where the Markov Link Method may apply

The validity of the Markov Link Method assumption for a given situation should be closely contemplated. Let us consider a few real-world examples where this assumption may apply.

- Quality control for manufacturing. One way to test the reliability of a part is to construct a machine that pushes the part until it breaks. However, how can we test the reliability of the machine that performs the test? In each test run there will be some variability induced by the machine itself, which induces a measurement error. In practice, some kind of assumptions about part homogeneity are used to approximate this error (cf. (3)). However, if we have two testing machines we can use the MLM to obtain a calibration between the machines, even though we can never test the same part with both machines. This enables us to bound the overall measurement error. In this case, ℓ might indicate the type of a part being tested, X would indicate the reliability of a part as measured by one machine, and Y would indicate the reliability

of a part as measured by another machine. If the error in machine Y is not correlated to the part type ℓ, then the MLM assumption certainly holds. Even if the error is correlated, the MLM assumption may still hold. For example, imagine that the Y error is correlated with the absolute reliability of the part; this may pose no problem if that reliability is adequately measured by X.

- 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. (4)). 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. (5)). However, we know that cultured cell-lines look quite different from in-vivo cells (cf. (6; 7)). 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. (8)). 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 (9). 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. (10; 11)), 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 Mathematical Results and Simulations

There are three desiderata we might hope the Markov Link Method estimators would achieve:

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

Interval concentration As we obtain more samples, the interval $C_{x,y}$ should get smaller. Asymptotically it should include nothing but the point $g(y|x)$.

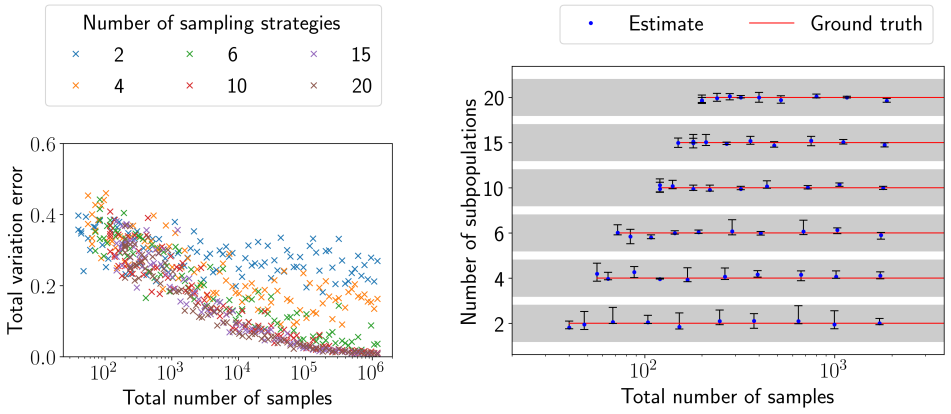
Conservative coverage $g(y|x) \in C_{x,y}$ with high probability.

In many cases, the MLM achieves all three of these conditions. In other cases, we can show that an issue known as 'identifiability' blocks even the possibility of estimator convergence or interval concentration. However, even when interval concentration is impossible, the MLM appears to do as well as it possibly can: the intervals $C_{x,y}$ close around the bounds of what can possibly be learned from the data we observe. We have begun to develop a theory of the relevant issues, which we detail in Appendix A. We give a few of the interesting corner-cases and prove a few surprising results. A complete story remains elusive. 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 selection 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 selection procedures and pick a 'ground truth' link uniformly at random from the parameter space. We then simulate a dataset from this ground truth link, fixing the total number of samples (spreading these samples equally among all combinations of selection 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 total variation distance:

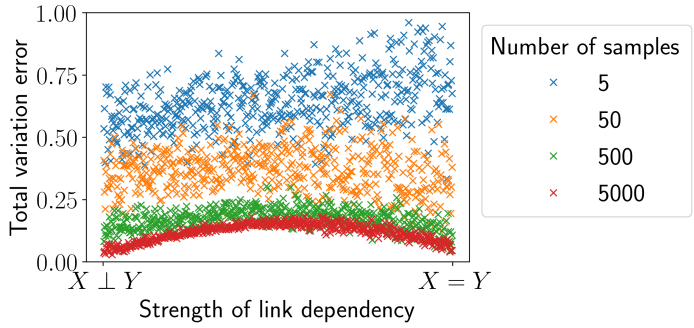
$$\text{Error}(\tilde{q}, g) = \frac{1}{12} \sum_{x=1}^{6} \sum_{y=1}^{6} |\tilde{q}(y|x) - g(y|x)|$$

  This error ranges between zero and one. Zero error indicates that $\tilde{q} = g$. An error of one indicates that the estimate has completely incorrect beliefs about the probability



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

(b) Credible intervals, varying number of samples and number of selection 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, selection procedures, and the link itself. In each trial our 'ground truth' link 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 selection 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 $g(y|x)$ with the MLM interval $C_{x,y}$. We show the interval (in black) centered around the ground truth (in red). Since the link probabilities 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 selection procedures. In subfigure (c) we focus on the case of exactly 4 selection 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 selection procedures may make it impossible to determine the true link; this issue is discussed in detail in Section C.

mass, i.e. $g(y|x) = 0$ 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 selection procedures. This figure also shows that the intervals work correctly regardless of of the number of samples or selection procedures; they include the ground truth with high probability. With many samples and selection procedures, the intervals are small and concentrated around the truth. With fewer samples or selection procedures, the intervals are sometimes forced to be larger. However, our uncertainty is not the same for every aspect of the link. In some cases we are able to obtain a very tight credible interval for one particular value even though the overall estimator error is high.

- How does estimator convergence depend upon the link itself? In each trial of this simulation we use four selection 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 selection procedures. However, in our previous simulations we picked the link uniformly from its parameter space. Now we will be more choosey. On one extreme, we will produce trials where the link makes X, Y independent, i.e. $g(y|x) = g(y|x')$ 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 selection 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 C, 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 a degraded transcriptomic signal, requiring new methods to estimate the cell-type. We refer the reader to (12) 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"

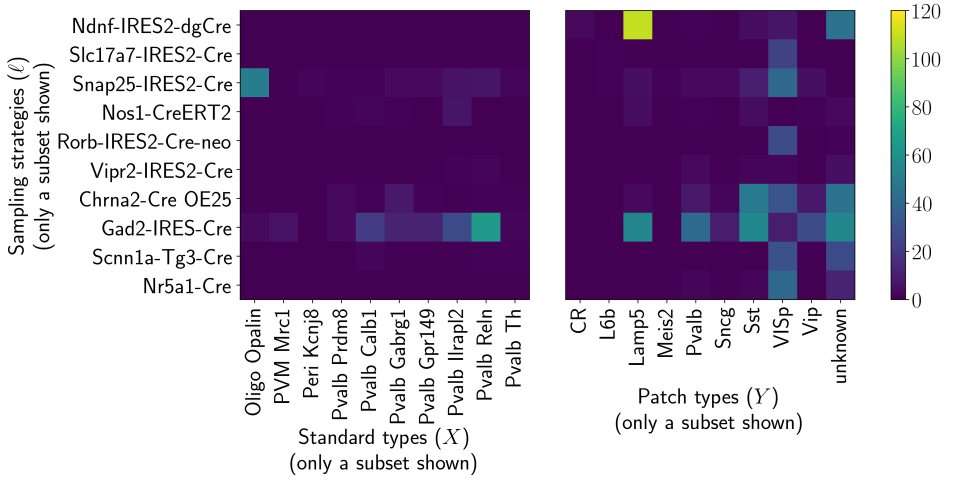


Figure 3: The input to the Markov Link Method: one experiment for each measurement procedure and each selection procedure. Here we show a portion of the real-world data to which we applied the MLM. The Allen Institute gathered neurons from the visual cortex of mice, using a variety of cre/lox-based selection procedure (cf. (12)). 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?"

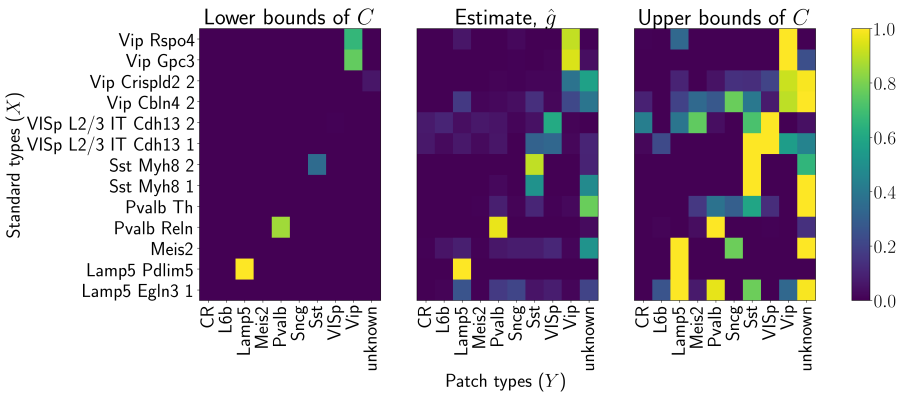


Figure 4: MLM estimation of the link, with credible interval. The central plot shows a portion of the MLM estimator applied to the Allen Institute data. We look at a subselection of Standard types (x) and all of the Patch types (y). For each combination x, y we draw a rectangle whose color indicates the value of the estimator $\hat{g}(y|x)$. 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 aspects of the link 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 $g(y|\text{Meis2})$ could be nearly 1 for many different Patch types. Obviously it cannot be 1 for all of those types simultaneously, since $\sum_y g(y|\text{Meis2}) = 1$, but the data simply doesn't tell us which y carries the mass.

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 selection procedures. Using a cre/lox system (cf. (12)) 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 selection 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 $\hat{g}(y|x)$ and credible intervals C_{xy}. In Figure 4 we visualize these objects for selected values of x, y. Each part of the link has a slightly different story. For example, for the Standard type $x =$'Vip Rspo4' and the Patch type $y =$'Vip,' we have that $C_{x,y} = [.88, 1.0]$. This supports the idea that the true link satisfies $g(y|x) \geq .88$: at least 88% 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, g.

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. (11)). 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. (13)). From the most general point of view, we are engaged in meta-analysis; we refer the reader to (14) 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 (15) for an introduction to what we mean by identifiability. We discuss the identifiability issues for our problem in particular in Appendix C. Our only assumption is that the selection 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 (16) 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 (17). More generally, the physics literature has contributed many key assumptions that bound unidentifiable parameters (cf. (18), (19), and the references therein). Perhaps the closest work to this one would be (20), 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 $q(y|x)$ 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 selection 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 gives provable guarantees, provides 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.

Another direction for future work would be to take a frequentist point of view. An obvious direction would be to use profile likelihoods as statistics to define confidence intervals for the link. However, the distribution of these statistic is difficult to pin down. Traditional techniques based on asymptotic normality may fail dramatically, because the true link may lie at the boundary of the parameter space. In particular, there may be some x, y for which $g(y|x) = 0$. Even approximate methods are difficult to apply because the parameter space can be quite high-dimensional. Perhaps future work could uncover a way to overcome these challenges.

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, we can bound our uncertainty. By probing this uncertainty carefully, we can learn what the data actually has to say and what experiments will help us learn more.

References

[1] Nathan W Gouwens, Staci A Sorensen, Jim Berg, Changkyu Lee, Tim Jarsky, Jonathan Ting, Susan M Sunkin, David Feng, Costas Anastassiou, Eliza Barkan, et al. Classifica-

tion of electrophysiological and morphological types in mouse visual cortex. *bioRxiv*, page 368456, 2018.

[2] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

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

[4] 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.

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

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

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

[17] 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.

[18] 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.

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

[20] 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.

[21] Jyrki Kivinen and Manfred K Warmuth. Additive versus exponentiated gradient updates for linear prediction. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pages 209–218. ACM, 1995.

A Exact details of the Markov Link Method

Consider experiments yielding an $\Omega_\ell \times \Omega_X$ matrix $\mathscr{D}_X$ and an $\Omega_\ell \times \Omega_Y$ matrix $\mathscr{D}_Y$, carrying the distribution

$$(\mathscr{D}_{X\ell 1}, \mathscr{D}_{X\ell 2} \cdots \mathscr{D}_{X\ell\Omega_X}) \sim \text{Multinomial}(n_\ell, f(\cdot|\ell))$$
$$(\mathscr{D}_{Y\ell 1}, \mathscr{D}_{Y\ell 2} \cdots \mathscr{D}_{Y\ell\Omega_Y}) \sim \text{Multinomial}(m_\ell, h(\cdot|\ell))$$

where $f(x|\ell), h(y|\ell)$ are conditional distributions and there is some $g(y|x)$ such that $h(y|\ell) = \sum_x f(x|\ell)g(y|x)$. This is simply a restatement of the assumptions we have made throughout this paper about how our objects of interest (f, g, h) are related to the data we can observe $(\mathscr{D}_X, \mathscr{D}_Y)$.

The purpose of the MLM is to make estimates about g using the data $\mathscr{D}_X, \mathscr{D}_Y$. Unfortunately, g cannot be directly determined from the data. Even perfect knowledge of f, h may be insufficient to determine the true value of g. Some examples are detailed in Appendix C. This problem is called 'nonidentifiability,' and it can have some troubling consequences. For example, standard Bayesian analyses applied to nonidentifiable parameters will be extremely sensitive to the precise choice of our prior beliefs. Even with infinite data, the prior beliefs may have a significant impact on inferences. To avoid these difficulties, we focus on objects that we know we can identify from data. In particular, we will look at lower bounds, upper bounds, and something in-between.

Let $\Theta(f,h) = \{g : h(y|\ell) = \sum_x f(x|\ell)g(y|x)\}$ denote the set of links which are consistent with f, h and the Markov Link Method assumption. We define

- $g_{\text{lo},f,h}(y|x) \triangleq \min_{g\in\Theta(f,h)} g(y|x)$
- $g_{\text{hi},f,h}(y|x) \triangleq \max_{g\in\Theta(f,h)} g(y|x)$
- $\hat{g}_{f,h} \triangleq \arg\min_{g\in\Theta(f,h)} D_f(\text{Uniform}||g)$, where D_f is some f-divergence.[1]

Even if g is nonidentifiable, we can still be assured that $g_{\text{lo}}, g_{\text{hi}}$ are identifiable and $q_{\text{lo}}(y|x) \leq g(y|x) \leq g_{\text{hi}}(y|x)$. The estimator $\hat{g}$ is also identifiable and we can also hope it will strike a middle ground. In producing this single point estimate we had to decide how to deal with the fundamental fact that actually any $g \in \Theta$ might be correct. At a basic level, we could make two kinds of mistakes. We might claim a very strong association between the measurement procedures even though actually there is none. We might claim a very weak association even though actually there is a strong association. We choose to err on the side

[1] In practice, we choose a χ^2 divergence because it makes the minimization problem a quadratic program; this makes it particularly easy to solve. See Appendix B for details.

of asserting weak associations, by choosing the g which is as close as possible to uniform. We made this choice in the spirit of the Maximum Entropy Principle, i.e. that in the absence of other information we assume X is associated with each Y equally. This is perhaps as reasonable as any way to pick a particular $\hat{g}$. However, we reiterate that $\hat{g}$ is just one possibility among many. It is safest to consider the full spectrum of possibilities by looking at the extremes $g_{\text{lo}}, g_{\text{hi}}$.

If we had perfect knowledge of f, h, the objects $g_{\text{lo},f,h}, g_{\text{hi},f,h}, \hat{g}_{f,h}$ would give us a reasonable understanding of what we can know about the link g. However, in practice we do not have access to f, h. Instead, we have access to the data $\mathscr{D}_X, \mathscr{D}_Y$ which enables us to estimate f, h. To account for our uncertainty about these estimates, we take a Bayesian perspective. For prior beliefs about f, h, we take a noninformative uniform prior $\mathbb{P}$:

$$\mathbb{P}(f, h) \propto 1$$

Following the Bayesian philosophy, we then incorporate new knowledge by conditioning. We have two important pieces of knowledge about f, h. First, we have observed the data, $\mathscr{D}_X, \mathscr{D}_Y$. Second, we know from the MLM assumption that there exists *some* value g such that Equation (1) holds. We would like to condition on both of these facts. However, due to the Borel-Kolmogorov paradox, 'conditioning on the MLM assumption' is not a meaningful idea. Instead, it is necessary to define a variable indicating how much the MLM assumption fails, and condition on this variable being zero. In particular, let $D(h||h')$ denote the Kullback-Leibler divergence and $\Gamma(f, h) = \min_{h':\ \Theta(h,h')\neq\emptyset} D(h||h')$. Our posterior uncertainty about f, h can then be articulated through the distribution

$$\mathbb{P}\left(f, h | \mathscr{D}_X, \mathscr{D}_Y, \Gamma(f, h) = 0\right)$$

In terms of this posterior, we define our final point estimate $\hat{g}$ and uncertainty bounds C as follows:

1. $\hat{g}$ is calculated using posterior expectation:

$$\hat{g} \triangleq \mathbb{E}\left[\hat{g}_{f,h} | \mathscr{D}_X, \mathscr{D}_Y, \Gamma(f, h) = 0\right]$$

2. $C_{x,y}$ is calculated as credible intervals. For each x, y, we define $C_{x,y}$ to be the largest interval such that

$$\mathbb{P}\left(g_{\text{lo},f,h}(y|x) \in C_{x,y} | \mathscr{D}_X, \mathscr{D}_Y, \Gamma(f, h) = 0\right) \geq 97.5\%$$
$$\mathbb{P}\left(g_{\text{hi},f,h}(y|x) \in C_{x,y} | \mathscr{D}_X, \mathscr{D}_Y, \Gamma(f, h) = 0\right) \geq 97.5\%$$

In practice, we were not able to find a way to compute these objects exactly. Given samples from the posterior distribution $\mathbb{P}(f, h | \mathscr{D}_X, \mathscr{D}_Y, \Gamma(f, h) = 0)$, it would be straightforward to get good estimates. As seen in Appendix B, it is straightforward to compute $g_{\text{lo}}, g_{\text{hi}}, \hat{g}$ from samples of f, h, so we could use use monte-carlo approximations for our objects of interest. Unfortunately, it seems difficult to obtain samples from this posterior distribution. 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 $\Gamma(p, h) = 0$ that formalizes the MLM assumption. We instead take a somewhat naïve approach. We start by drawing samples according to

$$F, H \ \sim \mathbb{P}(f, h | \mathscr{D}_X, \mathscr{D}_Y)$$

This can be achieved exactly, using the the conjugacy between our prior and the Multinomial distribution. Notice that these samples do not incorporate our knowledge of the MLM assumption, insofar as they are not conditioned on the event $\Gamma(f, h) = 0$. To approximately remedy this, we define $\tilde{H}$ as the solution of $\min_{h'} D(H|h')$, subject to the constraint that

$\Gamma(F,h') = 0$. Optimization details can be found in Appendix B. We use the pair $F,\tilde{H}$ as *approximate* samples for the distribution $\mathbb{P}(f,h|\mathscr{D}_X,\mathscr{D}_Y,\Gamma(f,h)=0)$. We can repeat this process to produce many samples of $(F,\tilde{H})$ and use those samples to produce approximate monte-carlo estimates for $\hat{g}(y|x), C_{x,y}$. We asymptotically expect that F,H will nearly satisfy the MLM assumption in any case, so this approximation should not make a large difference. For example, on the Allen Institute data we found that the total variation distance between $H(\cdot|\ell)$ and $\tilde{H}(\cdot|\ell)$ was about 15% (averaging over all selection procedures ℓ and various samples of H). For comparison, this is about three times smaller than the average total variation distance between $H(\cdot|\ell)$ and $H(\cdot|\ell')$ for a randomly selected pair of selection procedures (ℓ,ℓ'), which averages out to around 50%. In any case, it is a practical solution to a difficult problem.

B 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 values for F,H. One step in the MLM involves projecting H to the set of distributions which are consistent with F 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

$$\min_h D(H|h)$$

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

$$\max_g \sum_{\ell,y} H(y|\ell)\log\left(\sum_x F(x|\ell)g(y|x)\right)$$

   Taking derivatives one can readily show that this problem is convex. We solve it using exponentiated gradient ascent (cf. (21)). We initially guess that g is uniform. We then repeatedly make the updates

$$g(y|x) \propto g(y|x)\sum_\ell F(x|\ell)\frac{H(y|\ell)}{\sum_x F(x|\ell)g(y|x)}$$

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

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

$$\min_{g\in\Theta(f,h)} g(y|x)$$

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

3. Quadratic programming. To obtain the minimum χ^2 divergence to uniform, the MLM also requires us to solve quadratic optimization problems within Θ:

$$\min_{g \in \Theta(f,h)} \sum g(y|x)^2$$

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

C 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. Here we will use the notation introduced in Appendix A.

First note that we can obtain arbitrarily good estimates of f, 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 f, h. Even so, the data does not necessarily tell us the value of the link g. There may be many possible links, g, which are all equally consistent with f, h. That is, we may have g_1, g_2 such that $h(y|\ell) = \sum_x f(x|\ell) g_1(y|x) = h(y|\ell) = \sum_x f(x|\ell) g_2(y|x)$. Both links yield 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 g might be.

We will now look at some examples:

C.1 A simple failure case

Consider the case that $\Omega_\ell = \Omega_Y = 2$ and $\Omega_X = 3$. That is, there are 2 separate selection procedures, tool I recognizes 3 categories and tool II recognizes 2 categories. In particular, let us imagine that $f(x|\ell) = A_{\ell x}$ 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 selection procedures and columns correspond to a different measurement outcome. Now let $g(y|x) = C_{xy}$, 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 $\sum_y g(y|x) = 1$, 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 g that we simply cannot resolve. For example, here are two choices of C which are both consistent with the equation $A \times C = B$:

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

C.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

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

and $f(x|\ell) = A_{\ell,x}$, 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 $f(x|\ell)$ and $h(y|\ell) = \sum_x f(y|\ell)g(y|x)$. Notice that due to the simple structure of g we obtain $h(y|\ell) = A^{\ell,y}$. However, let us imagine we know nothing about the true value of g.

How much can we say about g, if we only had knowledge of f and h? On the one hand, we observe that in the absence of any other constraints, the object g has 2^{2k-3} degrees of freedom. This is because there are 2^{k-1} values of ℓ and for each subpopulation $g(\cdot|\ell)$ 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 $\sum_y g(y|x) = 1$. 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 $\sum_x f(y|\ell)q(y|x) = h(y|\ell)$. Indeed, we have *exponentially many* more degrees of freedom than we have constraints.

However, the reality is that q is exactly determined by f, h. This is possible because there are inequality constraints which also govern q, namely $g(y|x) \geq 0$. 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 f, h be as they are defined above. Then there is exactly one g that is consistent with f, h and the Markov Link assumption. That is, g is the only possible value satisfying*

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

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

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

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

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

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

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

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 $g(y|x)$ is uniquely determined for the first 2^{k-2} values of x, y. Moreover, since $\sum_{y \leq 2^{k-2}} g(y|x) = 1$, we see that g must also satisfy $g(y|x) = 0$ for $y > 2^{k-2}$ and $x \leq 2^{k-2}$. Thus we have seen that g is uniquely identified 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 $\sum_{x \leq 2^{k-2}} g(y|x) = 0$ for each $y > 2^{k-2}$. We can then use the same reasoning to obtain that g is uniquely identified for the remaining values of x, y. □

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

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

C.3 Discontinuity

Consider the case that $\Omega_\ell = 1$ and $\Omega_X = \Omega_Y = 2$. That is, there is only one selection 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
   - $\mathbb{P}(X = 1) = f(1) = 0$
   - $\mathbb{P}(X = 2) = f(2) = 1$
   - $\mathbb{P}(Y = 1) = h(1) = 0$
   - $\mathbb{P}(Y = 2) = h(2) = 1$

   In this case the MLM assumption $\sum_x f(x)g(y|x) = h(y)$ can be used to prove that $g(1|2) = 0, g(2|2) = 1$, but we now have *absolutely no* knowledge of $g(1|1), g(2|1)$. 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 $g(y|x)$ for $x = 1$.
2. Now let us take a slight variation:
   - $\mathbb{P}(X = 1) = f(1) = 0.01$

- $\mathbb{P}(X=2)=f(2)=0.99$
- $\mathbb{P}(Y=1)=h(1)=0$
- $\mathbb{P}(Y=2)=h(2)=1$

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

3. Now we take yet another slight variation:
   - $\mathbb{P}(X=1)=f(1)=0.01$
   - $\mathbb{P}(X=2)=f(2)=0.99$
   - $\mathbb{P}(Y=1)=h(1)=0.01$
   - $\mathbb{P}(Y=2)=h(2)=0.99$

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

The disturbing thing about this example is that by making infinitesimal perturbations to f 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 f,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 g is bounded away from zero i.e. $g(y|x)>c$ for every x,y for some $c>0$. If this holds, together with a certain linear independence assumption, we can guarantee that the true g is close to the set $\Theta(\hat{f},\hat{h})$ when $\hat{f},\hat{h}$ are good approximations to f,h, where

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

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

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

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

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

Since it is easy to find consistent estimators for f,h, this theorem suggests we can use those estimators to get good estimates for our uncertainty about g. In particular, subject to the conditions of the theorem, we have that the bounds of $C_{x,y}$ converge to the bounds of what is identifiable about g. It is certainly a step in the right direction, but we emphasize

that the theorem's conditions are nontrivial. It is our opinion that the linear independence condition is fairly mild (if it is not met then subpopulations can simply be merged together). However, the positivity condition is quite troubling. In many cases of interest the true link g 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 ($g(x|y) > c > 0$) is much stronger than is necessary. A precise understanding of the discontinuity example above has remained elusive. Better understanding could lead to more accurate estimates. We leave it for future work.