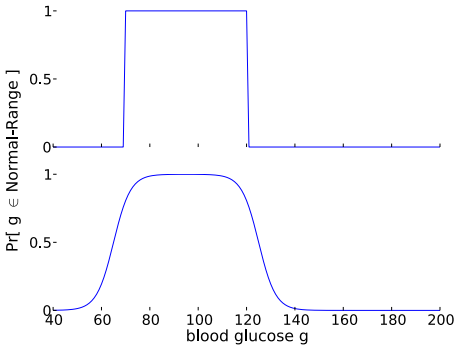
Causal Inference under Uncertainty via Adjustments and SOPDs

Dylan Hutchison and Samantha Kleinberg
Stevens Institute of Technology

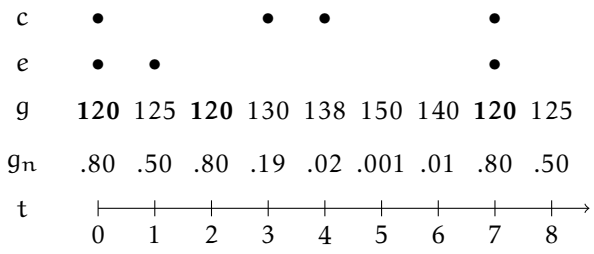
One challenge for inferring causal relationships is that the observations from which probabilities are calculated are often uncertain. A patient's medical record, for instance, has omissions, incorrect measurements and outdated information so making assumptions such as that patients only take aspirin when explicitly mentioned will lead to incorrect inferences about their risk of heart attack. More generally, the resulting errors in estimating probabilities will lead to false discoveries or non-discoveries in causal inference.

Our solution is to assign less weight to uncertain values, so that conclusions like conditional independence are weaker when based on error-prone measurements instead of reliable ones. We propose a new approach based on second order probability distributions (SOPDs) and Bayesian updating to allow observation-specific uncertainty levels that can be determined through background knowledge (e.g. device error rates, logs) or experiment (multiple variable measurements). Here an observation's impact is given by a prior distribution that is updated with meta-information about the observation, resulting in a shift to the distribution's center (most likely value) and spread (degree of certainty).

This framework addresses uncertainty generally, but we focus here on how it can be used for discretization and to improve causal inference. Discretization involves binning continuous variable measurements into a collection of discrete states, which can be useful for reducing noise. For instance, blood glucose measurements are often categorized into {normal, too-high, too-low}, where values in [70,120] are considered normal. Rather than adopt this strict partition, which is often within a device's margin of error and seldom used in this way by patients and clinicians, we create "fuzzy" partitions with probability distributions. Traditional and new discretization functions are shown below left. Below right is a set of corresponding timeseries data with discrete variables c (carb-heavy meal) and e (high-level of exercise), and the continuous variable g (blood glucose measurement). The traditional approach classifies bold g values as normal. Our approach maps each g value to the indicated probability of normalcy g_n.



P(g=normal)



Sample data

We motivate the need to account for uncertainty with a simple example: does exercise affect glucose levels in one time unit? We use an algorithm developed by Kleinberg that finds the average difference a cause makes to the probability of an effect by holding fixed other variables. Using traditional discretization, the effect of e on g_n in one time unit is:

$$P(g_n|c,e) - P(g_n|c,\neg e) = 0/2 - 0/2 = 0. \quad (1)$$

Here it seems that exercise does not regulate glucose, since values close to normal (125) following $c \wedge e$ are categorized as abnormal and indistinguishable from the higher values following $c \wedge \neg e$ (138 and 150).

Instead, our approach uses the distributions to assign a probability of g_n given the actual value of g. We assign base weights to data values according to certainty in their observation, akin to adjusting sample data

base weights based on data representativeness and nonresponse. The effect of e on g_n in one time unit is then:

$$\begin{aligned} P(g_n|c,e) - P(g_n|c,\neg e) =& \frac{0.50+0.50}{1+1} - \frac{0.02+0.001}{1+1} \\ =& 0.50 - 0.01 = 0.49. \end{aligned} \tag{2}$$

Now e correctly has a potentially significant impact on g. The base weights (real-valued numbers $\in [0,1]$, as in equation (2)) hold more information than binary indicators (0 or 1, as in equation (1)) of whether an event occurred. One could also account for uncertainty in c and e by similarly mapping raw values for meal carbohydrate content and exercise intensity to probabilities.

To determine the base weight for events in general we maintain SOPDs and update them as new information arrives. An SOPD represents belief (alternatively, uncertainty) in the first order probabilities that a variable has a particular value. Thus we have an SOPD for each possible glucose value. One for $g = 130$ is shown in figure 1, where the x-axis corresponds to probabilities that $g = 130$ indicates normal glucose, and the y-axis to belief (in the form of a probability distribution) that the corresponding x-axis probability is the true probability. The most likely value of the graph (mean of the SOPD) is low, reflecting that a $g = 130$ is usually considered high. However, if we know that for a particular reading the subject dropped his test strip on the ground before inserting the sample into his glucometer, we can account for this contamination. Assuming ground materials could lead to any blood glucose reading uniformly through their interaction with the blood glucose meter, we represent the new information with the uniform data distribution.

To form the posterior distribution we take the weighted sum of the prior and data distributions, assigning weight to the prior corresponding to how strongly we trust the new data. If we know the blood glucose meter has high sensitivity to contaminants, we would assign higher weight to the prior, because we cannot trust the data as much as data from a more robust device. Supposing a confidence of 0.3, we form the posterior distribution shown between the prior and data distributions in figure 1. The end result is a "weakening" of the observation base weight via the flattening of the SOPD. Compared with the prior, the posterior reflects our higher uncertainty via increased distribution spread and our stronger most likely belief that the 130 reading could be normal by pushing the distribution center toward 0.5, indifference between a value being normal/abnormal.

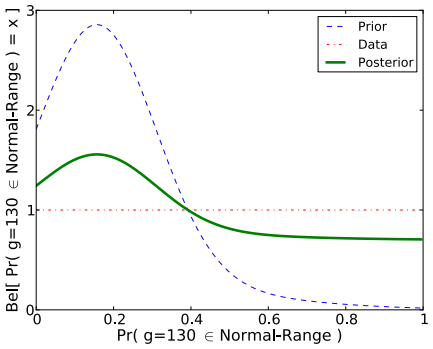


Figure 1: SOPD formulation for g=130

In general, we call our procedure for updating base weights *adjustment*, as the *center* and *spread* of data is shifted in response to meta-information. Adjustments to the center are critical to causal inference accuracy, whereas adjustments to the spread reflect inference precision. Precision is important in belief updating (the propensity to change beliefs in response to new information) and risk analysis (supplementing inferences with sensitivity analysis and confidence intervals). Beliefs formed from comprehensive, irrefutable evidence have smaller spread and therefore more resistance to change than beliefs formed from speculatively extrapolating a handful of data points.

Our proposed approach delivers detailed uncertainty representations that result in more accurate causal inference by representing beliefs with SOPDs, which can be manipulated through adjustments. We highlighted the problem of discretization and measurement error, but more general adjustments discussed in the full paper will address incorrect, incomplete, and outdated data as well as an application of the approach to actual diabetic physiological data. A key area for future work is automating information translation into SOPDs and proving robustness against minor translation errors.