CONDITIONAL PROBABILITIES FOR A WORLD MODEL

Adarsh Pyarelal, *Clayton Morrison*, *Kobus Barnard*

January 7, 2019

Abstract

The level of food insecurity in a region is a complex function of economic, political, and environmental factors, whose temporal evolution can be modeled using a dynamic Bayesian network. The transition probabilities associated with the latent states in such a network can in principle be inferred from sentences extracted from relevant literature through machine reading. However, the interpretation of these sentences involves a great deal of uncertainty. In this paper, we present a principled approach for dealing with this uncertainty. The transition model itself is taken to be a random variable, and we construct its distribution empirically using evidence sentences and crowdsourced responses about the magnitude of gradable adjectives.

CONTENTS

1 MODEL CONSTRUCTION 2
- 1.1 Introduction . . . 2
- 1.2 Causal Dynamic Bayes Networks . . . 2
- 1.3 A simple causal linear dynamical model . . . 6
- 1.4 Going beyond two nodes . . . 7
- 1.5 Implications of the simple linear model . . . 9
- 1.6 Grounding gradable adjectives . . . 10
- 1.7 Constructing empirical probability densities . . . 12
- 1.8 Prior . . . 14
- 1.9 Emission model . . . 15
- 1.10 Relaxing our assumptions . . . 16

2 INFERENCE 17

3 SOFTWARE CONTRIBUTION 18

BIBLIOGRAPHY 19

1 MODEL CONSTRUCTION

1.1 INTRODUCTION

Food insecurity, defined in [1] as the 'limited or uncertain availability of nutritionally adequate and safe foods or limited or uncertain ability to to acquire acceptable foods in socially acceptable ways', is affected by many factors that interact in complex ways. For example, basic microeconomic theory implies an inverse relationship between crop yields and crop prices. Crop yields can in turn be affected by precipitation levels, fertilizer usage, pollution and technological advances, while crop prices can be affected by subsidies.

The complexity of these interactions makes it challenging to make predictions about food insecurity in the near future. The current state of the art for making such predictions involves comprehensive analyses prepared by domain experts. Typically, there is a tradeoff between making an analysis comprehensive (i.e., considering all the relevant factors), quantitative, and timely. Analyses that are both quantitative and comprehensive can take on the order of two years to complete [6]. However, the time scale of fluctuations in the state of food security is much shorter, on the order of months. There is therefore an urgent need for tools that reduce the time it takes to prepare these analyses so that timely recommendations can be made and appropriate interventions performed. The development of these tools is the goal of DARPA's World Modelers program [9].

1.2 CAUSAL DYNAMIC BAYES NETWORKS

Causal Analysis Graphs

The first step in the automation of these analyses is the extraction of information from relevant text sources, such as reports and news articles. Within these sources, we are particularly interested in identifying sentences that describe *causal* relationships between entities, of the form

> A small increase in X causes a large decrease in Y. (e_1)

Here, X and Y are factors affecting food insecurity. In this sentence, for example, X might represent the yield of a certain crop yield, while Y might represent its price.

Collecting multiple sentences like e_1 will allow us to build a *causal analysis graph* - a graph where the nodes represent entities and the directed edges represent the causal relationships between them. For example, the evidence sentence above might map to a directed edge in

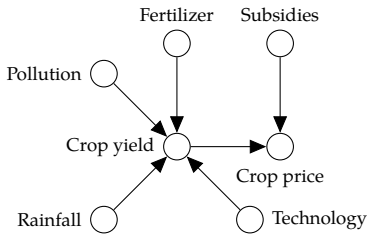

Figure 1.1: An example of a causal analysis graph that could potentially model some of the factors that affect food security.

a causal analysis graph going from node X to node Y, and the factors and relationships mentioned in § 1.1 could potentially be encoded in the causal analysis graph in figure 1.1.

Bayes Networks

A natural way to obtain a quantitative model from a causal influence network is to map it to a *Bayes network*. A Bayes network is a graphical representation of a probability model, with the nodes representing random variables [1] and the link structure representing how the joint probability can be factored into a product of conditional probabilities.

The process of inferring the link structure of such a Bayes net from a collection of evidence sentences is distinct from the process of specifying the conditional probability distributions it represents. In this paper, we will focus on the latter. The notation that we use in the rest of the document is as follows.

- Lowercase bold letters represent vectors (e.g. $\mathbf{v}$).
- Uppercase bold letters represent matrices (e.g. $\mathbf{M}$).
- Subscripts indicate the discrete time slice corresponding to a quantity (e.g. $\mathbf{v}_n$ is $\mathbf{v}$ at time slice n.)
- Factors that affect food insecurity, such as crop yields, subsidies, etc, are denoted by italicized capital letters (e.g. X)
- A dot above a factor represents its partial derivative with respect to time:

$$\dot{X} = \frac{\partial X}{\partial t}. \tag{1.1}$$

- A vertical line with a subscript, placed to the right of a derivative, denotes that the derivative is evaluated at the time slice corresponding to that subscript. For example,

$$\left.\frac{\partial X}{\partial t}\right|_n \tag{1.2}$$

denotes $\frac{\partial X}{\partial t}$ evaluated at time slice n.

Dynamic Bayes Networks

While the example in figure 1.1 captures the causal relationships between the factors, it does not include the temporal dynamics needed to make predictions about them. The temporal evolution of the factors can be modeled by treating each observation of a random variable associated with a factor as a distinct random variable indexed by a discrete time index and conditioned on one or more *latent* variables. This kind of model is generally known as a *state space model* or *dynamic Bayesian network*[2]. Some well-known examples of such models include Hidden Markov Models (HMMs) for discrete latent variables and Linear Dynamical

[1] However, it is important to note that Bayes networks are constructs that are general enough to incorporate deterministic relations as well, such as systems of partial differential equations.

[2] At this point it bears emphasizing that the structure of the causal analysis graph is *not* the same as the structure of the Bayes network constructed from it. The former can have cycles, while the latter cannot.

Systems (LDSs) for continuous latent variables with linear-Gaussian distributions. At first glance, our problem seems like a good fit for an LDS, since many of our factors (prices, precipitation, etc.) are continuous quantities.

Review of Linear Dynamical Systems

Figure 1.2 shows an example of an LDS (example taken from [2]). A node labeled $\mathbf{s}_n$ represents the latent state of the system at time slice n, and a node labeled $\mathbf{o}_n$ represents an observation in the same time slice, conditioned on $\mathbf{s}_n$. The conditional probability distributions governing this system are defined by the relations:

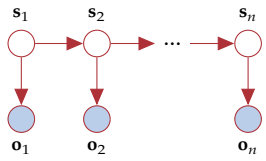

Figure 1.2: A Bayes network representing a linear dynamical system.

$$p(\mathbf{s}_n|\mathbf{s}_{n-1}) = \mathcal{N}(\mathbf{s}_n|\mathbf{A}\mathbf{s}_{n-1}, \mathbf{\Gamma}) \quad \text{(Transition probability)} \quad (1.3)$$

$$p(\mathbf{o}_n|\mathbf{s}_n) = \mathcal{N}(\mathbf{o}_n|\mathbf{C}\mathbf{s}_n, \mathbf{\Sigma}) \quad \text{(Emission probability)} \quad (1.4)$$

$$p(\mathbf{s}_1) = \mathcal{N}(\mathbf{s}_1|\boldsymbol{\mu}_0, \mathbf{V}_0) \quad \text{(Initial value)} \quad (1.5)$$

where the following notation has been used:

- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{\Sigma})$ denotes a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$.
- $\mathbf{A}$ represents the linear transition model.
- $\mathbf{C}$ represents the linear emission (sometimes referred to as observation) model.
- $\boldsymbol{\mu}_0$ and $\mathbf{V}_0$ denote the mean vector and covariance matrix for the distribution of the initial latent variable, $\mathbf{s}_1$.

The parameters of the model are collectively denoted by $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \{\mathbf{A}, \mathbf{\Gamma}, \mathbf{C}, \mathbf{\Sigma}, \boldsymbol{\mu}_0, \mathbf{V}_0\} \quad (1.6)$$

As indicated above, all the probability distributions in an LDS are Gaussian. Linear dynamical systems are often used to describe physical processes where the transition model is well understood. For example, the transition model that describes the evolution of the latent state of an object (comprised of its position and momentum vectors) that is moving at nonrelativistic speeds is fixed by Newton's laws of motion, modulo noise that arises from environmental factors (such as Brownian motion) and the limitations of the measurement apparatus.

In contrast, there are no models that describe abstract concepts like food security or conflict that are well understood and have undergone over three hundred years of testing. There are a number of sources of uncertainty inherent in constructing a model from textual evidence - for example, it is unclear which particular linear transition model $\mathbf{A}$ we should use, or whether a linear dynamical system is even a good choice for modeling the system in the first place. Furthermore, we have

no *a priori* reason to assume that the emission probabilities (1.4) are Gaussian.

Finally, for explanatory causal modeling, we do not necessarily want to absorb model uncertainty into stochastic state change (the stochastic part of (1.3)), since we will also be interested in inferring the model itself [5].

Below, we discuss potential sources of uncertainty involved in constructing models from text in more detail.

- The writer of a phrase such as: *"A small increase in X"* likely has a range of increases in X that they would describe as 'small', not to mention the possibility that this range might overlap with ranges corresponding to similar adjectives, such as 'little', or 'weak'.
- Different writers will disagree about the range of values that qualify as *small*.
- If we have only one sentence about the relationship between two quantities X and Y, such as *"A small increase in X causes a large decrease in Y"*, there is some uncertainty involved in extrapolating from this sentence to making predictions about things like the effect of a small *decrease* in X (which the sentence does not mention).
- Given multiple sentences about the same pair of quantities that can differ in their assessment of
  - the magnitude, e.g.:
    - *A small increase in X causes a large increase in Y.*
    - *A tiny increase in X causes a huge increase in Y.*
  - and the polarity, e.g.:
    - *A small increase in X causes a large decrease in Y* of the change in the quantities,
    - *A small increase in X causes a large increase in Y* of the change in the quantities,

  of the changes, it is uncertain how to combine and reconcile these differences.

Thus, instead of assuming Gaussian transition probabilities (we will elide the details of the emission probabilities for now, returning to them in § 1.9), we will take the *transition model* itself to be a random variable, $\boldsymbol{\theta}_T$ (similar to the definition in in (1.6), but now incorporating the distributions in addition to the parameters). This can be represented by the Bayes net in figure 1.3. In addition to the model $\boldsymbol{\theta}_T$, we also depict the text data (the collection of evidence sentences) by a shaded node labeled D.

In the next section, we will examine a simplified system with only two factors and a single evidence sentence about the relationship between them. We will begin by first treating the model as a known linear model, and then integrate some of the uncertainties from the list above.

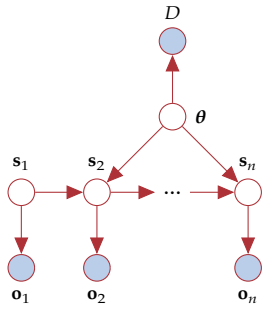

Figure 1.3: A Bayes network representing a dynamical system, but now with the transition probabilities sampled from a distribution.

1.3 A SIMPLE CAUSAL LINEAR DYNAMICAL MODEL

Consider two quantities X and Y that form a dynamical system, i.e. they are functions of each other and of time. A first approximation towards modeling this system involves treating the functions that describe their time evolutions as linear in the quantities and their partial derivatives w.r.t time:

$$\mathbf{s}_n = \mathbf{A}\mathbf{s}_{n-1}. \tag{1.7}$$

where $\mathbf{s}^T = (X, \dot{X}, Y, \dot{Y})$. Now, let us treat the sequence $\mathbf{s}_1, ..., \mathbf{s}_n$ as a sequence of latent variables in a dynamical Bayes net with the same structure as the one depicted in figure 1.2. The entries of the matrix $\mathbf{A}$ define the behavior of the dynamical system, and so we expect to fill them in based on the evidence sentences that we read. In the absence of any evidence sentences relating X and Y, we expect $\mathbf{A}$ to have the following structure:

$$\mathbf{A} = \begin{pmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{(In the absence of any evidence sentences).} \tag{1.8}$$

Now, however, suppose we run our machine reader over our corpus of literature and obtain exactly one sentence relating X and Y - the sentence e_1 introduced in § 1.2. Sentences such as this one suggest a causal relation between changes (or the rates thereof) in the quantities X and Y. Let us examine this sentence and what it might imply for $\mathbf{A}$.

Suppose we are given X_n, Y_n, and $\dot{X}_n$, with discrete time slices separated by Δt. We are interested in predicting Y_{n+1}. For a small enough time step Δt, we can approximate Y_{n+1} as follows:

$$\begin{aligned} Y_{n+1} &\cong Y_n + \left.\frac{dY}{dt}\right|_n \Delta t \\ &= Y_n + \left(\left.\frac{\partial Y}{\partial t}\right|_n + \left.\frac{\partial Y}{\partial X}\frac{\partial X}{\partial t}\right|_n \right) \Delta t \\ &= Y_n + \left(\dot{Y}_n + \left.\frac{\partial Y}{\partial X}\right|_n \dot{X}_n \right) \Delta t \end{aligned} \tag{1.9}$$

The term $\dot{Y}$ represents the *intrinsic* tendency of Y to change with time, that is, due to factors that are not explicitly specified elsewhere in our model. For example, the price of crops might have a natural upward trend over time due to inflation, but inflation may not be explicitly

represented as a factor for a particular model. For the purpose of this illustration, however, we will assume that

$$\frac{\partial Y}{\partial t} = 0. \tag{1.10}$$

which leaves us with

$$Y_{n+1} = Y_n + \left.\frac{\partial Y}{\partial X}\right|_n \dot{X}_n \Delta t \tag{1.11}$$

We can interpret the evidence sentence as telling us something about the form of $\partial Y / \partial X$. The simplest possible approach would be to take this quantity to be independent of X and t:

$$\frac{\partial Y}{\partial X} = \beta_{XY} \tag{1.12}$$

which changes (1.11) to:

$$Y_{n+1} = Y_n + \beta_{XY} \dot{X}_n \Delta t \tag{1.13}$$

Thus, given some ΔX, we would predict $\Delta Y = \beta_{XY} \Delta X$. Given the nature of the evidence sentence (an increase in X causes a decrease in Y), β_{XY} must be negative. Using (1.11) and (1.12), we can fill in some of the off-diagonal entries of $\mathbf{A}$:

$$\mathbf{A} = \begin{pmatrix} 1 & \Delta t & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \beta_{XY} \Delta t & 1 & \Delta t \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{(With evidence sentence, (1.11), and (1.12))} \tag{1.14}$$

So far, we have been treating our model $\boldsymbol{\theta}_T$ (consisting of the update rule in (1.7) and the parameter $\mathbf{A}$) as a deterministic parameter. Now, we will treat it as a random variable instead.

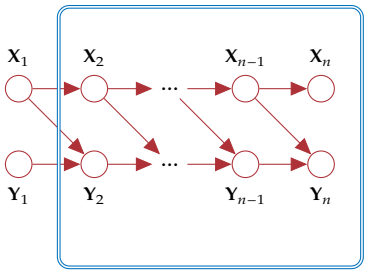

Figure 1.4: A DBN representing the latent state transitions for our example sentence. The box with the double-line border represents the fact that the model describes a *distribution* over sequences.

1.4 GOING BEYOND TWO NODES

Consider the causal analysis graph depicted in figure 1.5. Suppose we wish to model the time evolution of the random variable Z. Similar to our expression in (1.9), we would have

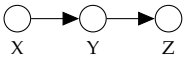

Figure 1.5: A 3-node causal analysis graph, with a chain-like structure.

$$Z_{n+1} \cong Z_n + \left.\frac{dZ}{dt}\right|_n \Delta t \tag{1.15}$$

If we interpret the CAG structure to mean that Z is a function of Y, which is in turn a function of X, the total derivative $\frac{dZ}{dt}$ further simplifies as follows:

$$\begin{aligned}\frac{dZ}{dt} &= \frac{\partial Z}{\partial t} + \frac{\partial Z}{\partial Y}\frac{dY}{dt} \\ &= \dot{Z} + \beta_{YZ}\left(\frac{\partial Y}{\partial t} + \frac{\partial Y}{\partial X}\frac{dX}{dt}\right) \\ &= \dot{Z} + \beta_{YZ}\left(\dot{Y} + \beta_{XY}\frac{\partial X}{\partial t}\right) \\ &= \dot{Z} + \beta_{YZ}\left(\dot{Y} + \beta_{XY}\dot{X}\right)\end{aligned} \tag{1.16}$$

Intuitively, a desirable feature of our model would be to have changes in X cause changes in Z, by inducing changes in Y. If we naively implement a version of (1.13), we would not get this behaviour, but we do by implementing (1.16).

In the same vein, consider the CAG in figure 1.6, where nodes X and Y influence a third node, Z, but not each other. In this case, the influences of X and Y upon Z can be written out as follows:

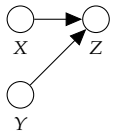

Figure 1.6: A 3-node CAG with two nodes influencing a third, but not each other.

$$\begin{aligned}\frac{dZ}{dt} &= \frac{\partial Z}{\partial t} + \frac{\partial Z}{\partial X}\frac{dX}{dt} + \frac{\partial Z}{\partial Y}\frac{dY}{dt} \\ &= \dot{Z} + \beta_{XZ}\frac{\partial X}{\partial t} + \beta_{YZ}\frac{\partial Y}{\partial t} \\ &= \dot{Z} + \beta_{XZ}\dot{X} + \beta_{YZ}\dot{Y}\end{aligned} \tag{1.17}$$

Next, consider the Y-shaped CAG in figure 1.7. Similar to the previous two cases, we can write out the total derivative of Z w.r.t time as:

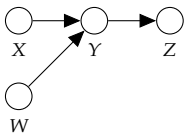

Figure 1.7: A 4-node Y-shaped CAG.

$$\frac{dZ}{dt} = \dot{Z} + \beta_{YZ}(\dot{Y} + \beta_{XY}\dot{X} + \beta_{WY}\dot{W}) \tag{1.18}$$

Finally, consider a causal analysis graph that contains a cycle.

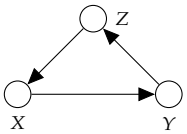

Figure 1.8: A 3-node CAG with two nodes influencing a third, but not each other.

$$\begin{aligned}\frac{dZ}{dt} &= \dot{Z} + \beta_{YZ}\left(\dot{Y} + \beta_{XY}\dot{X}\right) \\ \frac{dX}{dt} &= \dot{X} + \beta_{ZX}\left(\dot{Z} + \beta_{YZ}\dot{Y}\right) \\ \frac{dY}{dt} &= \dot{Y} + \beta_{XY}\left(\dot{X} + \beta_{ZX}\dot{Z}\right)\end{aligned} \tag{1.19}$$

In this case, each of the nodes would influence the others. The reason we do not infinitely recurse in a loop is because of the implicit temporal separation of the influence. That is, the value of Z at time step n will be affected by the values of the partial derivatives of the variables with respect to each other and w.r.t time, evaluated at the previous time step $(n-1)$.

1.5 IMPLICATIONS OF THE SIMPLE LINEAR MODEL

Note that the evidence sentence e_1 only talks about the case $\dot{X} > 0$. However, in using β_{XY} as we did above (independent of the sign of $\dot{X}$), we have implicitly implemented the following assumption:

If there exists only one evidence sentence that relates two quantities X and Y, then the following assumption holds:

Assumption 1 (Implication of the dual): the evidence sentence implies its 'dual' constructed by reversing the 'polarity' of the changes described in the sentence, i.e, if we change 'increase' to 'decrease' and vice-versa.

Example:

A small increase in X causes a large decrease in Y.

⇕

A small decrease in X causes a large increase in Y.

This also implies 'isotropic' sensitivity, i.e. if Y is sensitive to changes in X, then the adjective describing $|\Delta Y|$ depends only on the adjective describing $|\Delta X|$, as can be seen in the example above (the adjectives *small* and *large* remain unchanged upon reversing the 'polarity').

The magnitude of β_{XY} represents the sensitivity of Y to fluctuations in X and depends on some notion of the 'ratio' of the two adjectives 'small' and 'large'. As mentioned in § 1.2, there is some level of uncertainty involved in interpreting these adjectives - thus, there is some uncertainty in β_{XY}, which leads to uncertainty in $\mathbf{A}$, which in turn leads to uncertainty in $\boldsymbol{\theta}_T$ (nicely motivating our treatment of $\boldsymbol{\theta}_T$ as a random variable). Since the shape of the distribution of β_{XY} is not known to us *a priori*, we propose to estimate it empirically, using evidence sentences and data gathered about human intuitions about adjectives such as *small* and *large*, known collectively as *gradable adjectives*.

Gradable adjectives are often used to qualitatively describe magnitudes of changes in the literature on food insecurity. However, a small change in one domain can mean something very different from a small change in another. For example, a 15% drop in the average global temperature within a year would plunge the world into a little ice age [3], while a 15% change in precipitation in Hawaii over the same time period is hardly remarkable [4].

Thus, describing a change in some quantity with a gradable adjective requires some knowledge of the typical distribution of that quantity. In § 1.6, we provide a brief review of the CLULAB's approach to quantifying the effect of gradable adjectives through crowdsourcing.

1.6 GROUNDING GRADABLE ADJECTIVES

To quantify the size of the changes implied by different gradable adjectives, the CLULAB crowdsourced responses from a number of Amazon Mechanical Turk workers (henceforth referred to as Turkers) [8].

The turkers were provided with the name of a quantity (a nonword), its value for a particular group, (corresponding to the mean), and the typical range of values (corresponding to two standard deviations, assuming the data is normally distributed, to capture the intuitive sense of the word 'most' [3]) of some quantity and an adjective that described the magnitude of some change to that quantity. They were then tasked to give their best estimate of what that adjective implied for the magnitude of change in the quantity. For example, a task might look like this (taken verbatim from [8]):

> Most groups contain between 1470 to 2770 *mards* A particular group has 2120 mards. There is a *prominent* increase in this group. How many mards are there? (Please enter a number response).

From the response r given by the turker, the normalized quantity $\bar{r}$ was calculated:

$$\bar{r} = \frac{|r - \mu|}{\sigma}, \tag{1.20}$$

where μ is the mean and σ is the standard deviation (in the above example, 2120 and 325 respectively). This was then used as the dependent variable of the model.

Collecting the responses of multiple turkers for multiple adjectives, we obtain the distributions shown in table 1.1.

[3] This assumption can be adjusted - for example, if we take 'most' to mean three standard deviations instead of two, the points in figure 1.9 would be more spread out.

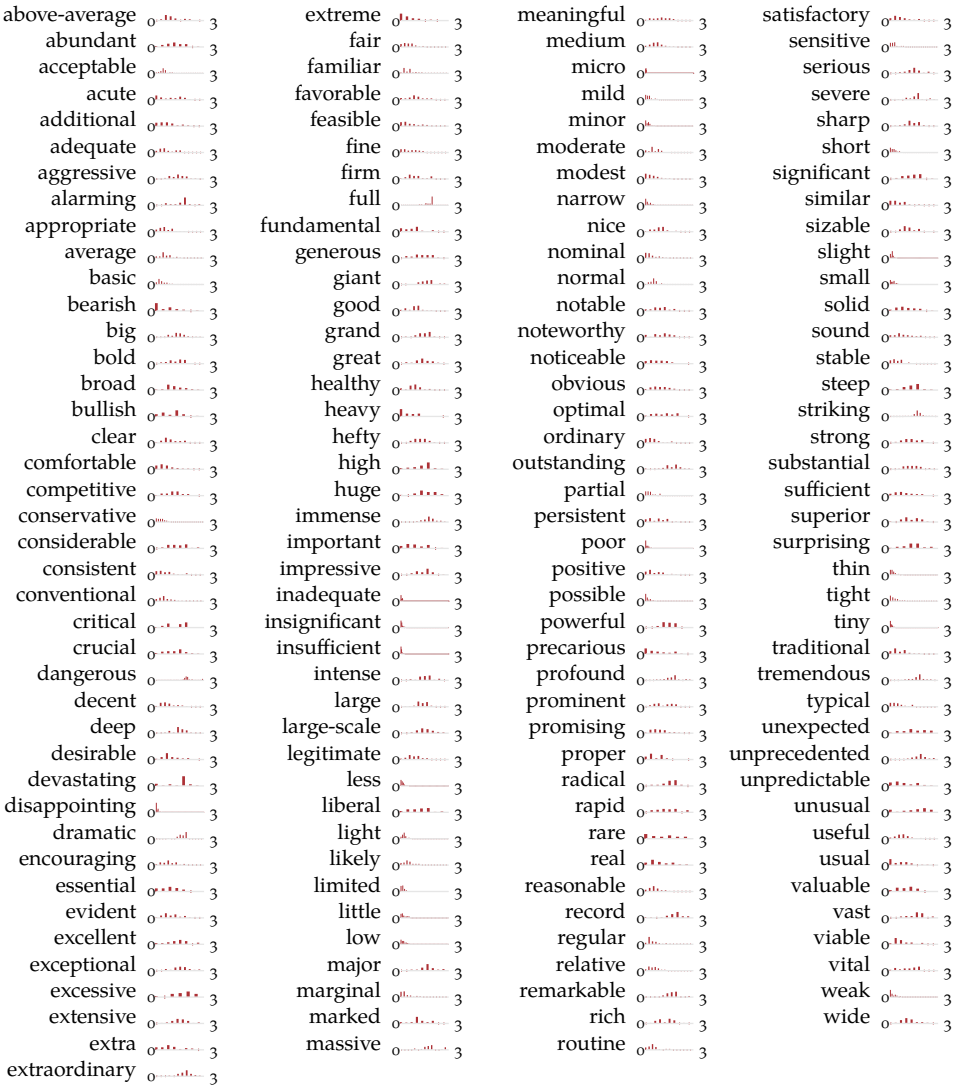

Table 1.1: Distributions of responses by adjective (as multiples of the standard deviation σ). For clarity, we set a common range of 0-3.

We can also visualize the joint distribution of $\bar{r}$ for two adjectives simultaneously. For example, figure 1.9 shows the cartesian product of the distributions of $\bar{r}$ for the adjectives *small* and *large* (assuming a small increase and large decrease, and the assumptions described in § 1.3), and a kernel density estimate. In visualizing this joint distribution, we implicitly make the following assumption:

Assumption 2 (Symmetry under polarity reversal): The magnitude of a change described by a gradable adjective is independent of whether the change is an increase or a decrease.

This assumption is supported by a pilot study conducted by the CLULAB, where it was found that there was no significant difference in the magnitude of changes for a gradable adjective when an decrease was specified as opposed to an increase. This observation also influenced the design of the task in [8] - turkers were only asked to estimate increases and not decreases associated with gradable adjectives. The effect of this assumption, along with Assumption 1, is that the data in the second quadrant is 'reflected' into the fourth.

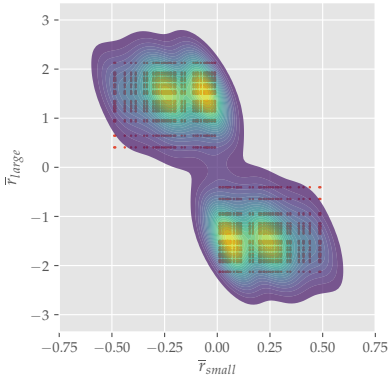

Figure 1.9: Figure showing the cartesian product of the distributions of $\bar{r}$ for the adjectives *small* and *large*, the kernel density estimate, and the best fit line to the data.

1.7 CONSTRUCTING EMPIRICAL PROBABILITY DENSITIES

If all the respondents agreed on the value of $\bar{r}_{\text{small}}$ and $\bar{r}_{\text{large}}$, we would be certain of the value of β_{XY} in (1.12) - it would simply be given by:

$$\beta_{XY} = \frac{\Delta Y}{\Delta X} = \frac{\sigma_Y \bar{r}_{\text{large}}}{\sigma_X \bar{r}_{\text{small}}} \tag{1.21}$$

However, as can be seen in figure 1.9, we have a distribution of values of β_{XY}, with $\sigma_Y/\sigma_X = 1$, each corresponding to the slope of a line through the origin. The distribution is shown in figure 1.10, which is constructed using the following procedure:

1. Take the cartesian product of the crowdsourced values of $\bar{r}_{small}$ and $\bar{r}_{large}$. This gives us a collection of points in the second quadrant of the $\bar{r}_{small} - \bar{r}_{large}$ plane.
2. Since there is only one evidence sentence in this example, reflect the points into the opposite quadrant.
3. Calculate the kernel density estimate (KDE) using Gaussian kernels, and bandwidth chosen using Scott's Rule [7]. This is done using the `scipy.stats.gaussian_kde` class from the `scipy` Python package.
4. Sample a dataset from the KDE.

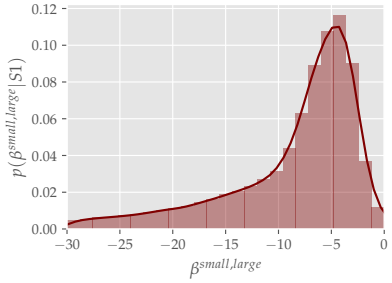

Figure 1.10: $\tilde{\beta}_{XY}$ distribution (values are restricted to $\tilde{\beta}_{XY} \in [-30, 0]$ for clarity).

5. For each point in the sampled dataset, treat it as if it lies on a line through the origin, then calculate the slope of this line:

$$\tilde{\beta}_{XY} = \bar{r}_{large}/\bar{r}_{small} \tag{1.22}$$

   (The tilde over the β indicates that this slope must be rescaled by σ_Y/σ_X when it is actually used in the transition matrix $\mathbf{A}$. In the examples in the rest of the document, we take $\sigma_Y/\sigma_X = 1$.)
6. Plot a histogram of these values of $\tilde{\beta}_{XY}$.

This distribution is clustered around the slopes of the lines passing through the yellow regions. A point estimate of β_{XY} can be obtained by taking the best-fit line through the data. While appealing in its simplicity, this approach will fail to capture the salient features of the distribution. For example, a best-fit line through the origin will not be able to pass through the regions with high probability densities (yellow). For this reason, we treat β_{XY} as a random variable.

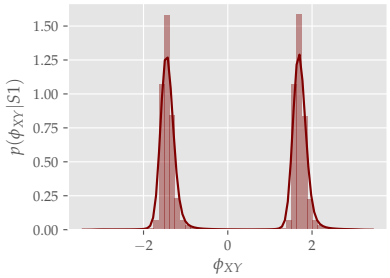

Figure 1.11: The distribution of θ_{XY}.

A compact reparameterization

In practice, it will be more convenient to work with θ_{XY}, defined by

$$\theta_{XY} = \text{arctan2}(\sigma_Y \bar{r}_{\text{large}}, \sigma_X \bar{r}_{\text{small}}) \tag{1.23}$$

where $\text{arctan2}(x_1, x_2)$ is the elementwise tangent[4] of x_1/x_2, instead of β_{XY}. The advantage of this reparameterization is that it allows the conditional probability density to have compact support. Substituting (1.23) in (1.13), we obtain a rule to update Y_t once we sample $\dot{X}_t$ and θ_{XY}:

$$Y_{n+1} = Y_n + \tan\theta_{XY}\dot{X}_n\Delta t. \tag{1.24}$$

The distribution of θ_{XY} resulting from e_1 is shown in figure 1.11.

[4] https://docs.scipy.org/doc/numpy/reference/generated/numpy.arctan2.html

Combining multiple sentences

In the previous section, we saw what the distribution of θ_{XY} looks like given a single evidence sentence. In this section, we will provide a method to construct the probability distribution function of θ_{XY} from multiple evidence sentences relating a pair of quantities.

1. For each evidence sentence implying that X causally influences Y:
   (a) Take the cartesian product of the responses corresponding to the pair of gradable adjectives in the sentence, and place the collection of points in the appropriate quadrant based on the polarity of the changes described in the sentence.
   (b) Calculate the kernel density estimate[5] for the data and nor-

[5] Gaussian kernel, with bandwidth chosen using Scott's Rule.

malize it so that it is only supported within that quadrant.

2. If the kernel density is restricted to either side of the y-axis, perform a 'reflection' such that the first and third quadrants are symmetric about the line $y = x$ and the second and fourth quadrant are symmetric about $y = -x$.
3. Normalize the sum of the kernel densities so that it forms a valid probability distribution.

We would then 'plug-in' this probability distribution into the Bayes network in figure 1.3.

1.8 PRIOR

At this stage, we are working on formulating a sensible prior as well, which will help when we have a paucity of evidence sentences, gradable adjectives, crowdsourced responses for the gradable adjectives, or any combination of these. Consider the following collection of evidence sentences, which forms a *representative agricultural pathway* (RAP).

> The government promotes improved cultivar to boost agricultural production for ensuring food security. However, the policy to seriously cut down the use of inorganic fertilizer and phase out the fertilizer subsidy results in deteriorating biophysical conditions, low use of inorganic fertilizer, less water, significantly reduced farm sizes which lead to low benefit from the improved cultivar.

The first sentence implies a positive correlation between the quality of the cultivar and agricultural production. However, it does not describe the extent of the improvement in the cultivar, or the amount that the agricultural production is boosted. To deal with this case, we introduce a *prior* over the distribution of $\theta_{\text{cultivar, agricultural production}}$ (abbreviated as θ below). We can construct the prior from empirical data, using the following procedure.

1. For each adjective:
   a) Calculate the kernel density estimator for the collection of responses $\bar{r}$ corresponding to that adjective.
   b) Sample n points from the KDE. The reason for doing this is that different adjectives might have different numbers of valid crowdsourced responses, and we wish to account for that [6].
2. Collect all the sampled points into a set S. Also construct a set $-S$ containing the negative of the points in S.
3. Construct the set of points $(S \times S) \cup (-S \times -S)$. The distribution of this set of points is shown in figure 1.12, for $n = 30$, using a 2-D histogram, with hexagonal bins. The marginal distributions are also shown.

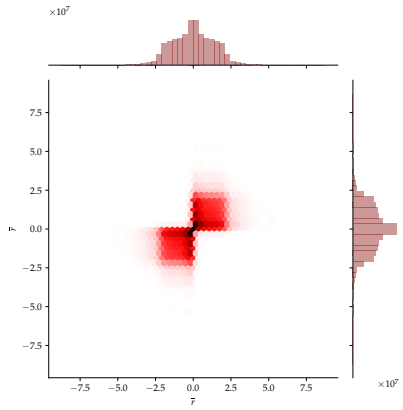

Figure 1.12: Distribution of $(S \times S) \cup (-S \times -S)$. This depicts the distribution of all combinations of responses $\bar{r}$, across all gradable adjectives studied in [8].

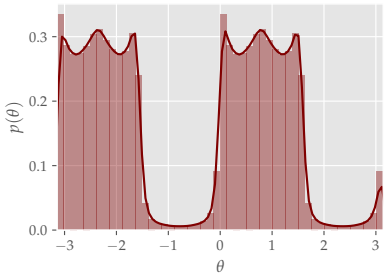

Figure 1.13: Distribution of θ generated using empirical data, for a positive correlation.

[6] Implicit in this step is the assumption that the collection of gradable adjectives studied in [8] is not inherently biased towards adjectives with particular distributions of $\bar{r}$.

4. Take the inverse tangent (restricted to $(-\pi, \pi)$) of each point in the set constructed in the previous step. The distribution of the resulting angles θ is shown in figure 1.13, for $n = 80$.

The shape of the prior over θ in figure 1.13 is also supported by our intuition that in the absence of any knowledge about the magnitudes of the changes in the factors, we expect that the sensitivity of the influenced factor to changes in the influencing factor will not be arbitrarily large. For a sentence that implies a positive correlation, we would simply replace $(S \times S) \cup (-S \times -S)$ with $(S \times -S) \cup (-S \times S)$ in steps 3 and 4 above.

1.9 EMISSION MODEL

We now return to the emission model, briefly mentioned in § 1.2. Abstract concepts like food security and conflict are not directly measurable quantities like mass, length, time, etc. However, there are certain measurable quantities that can often act as proxies for the abstract concepts. For example, intuitively, we would expect to see the number of people displaced by violence, as well as the number of battle-related fatalities increase as conflict increases. We term these measurable, proxy quantities as *indicators*, and use them as a starting point for briding the semantic gap between qualitative causal analysis graphs and quantitative models built on those analysis graphs.

Abstract concepts and indicators fit in nicely in a state-space modeling framework, since they can readily be associated with latent and observed variables. Our emission model, then, is as follows. Let n_{ij} be the value of the j^{th} indicator for the i^{th} component of a latent state vector $\mathbf{s}$, denoted s_i. Then

$$n_{ij} \sim \mathcal{N}(s_i \mu_{ij}, \sigma_{ij}) \tag{1.25}$$

where μ_{ij}, σ_{ij} are the mean and standard deviation of the indicator variable. In principle, this amounts to a systematic rescaling of a given value for an indicator variable according to the change in the corresponding latent variable. Two caveats must be kept in mind for this model.

The first is that this model implicitly assumes that the 'polarity' of the indicator is the same as the polarity of the abstract concept that it acts as a proxy for. That is, the natural language interpretation of an 'increase' in the abstract concept corresponds to an increase in the indicator variable, and vice versa, as in the example with conflict and battle-related fatalities described earlier. The case where the indicator and the abstract concept it represents have opposite polarities (for example, food insecurity and average number of kCal consumed per day) would require either an extension to this model or linguistic normalization of the causal analysis graph (i.e. replacing food insecurity with food security).

The second is that this simple model does not address the case of 'shared' indicators, that is, when an indicator can act as a proxy for multiple abstract concepts. Intuitively, we would tend to disfavor this kind of model, since if an indicator represents two concepts simultaneously, we think that the two concepts are linked by polarity reversal (like food security and food insecurity), or by a causal relation. For example, the average number of kCal consumed per day could be an indicator for both the abstract concepts drought and food security. However, drought influences food security, thus if both concepts are in the causal analysis graph, the indicator variable should ideally represent only food security and not drought. These kinds of 'global hints' can be implemented in an ensemble modeling framework, in which we marginalize over a distribution of causal analysis graph link structures instead of a single one.

Indicators with polarities opposite to the concept

Often we will encounter indicators that have a polarity opposite to the concept that they are associated with. For example, drought and rainfall.

1.10 RELAXING OUR ASSUMPTIONS

As we gain more knowledge about the domain, we could replace the simple transition model we have described here with more complex versions, informed by subject matter experts.

2

INFERENCE

$$\frac{}{x \leftarrow y}$$

3 SOFTWARE CONTRIBUTION

In this section, we document the design of our open-source framework for assembling causal dynamic bayes networks from reading evidence, parameterizing the models with real-world data, and 'executing' the models, i.e. running sampling and inference to determine probability distributions for the random variables in the model.

The framework is named Delphi, since one of the primary intended capabilities of the software is to be able to make quantitative, probabilistic predictions. Delphi can be found at `https://github.com/ml4ai/delphi`.

Delphi implements algorithms to compute the probability distributions over θ_{XY} between pairs of quantities X and Y given a collection of evidence sentences relating them. These distributions can then be plugged into dynamical Bayesian networks such as the one depicted in figure 1.3 to perform inference and make predictions. In addition, Delphi provides functionality to analyze the structure of FORTRAN programs, convert them into causal analysis graphs, and perform sensitivity analysis on them.

Delphi is implemented in Python, to provide a user-friendly and easily extensible interface. In the future, we plan to extend Delphi using C/C++ in the backend to be able to perform inference at scale with large causal analysis graphs. The main data structure is the `AnalysisGraph` class, which are acted upon by functions for visualization, inspection, manipulation, and 'execution', i.e. sampling and inference, that reside in eponymous modules. The API reference can be found at `http://delphi.readthedocs.io`.

BIBLIOGRAPHY

[1] Sue Ann Anderson, *Core indicators of nutritional state for difficult-to-sample populations*, The Journal of nutrition (USA) (1990) (cited on page 2).

[2] C Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn*, Springer, New York (2007) (cited on page 4).

[3] Michael Carlowicz, *Global Temperatures*, URL: https://earthobservatory.nasa.gov/Features/WorldOfChange/decadaltemp.php (cited on page 9).

[4] *Climate at a Glance*, URL: https://www.ncdc.noaa.gov/cag/ (cited on page 9).

[5] Jinyan Guan, Kyle Simek, Ernesto Brau, Clayton Morrison, Emily Butler, and Kobus Barnard, "Moderated and Drifting Linear Dynamical Systems", *Proceedings of the 32nd International Conference on Machine Learning*, ed. by Francis Bach and David Blei, vol. 37, Proceedings of Machine Learning Research, Lille, France: PMLR, July 2015, pp. 2473–2482 (cited on page 5).

[6] Steve Hatfield-Dodds, Heinz Schandl, Philip D Adams, Timothy M Baynes, Thomas S Brinsmead, Brett A Bryan, Francis HS Chiew, Paul W Graham, Mike Grundy, Tom Harwood, et al., *Australia is 'free to choose' economic growth and falling environmental pressures*, Nature **527**. (2015), pp. 49–53 (cited on page 2).

[7] David W. Scott, "Multivariate Density Estimation and Visualization", *Handbook of Computational Statistics: Concepts and Methods*, ed. by James E. Gentle, Wolfgang Karl Härdle, and Yuichi Mori, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 549–569, ISBN: 978-3-642-21551-3 (cited on page 12).

[8] Rebecca Sharp, Ajay Nagesh, Dane Bell, and Mihai Surdeanu, *Grounding Gradable Adjectives through Crowdsourcing*, (Manuscript submitted for publication) (2017) (cited on pages 10, 12, 14).

[9] *World Modelers Broad Agency Announcement*, HR001117S0017, DARPA, 2017 (cited on page 2).