BRAINS, VOCAL BEHAVIOR AND MISSING DATA

RICHARD MCELREATH

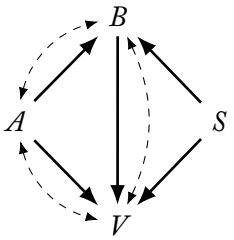
1. Purpose

Brain structure changes with age and varies by sex and population. Vocal behavior also changes with age and varies by sex and population. Under the hypothesis that brain structure influences vocal behavior, we need to estimate any causal influence of brain structure while controlling for unobserved confounds associated with age and sex and population. For example, if brain structure and vocal behavior both are influenced by other age-related factors, then structure will be associated with vocal behavior, but the relationship may not be causal.

The problem is made more complex by the existence of extensive missing data. In the expected sample, there is extensive vocal data but much less brain structure data. A robust solution that allows the use of all data will require a strategy for the imputation of missing values.

2. Strategy

2.1. **Concept.** We begin by specifying the heuristic causal model of the system, in order to clarify the general conditions under which a causal estimate is possible. This is also sufficient to remind us why experiments were invented.



In this diagram, each letter is an observable variable: Age (A), Sex (S), Brain structure (B), and Vocal behavior (V). The arrows represent causal relationships. The dashed paths represent unobserved confounds. The directed arrow $B \to V$ is what we want to estimate.

Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany

E-mail address: richard_mcelreath@eva.mpg.de.

Date: January 5, 2022.

2.2. **Inference.** Using knowledge of d-separation and the backdoor criterion of do-calculus, we can derive the general conditions under which inference of the causal effect of B on V is possible.

First notice that there are backdoor paths through A and S. Stratification by A and S is sufficient to block these paths. So we must stratify by age and sex in order to estimate $B \to V$.

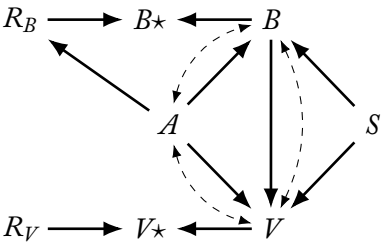
The unobserved confounds connecting A to B and V represent unknown cohort effects that mimic age but are unrelated to biological aging. These paths are problematic. If we stratify by age, then age becomes a collider on $B \longleftrightarrow A \longleftrightarrow V$. Therefore it is not possible to control for unknown cohort effects.

The unobserved confound connecting B and V obviously causes problems. This could arise from factors like common family exposure that is unobserved but generates association between brain structure and vocal behavior across families. It is impossible to exclude the possibility of unobserved confounds, but it is possible to calculate how strong the confounds must be to remove any apparently causal effect of B on V.

In summary, while it is not possible in theory to eliminate all plausible confounding, it is possible to interpret estimates correctly as mixes of causal and confound effects. For example, if we are willing to assume that cohort effects and unobserved confounds between B and V are absent, then stratification by A and S is sufficient to estimate the causal influence of B on V. Any estimated coefficient in such an analysis will likely be an upper bound, due to unobserved cohort or other confounds.

2.3. **Missing data.** When some data are missing, we must augment the causal model to simultaneously model to missingness mechanism. The cause of missing values tells us whether they are ignorable or rather how to impute them.

Consider the example below, in which the previous causal diagram is augmented with missingness nodes R that point into observed variables with stars $\star$ that indicate variables containing missing values. The corresponding variable without the $\star$ is unobserved (no missing values).



Consider V first. In this example, we observe only $V\star$, which is influenced by the true values V and the missingness mechanism R_V. This is the most benign form of missing data, *missing completely at random*. The association between B and V cannot be confounded by the missingness. It is only harder to measure. The causal model can be used to impute missing values, so that incomplete cases do not need to be dropped. But this is not strictly necessary for valid inference.

Now consider instead B. In this case, the missingness mechanism R_B is itself influenced by another variable, age (A). This is very plausible, since individuals of some ages are more likely to die and therefore contribute brain samples. In this case, there is a backdoor path through A that confounds inference of $B \to V$. We must condition of A to close that path, blocking the influence of R_B.

As explained before, if there are unobserved cohort effects (the dashed paths between B, A and V) then conditioning on A will introduce a new confound that cannot be adjusted for. But that situation is not new. It is not caused by the missing B values.

So what can we do? For each variable with missing values, we must state what influences missingness. If missingness is caused by an observed variable (like A), usually we can adjust for it and proceed as before. We can impute missing values as well, to retain as much efficiency in estimation as possible.

But if any pattern of missingness is caused by the variable itself, then we must model the missingness mechanism or otherwise give up. For example, in censored time-to-event data the values that are missing are missing exactly because of their value: the event happened after the sampling period ended. These missing values cannot be ignored. And no other observed variable can be conditioned on to resolve the problem. So instead we model the censoring process and use it to impute the missing values. That is indeed the standard justification for time-to-event analysis, where censored observations must be included in the analysis but modeled differently than observed values.

2.4. **Reciprocal causation.** An alternative causal model posits a feedback between brain structure and vocal behavior over time. In this case, any association between brain structure and vocal behavior cannot be interpreted as the influence of one on the other. Without a time series of brain and vocal measurements, it is not possible to do more. However, it is possible to specify a generative model, using for example ODEs, to express the hypothesis.

2.5. **Vocal variables.** The variables that stand for vocal behavior must be inferred from individual recordings. Individual chimpanzees are represented by different numbers of recordings. In many cases there are hundreds of recordings. But other individuals have only 1 recording. So ideally the vocal behavior is inferred from a separate set of latent factor equations. The resulting factor scores can be used as the outcome variable in the $B \to V$ model.

The simplest, and canonical, approach is to simply construct the factor scores as varying intercepts clustered on each individual chimpanzee. Each recording can be scored for whatever feature F is of interest. Then each feature is modeled as:

$$\begin{aligned} F_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha_{\text{ID}[i]} \\ \alpha_j &\sim \text{Normal}(\bar{\alpha}, \tau) \end{aligned}$$

The equation for α_i can be extended to include predictors for age, sex, or anything else. In that case:

$$\begin{aligned} F_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha_{\text{ID}[i]} \\ \alpha_j &= \bar{\alpha} + \gamma_j + ... \\ \gamma_j &\sim \text{Normal}(0, \tau) \end{aligned}$$

where the ... is a series of terms that stratify scores by demographic and other features.

When there are multiple vocal features, they can be simultaneously modeled using a multivariate normal, if strong correlations among them are expected. In that case, each individual is assigned a vector $\mathbf{a}_j = [\alpha_{j1}, \alpha_{j2}, ..., \alpha_{jn}]$ for n features and a covariance matrix models the structure among them.

3. Implementation

We implement a complete generative model, corresponding to the heuristic causal model, so that we can produce synthetic data that allows us to (1) clarify the hypothetical relationships among variables (through forward simulation) and (2) validate the statistical implementation (through inverse inference).

3.1. **Functional relationships.**

For each variable, it is necessary to specify how the other variables influence it. In the simplest models, all relationships are additive. But this is not necessary and rarely biologically realistic.

To motivate the analysis, we adopt simple non-linear relationships for brain structure and vocal behavior as functions of age, sex and brain (for vocal behavior).

For brain structure, suppose a logistic relationship with age, so that:

$$\begin{aligned} B_i &= \mu_i / (1 + \exp(-(\beta_{S[i]}(\mathcal{A}_i - \alpha) + \epsilon_i))) \\ \epsilon_i &\sim \text{Normal}(0, \sigma) \end{aligned}$$

The β parameters determine the rate of increase with age and α determines when the increase is maximized. The μ parameter is an individual-specific adult maximum. Obviously this cannot be estimated for young individuals. But a distribution could possibly be estimate for a population.

For vocal behavior, suppose a Poisson sample with a rate determined by brain structure, age and sex:

$$\begin{aligned} V_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &= B_i^{\gamma_{S[i]}} \mathcal{A}_i^{\delta} \end{aligned}$$

Here γ is a parameter (unique to each sex) that determines the influence of brain structure on the rate of vocal behavior. The parameter δ governs the rate of diminishing returns with age. When $\delta < 1$, the influence of age eventually goes flat. Age $\mathcal{A}_i$ in this function is a kind of exposure variable, and δ and γ are both effectively elasticities. AN economist will recognize the function for λ_i as a Cobb-Douglas

125 production function. It could also be expressed as a log-linear relationship, as is customary in GLMs:

$$\log \lambda_i = \gamma_{S[i]} \log B_i + \delta \log A_i$$

For real data, more thought needs to be given to how brain structure is measured and how it could theoretically influence behavior.

Finally, age and sex are given simple distributions:

$$A_i \sim \text{Uniform}(1, 20)$$
$$S_i \sim \text{Categorical}(\phi) \quad \text{for } S_i \in \{1, 2\}$$

130 The minimum age of 1 and maximum of 20 are arbitrary.

Code for simulating from this generative model is provided.

3.2. **Statistical inference.** The generative model above can be specified directly as a statistical model. Any probabilistic programming language is sufficient. We use Stan (mc-stan.org), because of its flexibility and availability in all scripting lan-
135 guages.

Bayesian imputation of missing values is straightforward for continuous variables. For discrete variables, like a Poisson count, the code will need to marginalize over unknowns. Such code can be slow, but given that the sample size is rather small, this is not an obstacle yet.

140 4. Validation

x