Dial $\hat{M}$ for Monotonic: A Kernel-Based Bayesian Approach to State-Trace Analysis

Gregory E. Cox and Michael L. Kalish
Syracuse University

Abstract

A monotonic state-trace implies that a single latent factor is sufficient to explain the joint variation between two outcome variables across a set of conditions. There are, however, few methods available for assessing how much evidence a sample of data provides about whether the variables are truly monotonically related or not. We present a model that allows researchers to estimate the statistic $\hat{M}$ which reflects the amount of evidence a dataset provides about whether two outcome variables are jointly monotonically related. This model is based on modeling the covariation between outcome measures in terms of a kernel function, which allows for computation of the latent derivatives of each outcome variable with respect to the other. $\hat{M}$ is the posterior odds that these derivatives are all of the same sign and are thus monotonic. Simulations show that $\hat{M}$ discriminates between monotonic and non-monotonic state traces and an example illustrates how the model can be applied to both continuous and binomial data from between-subjects, within-subjects, or mixed designs.

Keywords: State-trace analysis; Gaussian processes; Bayesian statistics.

Introduction

In many subfields of psychology and cognitive science, it is of critical theoretical importance to decide whether or not a set of results can be accounted for in terms of a single underlying factor or process. As a toy example, consider that foot size and vocabulary size, which are strongly correlated, can be jointly explained in terms of a single underlying factor, namely, age. More pertinent to psychologists, researchers want to know whether memory retrieval is a result of separate familiarity and recollection processes (Parks & Yonelinas, 2007; Brainerd & Reyna, 2010), whether categorization relies on separate rule-based and similarity-based systems (Ashby & Maddox, 2005), whether judgments arise from

---

This work was supported by grant 1461365 from the National Science Foundation.

Correspondence concerning this article should be addressed to Gregory E. Cox, Department of Psychology, 430 Huntington Hall, Syracuse University, Syracuse, NY 13244-2340. E-mail: gregcox7@gmail.com.

(a) Monotonic state-trace plots, indicating that a single factor is sufficient to explain how experimental manipulations affect the two outcome variables.

(b) Non-monotonic state-trace plots, indicating that a single factor is not sufficient to explain the effect of experimental manipulations on the two outcome variables.

Figure 1. Examples of state-trace plots illustrating both monotonic and non-monotonic state-traces.

a combination of rapid emotional reactions and slow deliberative reasoning (Evans, 2003), to cite just a few examples. The first—and most fundamental—step toward addressing these theoretical questions is to assess whether or not a single factor is sufficient to explain extant results. If so, then additional factors need not be assumed, leading to a more parsimonious account. If not, then researchers must posit additional factors to explain their results. Although a number of methods exist by which a researcher could take this step, many of them force researchers to make various strong assumptions about how underlying psychological factors are related to one another and to the data, such as orthogonality in the case of Principal Components Analysis or linearity in the case of factor analysis or analysis of variance. One method that does not require such assumptions is *state-trace analysis* (Bamber, 1979).

A state-trace is a plot in which data on a set of outcome variables are plotted against each other, and where the measurements were obtained in some number of different experimental conditions. Figure 1 shows some example state-trace plots each with two outcome

Figure 2. Measurement variability can make it difficult to determine whether a state-trace is actually monotonic or not, as shown by two samples drawn from an underlying monotonic state trace (a), one of which appears monotonic (b) while the other does not (c).

variables¹. The motivation for making a state-trace plot is to see whether the outcome variables are affected by experimental manipulations in a common manner. If so, then any time a manipulation increases the value of one outcome, it should either always increase or always decrease the value of the other outcome. This kind of situation, depicted in Figure 1a, yields a *monotonic* state-trace. A monotonic state-trace indicates that the effect of the experimental manipulations on the two outcome measures can be explained in terms of a single underlying factor that is affected by those manipulations but which can manifest differently in the two outcome variables. If the state-trace is non-monotonic, examples of which are shown in Figure 1b, the two outcome measures are not affected by the manipulations in the same way, and cannot be explained by a single factor.

Although the qualitative distinction between monotonic and non-monotonic state-traces is conceptually clear, in practice, data are noisy, different numbers of measurements may be made in each condition, and data may only explore a small subregion of the space defined by the two outcome variables. For these reasons, statistics are needed to assess the degree to which the data used to produce a state trace plot actually provide evidence for or against monotonicity. For example, the state-traces in Figures 2b and 2c were both generated from an underlying set of means that are monotonically related (Figure 2a), but due to measurement variability, Figure 2b appears monotonic while Figure 2c does not. Recent work has investigated both frequentist (Kalish, Dunn, Burdakov, & Sysoev, 2016) and Bayesian (Prince, Brown, & Heathcote, 2012) methods for computing the degree to which a set of data yield a (non-)monotonic state-trace, but these methods can be computationally burdensome, limited to particular kinds of outcome measures (like only proportions; Prince et al., 2012), and have difficulty incorporating information about different sources of uncertainty. The goal of the present work is to develop an analysis model that yields a statistic $\hat{M}$ that assesses the degree to which a set of data yield evidence for a monotonic

¹For convenience, we focus on the case of just two outcome variables throughout, although we discuss the case of more than two outcomes later as the same logic applies.

state trace, where the model is not computationally prohibitive, can incorporate multiple sources of uncertainty, and is not limited to particular kinds of outcome measures. After describing the model, we show how it can yield a Bayesian estimate of $\hat{M}$ which can then be used to make principled inferences from the state-trace in a variety of situations. We provide code for the model using the R interface (R Core Team, 2017) for the Stan language (Carpenter et al., 2017) which allows it to be run efficiently.

Statistical Model for Assessing Monotonicity

Our analysis model for computing evidence for monotonicity $\hat{M}$ of a state-trace begins by assuming that measurements have been made on two outcome variables, which we refer to as x and y , in the same set of N_C conditions. These measurements need not have been made in the same experiment or with the same participants. The goal of the model is to simultaneously estimate the means of x and y in each condition, which reflect the effects of the experimental manipulations in each condition, as well as the *derivatives* of the means, with respect to each other, in each condition. Note that this is not the same thing as how x and y might covary *within* a condition if they were measured using the same participants; instead, it is a latent property of how x and y are jointly affected by the experimental manipulations in each condition. A monotonic state-trace is one where the derivatives of the means of x with respect to the means of y are all the same sign, as are the derivatives of the means of y with respect to the means of x . The quantity $\hat{M}$ is an estimate of how much evidence the data provide about whether the state-trace was generated by an underlying monotonic relationship between the two outcome measures. Estimating all these quantities in a way that properly reflects uncertainty requires the use of a *kernel* function which specifies how the means of x and y can covary. We now describe the model in detail.

Data representation

The data are defined by four vectors: $\mathbf{x}$ and $\mathbf{y}$, the individual observations of each variable; and $\mathbf{c}_x$ and $\mathbf{c}_y$, vectors that indicate to which experimental condition each observation belongs. This representation allows for different numbers of observations on each variable in each condition, as would be expected when observations come from different groups of participants. If observations can come from the same participants, then additional vectors, $\mathbf{g}_x$ and $\mathbf{g}_y$, are needed which indicate which participant generated which observation.

Standardized observations

Prior to analysis, we standardize the observations on each variable:

$$\begin{aligned} \mathbf{z}_x &= \frac{\mathbf{x} - \text{Mean}[\mathbf{x}]}{\text{SD}[\mathbf{x}]} \\ \mathbf{z}_y &= \frac{\mathbf{y} - \text{Mean}[\mathbf{y}]}{\text{SD}[\mathbf{y}]} \end{aligned}$$

where m_x and s_x are the mean and standard deviation of $\mathbf{x}$, and similarly for $\mathbf{y}$. The effect of standardization is to make it easier to place priors on model parameters, since priors on means can be centered around zero while those on scale parameters can be centered around

one, as described below. When data cannot be standardized in this way, such as when the data consist of a set of binary outcomes, different procedures are needed, as described in the worked example later in the paper.

Data model

The likelihood of each centered datum is based on vectors of means ($\mathbf{m}_x$ and $\mathbf{m}_y$) and standard deviations ($\mathbf{s}_x$ and $\mathbf{s}_y$). For now, we define the likelihood in terms of a Gaussian distributions, however we show in the example application below how the Gaussian likelihood can be replaced with a binomial likelihood for proportion data (compare the equations below with Equations 20 and 21):

$$\forall i \in 1 \dots N_x z_{x,i} \sim \mathcal{N}(m_{x,i}, s_{x,i}) \quad (1)$$

$$\forall j \in 1 \dots N_y z_{y,j} \sim \mathcal{N}(m_{y,j}, s_{y,j}) \quad (2)$$

where N_x and N_y are the total number of observations of the x and y variables, respectively, and the vectors of means and standard deviations are defined in terms of which condition each observation comes from:

$$\forall i \in 1 \dots N_x m_{x,i} := \beta_0 + \beta_d + \mu_{x,c_{x,i}}$$

$$\forall i \in 1 \dots N_x s_{x,i} := \sigma_{x,c_{x,i}}$$

$$\forall j \in 1 \dots N_y m_{y,j} := \beta_0 - \beta_d + \mu_{y,c_{y,j}}$$

$$\forall j \in 1 \dots N_y s_{y,j} := \sigma_{y,c_{y,j}}.$$

If there are multiple observations from the same participant, then each participant is assumed to affect the mean of the observations made on them, but not the measurement error associated with those observations:

$$\forall i \in 1 \dots N_x m_{x,i} := \beta_0 + \beta_d + \mu_{x,c_{x,i}} + \gamma_{g_{x,i},i}$$

$$\forall j \in 1 \dots N_y m_{y,j} := \beta_0 - \beta_d + \mu_{y,c_{y,j}} + \gamma_{g_{y,i},N_c+j}$$

where each participant is associated with a vector γ of $2N_c$ parameters corresponding to the two outcome variables in each of the N_c conditions (or as many conditions for which there are multiple observations from the same participant), as described in the next section.

Within-condition variability

Each condition c is associated with scale parameters for the x and y observations in that condition, $\sigma_{x,c}$ and $\sigma_{y,c}$. Because the data are scaled prior to analysis, it is natural to place a prior on scale parameters that is concentrated around 1 and does not permit very high values; to that end, we use Gamma distributions with shape and rate parameters chosen to yield a mode at 1 and a standard deviation of 1:

$$\forall c \in 1 \dots N_c \sigma_{x,c} \sim \Gamma\left(\frac{3 + \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2}\right) \quad (3)$$

$$\forall c \in 1 \dots N_c \sigma_{y,c} \sim \Gamma\left(\frac{3 + \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2}\right). \quad (4)$$

These parameters indicate the degree of measurement error in each condition, and are not *a priori* correlated either between conditions or between x and y within a condition.

If some observations come from the same participants, we assume that each participant $l = 1 \dots N_p$ is associated with a vector of $2N_c$ deviations from the group mean, as there is an average deviation for each outcome variable (of which there are 2) in each condition (of which there are N_c):

$$\forall_{l \in 1 \dots N_p} \gamma_l \sim \mathcal{N}(\mathbf{0}, \tau \Omega \tau^T) \quad (5)$$

where Ω is a correlation matrix that allows for covariance between participant-level deviations and τ is a vector of scales of participant-level variability in each condition, which are assigned priors

$$\Omega \sim \text{LKJ}(2) \quad (6)$$

$$\forall_{c \in 1 \dots 2N_c} \tau_c \sim \Gamma\left(\frac{3 + \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2}\right) \quad (7)$$

where the “LKJ” prior over correlation matrices is that given by Lewandowski, Kurowicka, and Joe (2009). Note that these “default” priors on scale parameters assign zero density to scales of zero, however it would be possible to adjust them if there was reason to believe variability within a condition and/or participant could be exactly zero (we prefer to allow variability to be very small just not zero).

Between-condition variability

Exactly as in effect-coding for regression, we model the overall mean across outcome measures as being sampled from a standard normal distribution:

$$\beta_0 \sim \mathcal{N}(0, 1)$$

and the overall means (across conditions) in each outcome measure are a deflection from μ_0 , also with a standard normal prior

$$\beta_d \sim \mathcal{N}(0, 1).$$

While the data are standardized prior to analysis in most applications, this is not always possible and, because the standardization is with respect to *all* observations, it may be biased if there are more observations in one condition. β_0 and β_d serve to correct any such biases.

We then model the effect of condition on each outcome measure as being sampled from a multivariate normal distribution centered around a vector of zeros and with covariance matrices K_x and K_y defined by a kernel:

$$\begin{aligned} \mu_{\mathbf{x}} &\sim \mathcal{N}(\mathbf{0}, K_x) \\ \mu_{\mathbf{y}} &\sim \mathcal{N}(\mathbf{0}, K_y). \end{aligned}$$

The kernel specifying each element of the covariance matrices K_x and K_y is chosen to meet two criterion. First, the covariance of the means of conditions c and d should increase as

the distance between their means on the *other* variable decreases. Second, recall that the motivating spirit of state-trace analysis is that a state-trace merely needs to be monotonic to meet the criterion of uni-dimensionality. We thus aim to choose kernels that impose the minimum smoothness constraint needed to obtain a distribution over their derivatives. The kernels chosen are of the Matérn family of symmetric kernels and are differentiable exactly once:

$$\begin{aligned} k(\mu_{x,c}, \mu_{x,d}) &= (\rho_x \alpha_x)^2 \left(1 + \frac{\sqrt{3} |\mu_{y,c} - \mu_{y,d}|}{\lambda_x} \right) \exp \left(-\frac{\sqrt{3} |\mu_{y,c} - \mu_{y,d}|}{\lambda_x} \right) \\ &\quad + [(1 - \rho_x) \alpha_x]^2 \delta(c - d) \\ k(\mu_{y,c}, \mu_{y,d}) &= (\rho_y \alpha_y)^2 \left(1 + \frac{\sqrt{3} |\mu_{x,c} - \mu_{x,d}|}{\lambda_y} \right) \exp \left(-\frac{\sqrt{3} |\mu_{x,c} - \mu_{x,d}|}{\lambda_y} \right) \\ &\quad + [(1 - \rho_y) \alpha_y]^2 \delta(c - d) \end{aligned}$$

where α_x and α_y represent the total amplitude of the variability in the means of x and y , λ_x and λ_y are length-scales, and ρ_x and ρ_y represent the proportions of variability that are governed by the kernel as opposed to unstructured noise on the diagonal (i.e., when $c - d = 0$ so the Dirac delta function $\delta(c - d) = 1$). Finally, to complete the model, we impose Gamma priors on the amplitudes and length-scales and boundary-avoiding weak Beta priors on the proportions:

$$\alpha_x \sim \Gamma \left(\frac{3 + \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2} \right) \quad (8)$$

$$\alpha_y \sim \Gamma \left(\frac{3 + \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2} \right) \quad (9)$$

$$\lambda_x \sim \Gamma \left(\frac{3 + \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2} \right) \quad (10)$$

$$\lambda_y \sim \Gamma \left(\frac{3 + \sqrt{5}}{2}, \frac{1 + \sqrt{5}}{2} \right) \quad (11)$$

$$\rho_x \sim \text{Beta}(2, 2) \quad (12)$$

$$\rho_y \sim \text{Beta}(2, 2). \quad (13)$$

Computing derivatives

The kernel that defines the Gaussian process for the means of each condition is once-differentiable, meaning that it imposes a Gaussian process on the first derivatives at each condition as well. The covariance kernels of these derivatives are

$$\begin{aligned} k \left(\frac{d\mu_{x,c}}{d\mu_{y,c}}, \frac{d\mu_{x,d}}{d\mu_{y,d}} \right) &= \frac{3(\alpha_x \rho_x)^2}{\lambda_x^2} \exp \left(-\frac{\sqrt{3} |\mu_{y,c} - \mu_{y,d}|}{\lambda_x} \right) \left(1 - \frac{\sqrt{3} |\mu_{y,c} - \mu_{y,d}|}{\lambda_x} \right) \\ k \left(\frac{d\mu_{y,c}}{d\mu_{x,c}}, \frac{d\mu_{y,d}}{d\mu_{x,d}} \right) &= \frac{3(\alpha_y \rho_y)^2}{\lambda_y^2} \exp \left(-\frac{\sqrt{3} |\mu_{x,c} - \mu_{x,d}|}{\lambda_y} \right) \left(1 - \frac{\sqrt{3} |\mu_{x,c} - \mu_{x,d}|}{\lambda_y} \right) \end{aligned}$$

and yield covariance matrices for the derivatives of the means of each condition, K'_x and K'_y . Matrices Q_x and Q_y expressing the covariance between the derivatives and the means themselves are given by the kernels

$$\begin{aligned} k\left(\frac{d\mu_{x,c}}{d\mu_{y,c}}, \mu_{x,d}\right) &= -\frac{3(\alpha_x\rho_x)^2}{\lambda_x^2}(\mu_{y,c} - \mu_{y,d}) \exp\left(-\frac{\sqrt{3}|\mu_{y,c} - \mu_{y,d}|}{\lambda_x}\right) \\ k\left(\frac{d\mu_{y,c}}{d\mu_{x,c}}, \mu_{y,d}\right) &= -\frac{3(\alpha_y\rho_y)^2}{\lambda_y^2}(\mu_{x,c} - \mu_{x,d}) \exp\left(-\frac{\sqrt{3}|\mu_{x,c} - \mu_{x,d}|}{\lambda_y}\right). \end{aligned}$$

Thus, the vector of derivatives has an implicit Gaussian prior

$$\begin{aligned} \mu'_{\mathbf{x}} &\sim \mathcal{N}(\mathbf{0}, K'_x) \\ \mu'_{\mathbf{y}} &\sim \mathcal{N}(\mathbf{0}, K'_y) \end{aligned}$$

which, given the estimated condition means yields the conditional posterior distribution

$$\widehat{\mu'_{\mathbf{x}}} \sim \mathcal{N}(Q_x K_x^{-1} \mu_{\mathbf{x}}, K'_x - Q_x K_x^{-1} Q_x^T) \quad (14)$$

$$\widehat{\mu'_{\mathbf{y}}} \sim \mathcal{N}(Q_y K_y^{-1} \mu_{\mathbf{y}}, K'_y - Q_y K_y^{-1} Q_y^T). \quad (15)$$

Parameter estimation

The locations of the condition means, the condition variability, and the kernel parameters that define the hyper-distribution on the condition means are all random variables that must be estimated using the data. Although this could be done via maximum likelihood methods to obtain a point estimate, we advocate using Bayesian methods to obtain an entire distribution over the joint parameter space. The present application uses “Stan” for this purpose (Carpenter et al., 2017), which estimates the joint posterior distribution by drawing a large number of representative *samples* of parameters from that distribution. If we draw, for example, 10,000 samples from the posterior, we then have 10,000 complete sets of model parameters, each of which is a representative sample of the posterior. The variability inherent in the data propagates into the variability between these samples and, therefore, reflects our residual uncertainty about the degree to which x and y are jointly monotonic.

Assessing monotonicity

Monotonicity occurs when, simultaneously, all the derivatives of x with respect to y are of the same sign *and* all the derivatives of y with respect to x are of the same sign. For each posterior sample of model parameters, recall that there is a conditional multivariate Gaussian posterior over the derivatives, as given by Equations 14 and 15. To estimate the probability that the derivatives have the same signs (given that there are no general closed-form methods for computing the multivariate Gaussian cumulative distribution), we draw samples of vectors of x derivatives (Equation 14) and vectors of y derivatives (Equation 15) for each sample from the posterior distribution of the other model parameters and count how many are such that they are all of the same sign. Let $M_{X,s} = 1$ if sample s of the vector of x derivatives are of the same sign (and $M_{X,s} = 0$ otherwise) and $M_{Y,s} = 1$ if

sample s of the vector of y derivatives are of the same sign (and $M_{Y,s} = 0$ otherwise). Then $M_{X,s} \times M_{Y,s} = 1$ only if samples s of x and y are jointly monotonic. Finally, we can define

$$\hat{p}_M = \frac{\sum_{s=1}^{N_S} M_{X,s} \times M_{Y,s}}{N_S} \quad (16)$$

as an estimate of the posterior probability of monotonicity.

While sampling is fast and scales well, it can be prone to error, especially when the true posterior probability of monotonicity is very low. Critically, this probability can be low even for a monotonic state trace simply because the *prior* probability of monotonicity falls rapidly as the number of conditions increases. Specifically, the prior probability of monotonicity is

$$\pi_M = \left(\frac{1}{2}\right)^{2(N_C-1)} \quad (17)$$

Although increasing N_S would result in $\hat{p}_M$ being a better estimate of the posterior probability of monotonicity, this would not necessarily scale as N_C increases. Therefore, we adopt a simple “smoothing” approach which treats each of the N_S samples of conditional derivatives as observations used to estimate parameters of a Beta distribution that reflects residual uncertainty about the true posterior probability. Our final estimate of the posterior probability of monotonicity is itself a sample from this Beta distribution, which is conjugate to prior distribution $\text{Beta}(\pi_M, 1 - \pi_M)$:

$$\tilde{p}_M \sim \text{Beta} \left(\pi_M + \sum_{s=1}^{N_S} M_{X,s} \times M_{Y,s}, 1 - \pi_M + N_S - \sum_{s=1}^{N_S} M_{X,s} \times M_{Y,s} \right). \quad (18)$$

Finally, we are in a position to compute the quantity $\hat{M}$ which reflects the degree to which the observed data increase or decrease the posterior belief that x and y are jointly monotonic. Conditioned on each sample from the posterior distribution of model parameters, we define

$$\hat{M} = \frac{\tilde{p}_M}{\tilde{p}_M + \pi_M} \quad (19)$$

that is, as the posterior *odds* that x and y are jointly monotonic. Each posterior sample of the model parameters thus yields its own value of $\hat{M}$, estimated as just described using samples of derivatives. Thus, over all samples drawn from the posterior, there is a distribution of $\hat{M}$ values that reflects the different contributions of different sources of uncertainty, including not just measurement error but between- and within-condition variability. The prior expectation of $\hat{M}$, where $\tilde{p}_M = \pi_M$, is just $\frac{1}{2}$, such that deviations above $\frac{1}{2}$ indicate evidence in favor of monotonicity while values of $\hat{M}$ below $\frac{1}{2}$ indicate evidence against monotonicity. If the bulk of the posterior distribution of $\hat{M}$ lies above $\frac{1}{2}$, we can consider this strong evidence that, even with the variability in the data, the plausible functional relationships between x and y are jointly monotonic. Conversely, if the bulk of the posterior distribution of $\hat{M}$ lies below $\frac{1}{2}$, this is evidence that none of the plausible functional relationships between x and y are jointly monotonic.

Kernel parameters

As described above, the estimate of monotonicity $\hat{M}$ is largely a function of the parameters of the covariance kernels that describe how the means of x covary with the means of y and vice versa. To support our understanding of the relationship between the kernels and monotonicity, we take this section to illustrate how the parameters of the kernel relate to the eventual estimate $\hat{M}$.

Two kernels define the covariance structure of the means across conditions: $k(\mu_{x,c}, \mu_{x,d})$ describes how the means of x in conditions c and d covary as a function of the means of y in conditions c and d ; $k(\mu_{y,c}, \mu_{y,d})$ describes how the means of y in conditions c and d covary as a function of the means of x in conditions c and d . Each kernel is associated with three parameters: the α parameter gives the total amount of variance in the means of the outcome measure; the ρ parameter dictates what proportion of the total variance is accounted for by the variation in the *other* outcome measure (i.e., how much of the variability in the means of x across conditions can be accounted for by variability in the mean of y across conditions, and vice versa); and the λ parameter describes the scale over which variation in the *other* outcome measure influences variability (i.e., how close do the means of y in conditions c and d need to be for $\mu_{x,c}$ and $\mu_{x,d}$ to be substantially correlated).

While the α parameters govern the overall variance in the means of each outcome measure, kernel parameters ρ and λ jointly capture the degree to which the means of one outcome measure can be predicted from the means of the other measure. If ρ is large, that indicates that most of the variability in the means can be explained by the variation in the means of the other measure. If λ is large, that signifies that variation in the means of the other measure exerts a broad influence on the means of the current measure. For example, the lower right panel of Figure X 3 depicts a situation in which both ρ_y and λ_y are large, such that most of the variation in the means of y is explicable by the variation of the means of x and variation in the means of x has a broad influence over many of the y means. If *either* ρ_y or λ_y is small, as in the other panels of Figures 3, the variation in the means of y is not well-accounted for by variation in the means of x . This is because either there is little variability to explain (low ρ_y) or because the range of influence is so small that little information is conveyed from one condition to the next (low λ_y).

Given that ρ and λ govern how closely related μ_x and μ_y tend to be, it would seem that ρ and λ should also be related to the degree of monotonicity between μ_x and μ_y , and indeed these kernel parameters are related to monotonicity. However, it is also clear from Figure 3 that even if ρ and λ are large, such that μ_y is closely related to μ_x (or vice versa), it does not follow that their relationship is monotonic. Large λ and ρ simply say that conditions with similar μ_x should have similar μ_y (or vice versa). In practice, the relationship between kernel parameters and monotonicity is not easy to quantify except in terms of $\hat{M}$, as defined above. While ρ has a relatively clear interpretation as a proportion of systematic covariance, λ lies on a partially infinite range $(0, \infty)$ where there is no fixed reference indicating whether it is large or small. Standardizing the data helps provide a “soft” reference, such that $\lambda = 1$ is equivalent to one standard error, but this is only an option when the data lie on an interval scale and cannot be done for, say, binomial data (see the example below for an application of the model to such data). Finally, because monotonicity is only *plausible* when both ρ and λ are large, even if there was a way to

Figure 3. Simulated values of μ_y using different parameters (λ_y and ρ_y) for the covariance kernel governing the relationship between μ_y and μ_x . In these examples, values of μ_x in each condition were randomly sampled from a normal distribution with mean 0 and variance 1, such that while μ_y depends on μ_x , the opposite is not the case. In our analyses, we assume that both mutually influence one another, the unidirectional influence of μ_x on μ_y is merely to illustrate the effect of different kernel parameters.

assess the magnitude of both ρ and λ and combine those two assessments, it is not clear that whatever the result was could be interpreted as a degree of support for or against monotonicity. Therefore, we believe the formulation of $\hat{M}$ above allows for the clearest interpretation of the resulting statistic while allowing it to be applied to a wide variety of data types.

Simulations

In this section, we demonstrate how the posterior distribution of $\hat{M}$ is affected by sample size, measurement error, number of observed conditions, and the true underlying degree of monotonicity between outcome variables x and y . For these simulations, we considered situations in which independent measurements were obtained on two outcome variables, x and y , in each of N_C experimental conditions. N_C could be 4, 9, or 16 conditions. For each condition, we randomly generated vectors of population means μ_x and μ_y , scaling each of them such that the smallest μ_x and μ_y were 0 and the largest μ_x and μ_y were 1. We then “shuffled” these two vectors of means until the Kendall’s τ rank correlation between them was equal to a target value, which was one of 0 (totally unrelated), $1/3$ (weakly correlated), $2/3$ (strongly correlated), or 1 (perfectly monotonic). Thus, a statistic designed to assess monotonicity should, ideally, discriminate situations where $\tau = 1$ from all other situations. We then simulated observations of x and y from each condition with means equal to μ_x and μ_y and standard deviation σ , where σ represents measurement error and could take values 0.32, 0.16, 0.08, or 0.04. The number of simulated observations per condition was either 20 or 80, representing different sample sizes. For each combination of N_C , τ , σ , and sample size, we simulated 500 datasets and applied the model described above to obtain a posterior distribution over $\hat{M}$. Each posterior distribution was estimated using 5000 samples from the posterior distribution over model parameters (1000 each from five parallel chains, following 1000 steps of adaptation/burn-in on each chain), with $N_S = 1000$ samples of derivatives per posterior sample.

The results of these simulations are summarized in Figure 4. Across all simulations, if $N_C = 4$, the posterior distribution of $\hat{M}$ tends to be near—and slightly above—0.5, which can be interpreted as the data not providing much evidence one way or the other about whether x and y are monotonically related. That $\hat{M}$ is centered slightly above 0.5 when $\tau = 0$ reflects the fact that, when there are only 4 conditions, it is plausible that the means of one or both outcome measure are just fluctuations around a single value, i.e., noise rather than structured covariance. Such a situation still qualifies as monotonic (or nearly so) since even derivatives that have essentially zero magnitude may be of the same sign. This is illustrated by comparing Figures 5 and 6, which show the simulated data, sample functions (μ_x vs. μ_y), inferred derivatives, and distributions of both $\tilde{p}_M$ (Equation 18) and $\hat{M}$ (Equation 19) for 4 conditions that are unrelated ($\tau = 0$, Figure fig:sim-0tau-4cond-20obs-08sigma) or monotonically related ($\tau = 1$, Figure 6). In both cases, there is a very weak tendency to consider the data as evidence in favor of joint monotonicity ($\hat{M} > 0.5$), but while it is clear that the plausible functional relationships in the monotonic case favor those that go near the observed means (Figure 6b), those in the nonmonotonic case favor “flat” functions where only one outcome varies systematically and the other can only vary due to noise (Figure 5b).

As the number of conditions increases beyond 4, the posterior distribution of $\hat{M}$

Figure 4. Average medians (lines and points) and inter-quartile ranges (shaded regions) for the posterior distribution over $\hat{M}$, averaged over 500 simulated datasets, as described in the main text.

Figure 5. Application of the analysis model to simulated data from $N_C = 4$ conditions, 20 observations per condition, $\sigma = 0.08$, and $\tau = 0$ (x and y are uncorrelated across conditions).

Figure 6. Application of the analysis model to simulated data from $N_C = 4$ conditions, 20 observations per condition, $\sigma = 0.08$, and $\tau = 1$ (x and y are jointly monotonic across conditions).

Figure 7. Application of the analysis model to simulated data from $N_C = 9$ conditions, 20 observations per condition, $\sigma = 0.32$, and $\tau = 0$ (x and y are uncorrelated across conditions).

tends to drop below 0.5 when $\tau < 1$ and rise above 0.5 when $\tau = 1$. When $\tau = 0$ and measurement noise is high ($\sigma = 0.32$), 9 conditions is not always sufficient to provide strong evidence against monotonicity, since variation between conditions is small relative to variability within conditions, making it hard to infer the form of any potential relationships between x and y (e.g., Figure 7). As measurement error decreases or there are a large number of conditions observed (e.g., $\sigma = 0.32$, $N_C = 16$, Figure 8), it is easier to identify when a state-trace provides evidence against joint monotonicity.

When the state trace evinces some degree of correlation ($0 < \tau < 1$), $\hat{M}$ correctly indicates that the evidence provided by the state trace is equivocal. For example, even in low noise conditions with large samples ($\tau = \frac{2}{3}$, $\sigma = 0.04$, $N_C = 16$, 80 observations per condition), it can be hard to tell whether the state trace is best accounted for by a monotonic function passing near each of the conditions means (which, in turn, vary in an

Figure 8. Application of the analysis model to simulated data from $N_C = 16$ conditions, 20 observations per condition, $\sigma = 0.32$, and $\tau = 0$ (x and y are uncorrelated across conditions).

Figure 9. Application of the analysis model to simulated data from $N_C = 16$ conditions, 80 observations per condition, $\sigma = 0.04$, and $\tau = \frac{2}{3}$ (x and y are partially correlated across conditions).

Figure 10. Application of the analysis model to simulated data from $N_C = 9$ conditions, 80 observations per condition, $\sigma = 0.16$, and $\tau = 1$ (x and y are monotonically related across conditions).

unsystematic fashion around this monotonic “spine”; Figure 9). Part of this reticence to reject monotonicity has to do with the extremely low prior on monotonicity (Equation 17) as the number of conditions increases—even when the state trace deviates from monotonicity, it may not provide much *additional* evidence *against* monotonicity.

When the state trace is produced from variables whose means are monotonically related ($\tau = 1$), 9 conditions are generally sufficient for the distribution of $\hat{M}$ to indicate evidence in favor of monotonicity (see, for example, Figure 10). The posterior distribution of $\hat{M}$ is not necessarily concentrated near one when $\tau = 1$, however, since even when there is considerable certainty about the relationship between the two sets of means μ_x and μ_y *within* the range observed, there can be much uncertainty about the relationship *beyond* the

Figure 11. Application of the analysis model to simulated data from $N_C = 16$ conditions, 80 observations per condition, $\sigma = 0.04$, and $\tau = 1$ (x and y are monotonically related across conditions).

Figure 12. An illustration of the problem of determining global monotonicity from local observations. If only the region in the dashed rectangle is observed, the local observations favor monotonicity even though the global relationship is non-monotonic.

observed range. This is shown in Figure 11, where samples of functions (μ_x vs. μ_y) from the posterior distribution over model parameters follow the observed data quite closely, but then “veer off” at the edge of the observed range, possibly violating monotonicity. This is an instance of a more general, almost philosophical, problem (see, e.g., Dunn & Kalish, 2018) of determining what monotonicity within a *sample* of data indicates about monotonicity of a relationship across *all possible data*, as illustrated schematically in Figure 12. Our analysis method could help address this problem, however, in that it provides not only a measure of how much evidence the data provide for monotonicity, but samples of plausible functions that could relate μ_x and μ_y in general. Thus, our methods could be a useful tool for guiding future research toward parts of the joint outcome space that are associated with greater uncertainty about their derivatives, and therefore greater uncertainty about the global relationship between the outcome variables.

In summary, we note the following characteristics of $\hat{M}$ based on these simulations:

- As the number of conditions observed increases, the distribution of $\hat{M}$ conditional on the data tends toward one for monotonically-related means ($\tau = 1$) and toward zero for sets of means that are not monotonically related ($\tau < 1$).
- The posterior distribution of $\hat{M}$ is sensitive to various kinds of uncertainty engendered by measurement error (σ), partial monotonicity ($0 < \tau < 1$), and sparsity of observations (low N_C), being wider as these different sources of uncertainty increase.
- $\hat{M}$ does not show any sign of being especially *biased*, though it can be *conservative* and “withhold judgment” (have greater posterior variability) in the presence of uncertainty.
- The number of observations per condition, so long as it is large enough (20 in these simulations) does not substantially affect the behavior of $\hat{M}$, suggesting that quality of measurement (σ) and number of conditions observed (N_C) are more important factors when trying to find decisive evidence about monotonicity.
- The ability to estimate not only monotonicity within a sample, but plausible functional relationships between sets of outcome measures, means our methods for estimating $\hat{M}$ is useful both in interpreting existing data and guiding future data collection.

Example application

Having established general features of our model for estimating $\hat{M}$ using simulations, we show how it can be applied in a practical setting using data from Dunn, Newell, and Kalish (2012). Two important extensions are made in this application: First, because the

data are proportions, we introduce a binomial likelihood function that replaces the Gaussian likelihood in the previous applications of the model. Second, the design is a mixed within- and between-subjects design, requiring an extension to model variability within individuals, as mentioned in the model description above. Finally, we note that different numbers of observations were made in each condition; while this does not affect the structure of the model in any way, this example therefore illustrates how it can be applied in such settings.

Participants in this experiment learned to assign a set of stimuli to one of four categories over the course of 4 experimental blocks of $N_T = 80$ trials each. On each trial, a stimulus would be presented, the participant would decide to which category the stimulus belonged, and the participant would then receive feedback about whether their choice was correct or not and, if not, which category had been the correct choice. For one group of participants, the categories were defined on the basis of single features such that they could perform the task using a set of simple rules. In contrast to this “rule-based” (RB) category structure, other participants had to learn a structure defined by the conjunction of two stimulus feature values, hence this structure is termed “information integration” (II), since two pieces of information must be integrated in order to categorize the stimuli correctly. Within these two groups, participants were additionally assigned to either a “no delay” condition, in which feedback was provided immediately after a category response was made, or a “delay” condition, in which a five second gap intervened between their response and feedback. It has been hypothesized that learning II category structures requires additional working memory resources that are consumed during the delay, such that learning in the delay condition should be slower for participants learning the II structure but not for those learning the RB structure.

The degree to which RB and II learning is differentially affected by delay can be assessed using a state trace plot: There are 8 experimental conditions defined by the combination of delay/no delay and learning block (1–4), and the state trace is constructed by plotting the proportion correct in RB against the proportion correct in II for each condition. This represents a mixed design, in that learning block (1–4) is a within-subjects variable, while delay/no delay and RB/II are manipulated between subjects.

Applying the model

As noted above, because the data are numbers correct out of $N_T = 80$ trials per block, we no longer standardize the data and instead express likelihood of the data in terms of a binomial distribution, rewriting Equations 1 and 2:

$$\forall_{i \in 1 \dots N_x} z_{x,i} \sim \text{Binomial}(\text{logit } m_{x,i}, N_T) \quad (20)$$

$$\forall_{j \in 1 \dots N_y} z_{y,j} \sim \text{Binomial}(\text{logit } m_{y,j}, N_T) \quad (21)$$

where each $z_{\cdot,\cdot}$ is the number correct between 0 and 80, “ x ” refers to the RB observations and “ y ” refers to the II observations, and the mean rates $m_{\cdot,\cdot}$ are transformed from $(-\infty, \infty)$ to $(0, 1)$ via the logit function. The parameters for the Gamma priors on scale parameters (as given in Equations 3, 4, 7, 8, 9, 10, and 11) were also changed to reflect the logistic scale, rather than a standardized scale, such that they have a mode and standard deviation of $\frac{\pi}{\sqrt{3}} \approx 1.81$, the scale of a standard logistic distribution, i.e., $\Gamma\left(\frac{3+\sqrt{5}}{2}, \frac{\sqrt{3}(1+\sqrt{5})}{2\pi}\right)$.

Figure 13. Results of estimating $\hat{M}$ on data from an experiment in which participants learned to categorize stimuli within either a rule-based (RB) or information-integration (II) setting. The samples of posterior derivatives in panel (b) are shown on logit scales, rather than probability scales. See main text for additional details.

To model the within-subject variation across blocks, each participant l was associated with a vector γ_l of 4 parameters reflecting their deviation from the overall mean in each block/condition, with a prior defined in terms of a vector of participant scales τ and matrix of within-participant across-block correlations Ω , as described above in the model specification (see Equation 5). To ensure that the model was identifiable, each block parameter was constrained to sum to zero across participants, such that these participant parameters indicate deflections from an overall group mean in each block.

Results

The results of applying our model to these data are shown in Figure 13, where $\hat{M}$ was estimated using 10 parallel Monte Carlo chains run for 2000 iterations each (10,000 total posterior samples), following 2000 iterations of adaptation/burn-in on each chain. These data provide strong support for a monotonic state trace, supporting the conclusion that a single factor is sufficient to explain the effects of RB and II learning regardless of delay. An important caveat, however, is that these data leave a large portion of the joint outcome space unexplored, and there remains considerable uncertainty in regions with either very high (more than 75%) or very low (less than 25%) accuracy. Thus, while the present data support a monotonic state trace, our analysis points to regions that future research can explore to determine whether monotonicity holds in a more global sense.

Partial order constraints

Because these data come from a learning experiment, it is reasonable to expect that performance should, at least on average, increase from block to block (or, at least, not *decrease* from block to block). This implies a *partial order constraint* on the means in each condition, such that $\mu_{x,i} \geq \mu_{x,j}$ if conditions i and j come from the same delay condition and block i is greater than block j . It is therefore reasonable to ask if the data still provide support for monotonicity when this constraint is applied².

²Note that this constraint is applied to the group means, not each participant. It is still possible for individual participants to violate the constraint.

Figure 14. Results of estimating $\hat{M}$ on data from an experiment in which participants learned to categorize stimuli within either a rule-based (RB) or information-integration (II) setting, under the constraint that mean performance cannot decrease from one block of the experiment to the next. The samples of posterior derivatives in panel (b) are shown on logit scales, rather than probability scales.

Applying a partial order constraint involves two steps: First, we must adjust our prior π_M to conform to the constraint, which in the case of this specific constraint implies a prior of $\pi_M = \frac{1}{16} = 0.0625$, considerably greater than the naive prior without the constraint (i.e., the prior used above, which was $2^{-14} \approx 0.00006$). Adjusting our prior not only reflects our additional knowledge expressed by the partial order constraint, it means the data must now provide even stronger evidence for monotonicity if $\hat{M}$ is to deviate from 0.5. Second, we must select only those posterior samples that conform to the given partial order constraint since, by assumption, any posterior sample that violates the constraint has zero probability. In this case, 6506 out of 10,000 samples conform to the constraint that mean performance should not decrease over blocks; this also serves as an estimate of the posterior probability (0.65) that the constraint is true about these data. We then compute $\hat{M} = \frac{\tilde{p}_M}{\tilde{p}_M + \pi_M}$ for each of those conforming samples (using the revised prior π_M).

When considering only those samples that conform to the given partial order constraint, we obtain the results in Figure 14. Even under this partial order constraint, the data provide compelling evidence for a monotonic state trace, though $\hat{M}$ is not so tightly clustered around one, owing to the fact that the prior probability of monotonicity is much higher and therefore stronger evidence is needed to achieve a more extreme $\hat{M}$. Aside from that, the difference between constrained and unconstrained analyses do not substantially differ given that most of the original posterior samples conformed to the constraint already.

In principle, any partial order constraint can be imposed by following the steps above: After running the unconstrained model, one first computes the prior probability of monotonicity under the constraint, then selects those samples from the unconstrained posterior that conform to the constraint, and finally recomputes the distribution of $\hat{M}$ using those samples and the revised prior. In practice, stricter constraints mean that fewer of the original posterior samples will be selected, potentially making it difficult to estimate the full distribution of $\hat{M}$. In such settings, an easy solution is simply to get more samples from the unconstrained posterior. But as noted above, the proportion of samples that conform to a constraint is also an estimate of the posterior probability (given the data) that the constraint is true. Thus, if few posterior samples conform to a constraint, it may be that

the constraint itself is ill-posed or impossible to realize.

Discussion

Obtaining an estimate of the amount of evidence a given data set provides for or against monotonicity of its underlying data generating process is challenging. We have presented a model for estimating the statistic $\hat{M}$, an estimate of how much evidence a set of data provides for the belief that two (or more) outcome measures are monotonically related, and can therefore be explained in terms of a single underlying factor or dimension. This model uses a Gaussian process representation of the underlying relationship between the two outcome measures, with this process parameterized by a covariance kernel. The kernel allows for joint estimation of the means of each outcome measure in each experimental condition as well as their derivatives with respect to one another. In a series of simulations, we showed that $\hat{M}$ is capable of discriminating between situations where the state-trace is and is not monotonic, and is sensitive to the degree of uncertainty inherent in finite data. In our supplemental materials, we provide model code using the R (R Core Team, 2017) interface to the Stan language (Carpenter et al., 2017) so that researchers can compute $\hat{M}$ using data from both between-subjects and within-subjects designs. This code includes our mixed-design example, as a reference aid for researchers to ensure that they are using the software correctly.

The kernel-based method for computing $\hat{M}$ is widely applicable across different experimental designs and kinds of data and, using Stan (Carpenter et al., 2017), can be done in a matter of minutes on modern personal computers, particularly if parallelization is available (all computations in this paper were done on a typical modern laptop with 8 cores). Our method thus represents an advance over both existing Bayesian (Prince et al., 2012) and frequentist (Kalish et al., 2016) methods for assessing monotonicity that can require a large degree of computation and have difficulty scaling up as the number of experimental conditions increases. Our methods are also not restricted to only a particular type of data, such as real-valued or proportional outcomes. Finally, by estimating candidate functions that relate the two sets of outcome measures, our methods can help guide future research toward areas of the joint outcome space that are associated with greater uncertainty.

Although our exposition has focused on cases with just two outcome measures—which is typical of most state-trace analyses—there is no reason why our methods should be restricted to the case. Indeed, the code we provide in the supplementary materials allows for any number of joint outcomes to be analyzed, such that $\hat{M}$ can be used to estimate the evidence for or against any set of outcome variables being jointly monotonic. As more outcome variables are included, however, the prior probability that they are all jointly monotonic plummets rapidly, to the point where minor numerical instabilities can have a strong influence on the analysis. Thus, it seems to us that situations with a large number of outcome variables all but require the imposition of partial order constraints, thus focusing the analysis on subregions of the joint outcome space that are likely to be more informative. Such constraints, of course, must be motivated by researchers’ domain-specific knowledge about the experimental conditions and data collection methods under analysis. Thus, the difficulties that arise with more than two outcome variables seem more to do with interpreting the results than with implementing the analysis.

Another potential realm of future work can focus on the eigenvalues and eigenvectors of the estimated kernels that describe the covariance between outcome measures. Just as principal components analysis can yield evidence regarding the number of dimensions needed to reconstruct a data matrix (Jolliffe, 2002), it may be possible to use the eigenvalue spectra of each kernel to decide how many latent dimensions are required to describe the covariance between a set of outcome measures, thus providing evidence not just about the adequacy of a single dimension (i.e., a monotonic state trace) but of *how many* dimensions (if greater than one) might be required. Unfortunately, analytic expressions of the eigenvalue spectrum of kernel functions are available only under highly restricted conditions, necessitating numeric approximations. Still, we expect this to be a fruitful avenue for future work, with the model presented here being an important first step along that route.

We close by emphasizing that, whether a state trace is or is not monotonic, $\hat{M}$ only tells a researcher about whether a single factor is *sufficient* to explain their data. It may be that multiple causal factors are actually impinging upon the data, but either their effects are small or they are not strongly affected by experimental manipulations, such that recourse to multiple factors provides no explanatory value for a given set of data. On the other hand, a non-monotonic state trace does not indicate how many factors are needed to explain the data—although it is at least two—nor does it say what those factors might be. In all these cases, it remains incumbent upon researchers to use their own domain expertise to provide a theoretically-motivated explanation for the results of any state-trace analysis. The present work helps to make such analysis accessible and to provide statistical tools with which a researcher can properly express their confidence in the results of the analysis.

References

- Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56, 149–178.
- Bamber, D. (1979). State-trace analysis: A method of testing simple theories of causation. *Journal of Mathematical Psychology*, 19(2), 137–181.
- Brainerd, C. J., & Reyna, V. F. (2010). Recollective and nonrecollective recall. *Journal of Memory and Language*, 63(3), 425–445.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).
- Dunn, J. C., & Kalish, M. L. (2018). *State-trace analysis*. Cham, Switzerland: Springer.
- Dunn, J. C., Newell, B. R., & Kalish, M. L. (2012). The effect of feedback delay and feedback type on perceptual category learning: The limits of multiple systems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(4), 840–859.
- Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10), 454–459.
- Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). New York, NY: Springer.
- Kalish, M. L., Dunn, J. C., Burdakov, O. P., & Sysoev, O. (2016). A statistical test of the equality of latent orders. *Journal of Mathematical Psychology*, 70, 1–11.
- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100, 1989–2001.
- Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted, 2007. *Psychological Review*, 114(1), 188–202.
- Prince, M., Brown, S., & Heathcote, A. (2012). The design and analysis of state-trace experiments. *Psychological Methods*, 17(1), 78–99.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>