Research Article

Statistics in Medicine

Received 4 December 2015, Accepted 27 September 2016 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.7151

Change-point detection of cognitive states across multiple trials in functional neuroimaging

F. Spencer Koerner,[a,b*†] John R. Anderson,[a,c,d] Jon M. Fincham[c] and Robert E. Kass[a,b,e]

Many functional neuroimaging-based studies involve repetitions of a task that may require several phases, or states, of mental activity. An appealing idea is to use relevant brain regions to identify the states. We developed a novel change-point methodology that adapts to the repeated trial structure of such experiments by assuming the number of states stays fixed across similar trials while allowing the timing of change-points to change across trials. Model fitting is based on reversible-jump MCMC. Simulation studies verified its ability to identify change-points successfully. We applied this technique to data collected via functional magnetic resonance imaging (fMRI) while each of 20 subjects solved unfamiliar arithmetic problems. Our methodology supplies both a summary of state dimensionality and uncertainty assessments about number of states and the timing of state transitions.

Keywords: change-point detection; Bayesian inference; reversible-jump MCMC; functional magnetic resonance imaging; segmentation

1. Introduction

A major use of functional neuroimaging is to describe brain activity, while subjects perform cognitive tasks carried out repeatedly across multiple trials. The object of this imaging is often diagnostic, sometimes relating functional activity to characteristics that can help identify or characterize diseases [1], but most applications aim to gain psychological insight into the processes that produce task-related behavior. Many tasks may be decomposed, intuitively, into discrete stages such as planning and execution, and a natural question is whether it is possible to identify discrete states of brain activity that might correspond to these discrete stages of task performance. To address the question, we developed a Bayesian change-point detection method in the context of multiple repeated trials of subjects solving arithmetic problems, where the trials may be clustered into groups having similar series of brain activity recordings. These clusters would be interpreted as trials on which the subject proceeded via the same sequence of brain states, with some variation in timing.

Statistical change-point detection methods have been studied thoroughly [2] and applied in a variety of biomedical contexts (e.g., [3]), including neuroimaging [4]. Much of the previous work in the field of change-point detection requires stringent assumptions, like the true number of change-points, in order to estimate distributions on the locations of the change-points. A recent contribution [5] treats the number of change-points as a random variable, the distribution of which is estimated using sequential Monte Carlo. In our situation, the data come in multiple trials, and there is substantial variation across trials, as may be seen in Figure 1. As the figure indicates, change-point detection in this context is challenging because of

[a]*Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[b]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[c]*Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[d]*Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[e]*Department of Machine Learning, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[*]*Correspondence to: F. S. Koerner, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA, U.S.A.*
[†]*E-mail: fsk@andrew.cmu.edu*

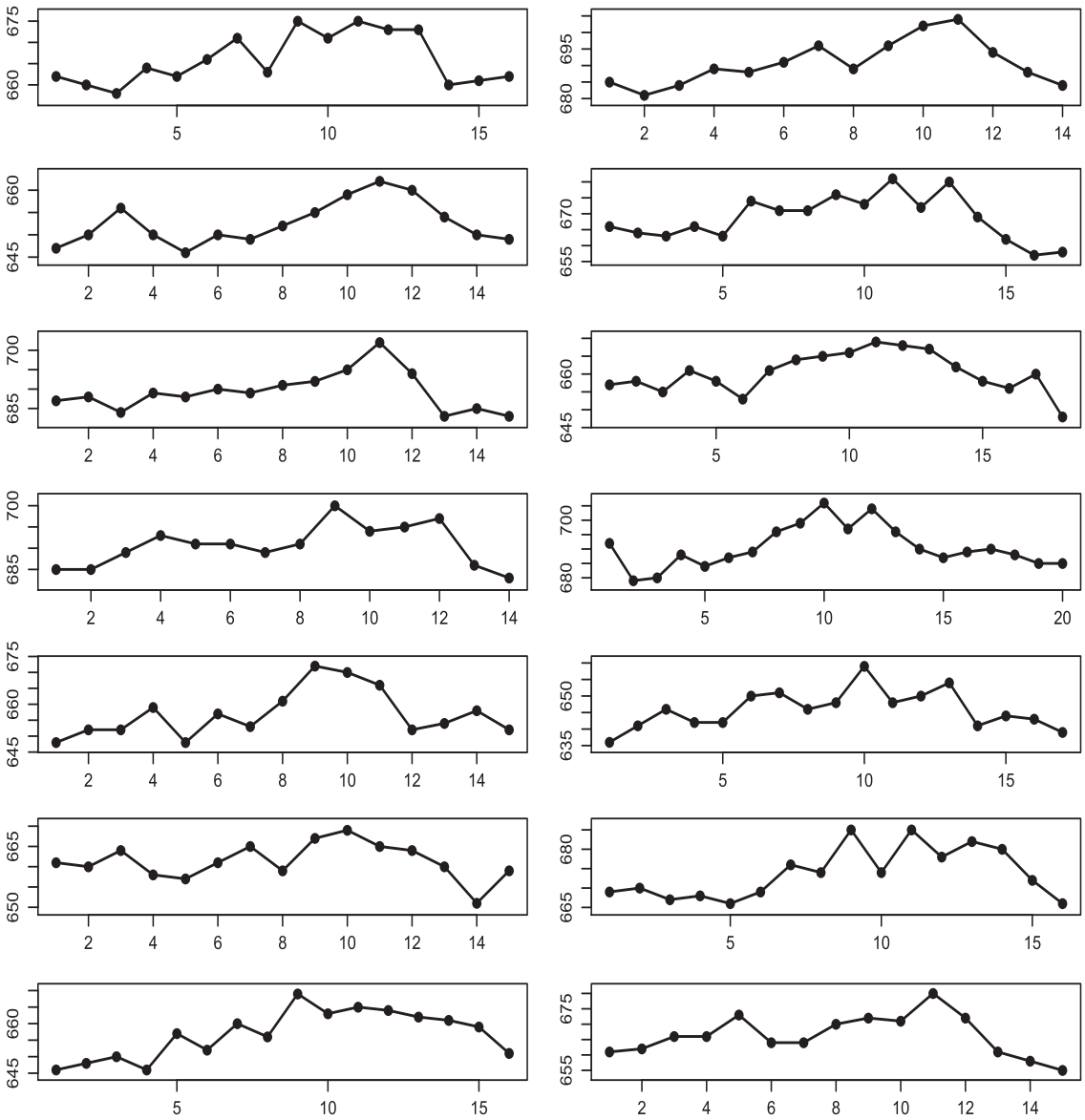

Figure 1. Activity of a voxel central within one subject's left horizontal intra-parietal sulcus for the duration of 14 similar arithmetic problems. Trials are substantially discrepant.

both noise and the natural variation in the way the task is performed across trials. Our approach uses a hierarchical model to leverage the information from multiple, similar trials while allowing the location of the change-points to vary across these similar trials. We implemented the method using reversible-jump MCMC [6] while incorporating a step to cluster the trials based on number of change-points. In addition to segmenting each trial into a small number of brain activity states, this method provides uncertainty in the form of posterior probability. In this article, we formulate two models, piecewise constant and continuous piecewise linear, derive the reversible-jump algorithms, and report results from a numerical simulation study showing that the method can correctly identify underlying states from reasonably sized data sets. We then illustrate the methodology by applying it to fMRI data collected during an experiment on solving algebraic problems [7, 8].

We first explain the experimental context for the fMRI data in Section 2, before we present the models and their implementations in Section 3. We describe our use of functional principal components to reduce dimensionality in Section 4 and report results from our simulation study in Section 5. We then analyze the original data in Section 6, and Section 7 provides our overall conclusions about the utility of the methodology.

2. The motivating problem and experiment

The *adaptive character of thought-rational* (ACT-R) model of cognition, as detailed and motivated in [9] and [10], accounts for the brain's progression through cognitive tasks as an optimal decision-theoretic production system. It also serves as the foundation for currently implemented technology. Cognitive tutors, as detailed in [11], help teach techniques like arithmetical manipulation; they automatically track a student's likely reasoning through an exercise via user input at each step. For effective use with problems allowing for more lateral thinking, we need to be able to determine with greater accuracy where in the problem-solving process the student is, in order to track effectively and provide appropriate feedback. Through functional imaging, we hope to be able to infer and/or verify the evolution of more general problem-solving processes and, through change-point detection, identify discrete states within the task that relate to concrete steps in arriving at an answer. As addressed in [12], models detecting discrete change-points can be used effectively for 'model tracing', that is, inferring a student's current progression and intentions based on his or her actions. However, many existing methods do not include probabilistic statements on the number and location of these changes. The methods proposed here specify where a change occurs in each trial's progression while capturing structure common to groups of similar problems.

In exploring how both familiar and unfamiliar problems are solved, the Anderson lab has used a novel type of arithmetic task based on what they call *pyramid problems*. These problems were developed for their facility in testing both cognitive processes and, mainly through exception problems, metacognitive processes, which are more abstract, and associated with higher order concepts, like reflection and incorporating current information to strategize [7]. In these problems, the subject solves for a variable in an equation involving the \$ operator. An expression of this form is evaluated as follows:

$$x\$y = x + (x-1) + \ldots + (x-y+1).$$

Subjects are given a problem of the form: $z = x\$y$ and are asked to solve for one of the variables (which have been deemed the *value*, *base*, and *height* of the problem, respectively). So, for example, given

$$z = 4\$3,$$

the solution would be

$$4 + 3 + 2 = 9.$$

This would be referred to as a *value* problem, whereas *base* and *height* problems would be those in which the question posed has the respective term given as a variable. Together, these three problem subtypes are called *regular* problems. On occasion, a subject may be given a slightly different type of problem termed *exception* problems. These problems are more difficult, as they might involve a larger number of terms, a larger magnitude of the terms, negative terms, fractional bases, or repeated variables. Each of the following three problems would be an example of an *exception* problem.

$$z = 127\$ - 3,$$

$$z = -9.5\$4,$$

$$z\$4 = z.$$

It would be particularly interesting to infer where change-points occur in these problems. The probable number of change-points and their locations could both reflect the differing strategies adopted by subjects in varying problem types. We consider both correctly solved regular and exception problems in our analyses.

The data were collected from 20 subjects each solving 120 problems in six blocks of 20 problems each. We have variables characterizing the nature of each problem solved: the subject, block, type of problem, the given base and height, whether the subject correctly solved the problem, and its starting and ending positions in a sequence of 41,635 scans. The imaging was performed via gradient echo planar image acquisition, using a Siemens 3T Allegra scanner with a standard radiofrequency head coil (quadrature birdcage). Repetition time was 2 s (30-ms echo time, 70° flip angle, and 20-cm field of view). The

experiment acquired 34, 64, and 64 (horizontal, coronal, and sagittal) slices per repetition time resulting in voxels of size $3\frac{1}{5}$mm $\times$ $3\frac{1}{8}$mm, $\times 3\frac{1}{8}$mm, respectively. The anterior commissure–posterior commissure line was on the 11th slice from the bottom scan slice. Functional images were motion corrected using the six-parameter 3D automated image registration, as via [13]. Images were co-registered to a common reference structural MRI via a 12-parameter 3D registration and smoothed with a 6mm full-width-half-maximum 3D Gaussian kernel. Each subject's entire sequence of scans was independently deconvolved with the canonical statistical parametric mapping (SPM) hemodynamic response function (a difference of four gamma densities), for each voxel independently, as there is no natural foothold for more rigorous stimulus-response-based hemodynamic response function (HRF) estimation. Each scan was down-sampled to 408 mega-voxels spanning the brain, a much lower spatial resolution than fMRI is capable of (for reference, about 47,000 voxels are active given the more canonical original resolution); however, the smoothing within mega-voxels can reduce noise and has been useful in relating ACT-R to brain activity (see [8] for a previous analysis of the data examined here, and more details of the data collection, as well as [14] for related background, including demarcations of the areas ACT-R implicates). The data have been deconvolved with a canonical HRF.

3. Change-point models

To segment out substantial changes in neural activity, we need to make precise what, if any, assumptions might be made about the relationships between trials. We could neglect the possibility of any shared information and treat each trial independently. However, determining change-points would then be subject to the sizable magnitude of noise within a single set of trajectories of BOLD over the course of one problem-solving process. Even within the same task-related megavoxel, as shown in Figure 1, trials elicit very different temporal activation. Instead, we use a functional version of principal components analysis to reduce the original 408-dimensional time series to much lower-dimensional series, maintaining a large proportion of variability across regions. An example of the first mode of variability in such a projection is shown in Figure 2, for 14 trials. We take advantage of the similarity in profiles across repetitions (trials) made apparent by this reduction. Rather than making strong assumptions about the relationship of each trial to prototypical functional forms, we allow trials to cluster into groups exhibiting similar temporal progressions.

First, we define some notation. Given a set of N univariate time series (of possibly differing lengths), we denote the observed value corresponding to trial i at time t by $X^{(i)}(t)$, and we let $\left\{\tau_j^{(i)}\right\}_{j=1}^{k}$ be the set of k values of t that define *change-points*, $\left\{\tau_j^{(i)}+\frac{1}{2}\right\}_{i=1}^{k}$, where the distribution of $X^{(i)}(t)$ changes. The collection of k change-points, $T^{(i)}$, segments the series into $k+1$ *states*. We use the convention $\tau_0 = 0$ and define $l^{(i)}$ to be the length (in images) of trial i. As we will require at least two data points occupy any state, we then know $\tau_j^{(i)} \in \{2, 3, \ldots, l^{(i)}-2\}$, for all i and j. To take a hierarchical approach, we write the mean functions, $\mathbb{E}[X^{(i)}(t)]$, as $f_j^{(i)}(t)$, for the jth state, and take them to be piecewise linear between change-points and continuous at change-points, with normal priors on the slope coefficients, $\beta_j^{(i)}$, and a normally distributed additive constant for each trial, $\gamma^{(i)}$. For the purposes of model fitting, each trial belongs to one of C disjoint collections of trials, specified by $\mathcal{C}_1, \ldots, \mathcal{C}_C$. We allow unique sets of priors on slopes, specific to each collection. Thus, parameters of trial i will be assumed generated by parameters of the cluster indicated by $c_i \in \{\mathcal{C}_1, \ldots, \mathcal{C}_C\}$. That is, trial-level slopes will be normally distributed around $\beta_j^{(c_i)}$, which are, in turn, assumed distributed via a Gaussian hyperprior. For cluster h, a Poisson prior is placed on the number of change-points, $n_{\mathcal{C}_h}$, a Dirichlet prior is placed on their locations with concentration parameters specified by the vector $\boldsymbol{\alpha} := \{\alpha_1, \ldots, \alpha_{n_{\mathcal{C}_h}}\}$, and for each trial, cluster membership is uniform, *a priori*. These clusters of trials allow us to fit models of differing dimensions across the entire set of trials. We, therefore, have the following model:

$$
\begin{aligned}
X^{(i)}(t) &\sim \mathrm{N}\left(f_j^{(i)}(t), \sigma^2\right) \text{ for } t \in \left\{\tau_{j-1}+1, \ldots, \tau_j\right\}, \\
f_j^{(i)}(t) &= \gamma^{(i)} + \sum_{k=1}^{j-1} \beta_k^{(i)} \left(\tau_k^{(i)} - \tau_{k-1}^{(i)}\right) + \beta_j^{(i)} \left(t - \tau_{j-1}^{(i)} + \frac{1}{2}\right), \\
\left(\tau_1^{(i)}, \ldots, \tau_{n_{c_i}}^{(i)}\right) \mid n_{c_i} &\sim l^{(i)} \cdot \mathrm{Dir}\left(\alpha_1, \ldots, \alpha_{n_{c_i}}\right), \\
\beta_j^{(i)} \mid \beta_j^{(c_i)} &\sim \mathrm{N}\left(\beta_j^{(c_i)}, \phi^2\right), \\
\gamma^{(i)} &\sim \mathrm{N}\left(0, \psi^2\right), \\
\beta_j^{(\mathcal{C}_h)} &\sim \mathrm{N}\left(0, \chi^2\right), \\
n_{\mathcal{C}_h} &\sim \mathrm{Pois}(\lambda), \\
c_i &\sim \mathrm{Unif}(\mathcal{C}_1, \ldots, \mathcal{C}_C),
\end{aligned}
\tag{1}
$$

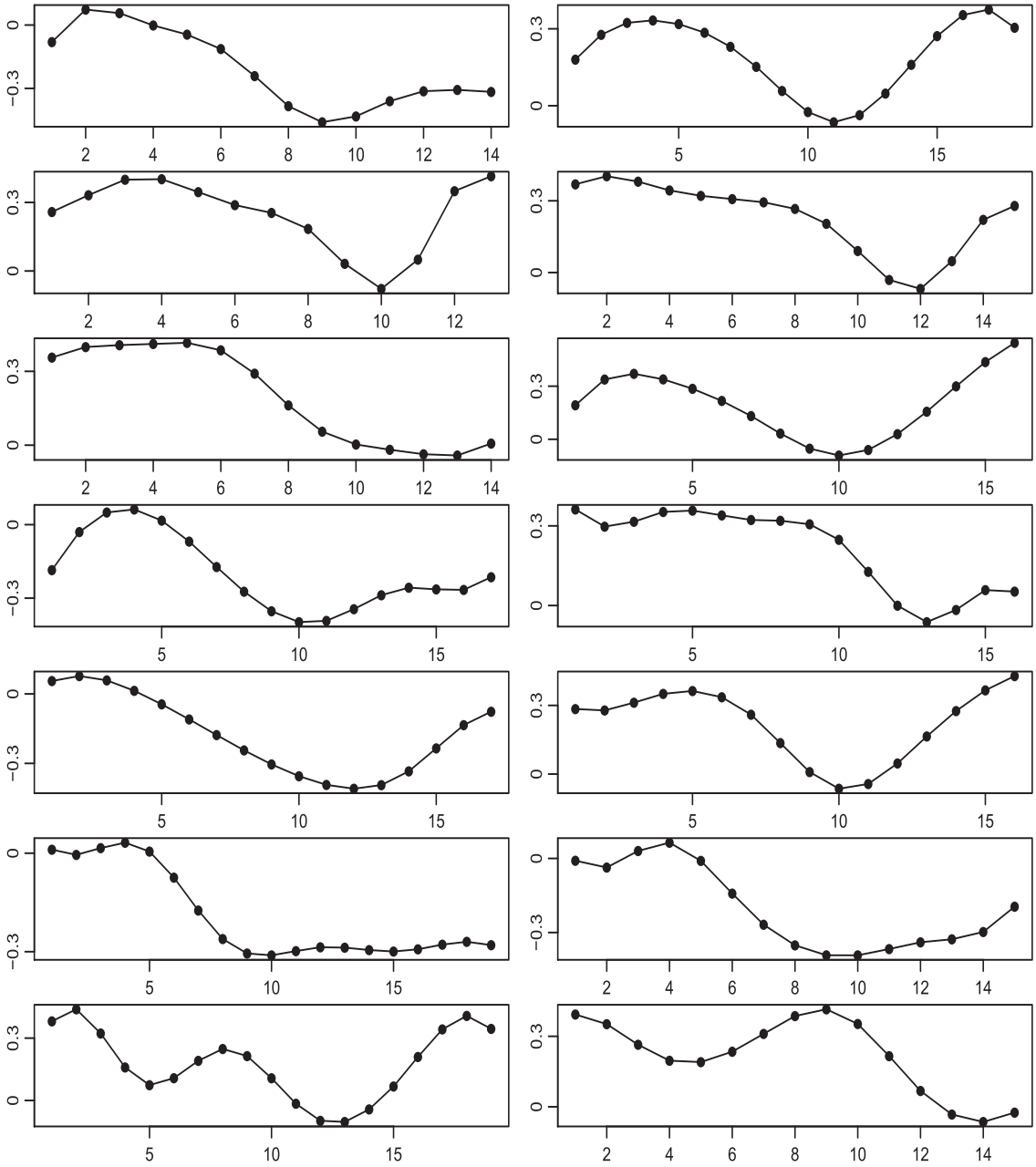

Figure 2. The first non-trivial functional principal component of the same 14 regular trials shown in Figure 1. Note that potential change-points are more readily apparent than in the much noisier raw data in Figure 1.

where for our purposes, $\sigma, \boldsymbol{\alpha}, \phi, \chi, \psi, \lambda,$ and C will be fixed; heuristics for finding appropriate hyperparameters will be discussed in Section 3.3. Given this hierarchy, if an estimated cluster contains multiple trials, we can leverage information between them to better estimate change-point locations, in isolation from other clusters.

In higher dimensions, continuity between states is a restrictive assumption. A geometric argument can reduce a model of piecewise hyper-planes that are continuous at change-points to a combination of a dimensionality reduction problem and fitting the continuous piecewise linear model to a one-dimensional projection, hence the importance of model (1). A useful generalization arises when we no longer require continuity; we can then specify hyper-planes that vary much more meaningfully. To be computationally tractable, we simplify to the constant case in higher dimensions, instead specifying $f_j^{(i)}(t) = \vec{\gamma}_j^{(i)}$ and retaining the same priors on change-point locations, number of states, and cluster membership, and now assuming multivariate Gaussian priors, where applicable:

$$\begin{aligned} \vec{X}^{(i)}(t) &\sim \mathrm{N}\left(\vec{\gamma}_j^{(i)}, \Sigma\right) \text{ for } t \in \{\tau_{j-1}+1, \ldots, \tau_j\}, \\ \vec{\gamma}_j^{(i)} \mid \vec{\gamma}_j^{(c_i)} &\sim \mathrm{N}\left(\vec{\gamma}_j^{(c_i)}, \Phi\right), \\ \vec{\gamma}_j^{(c_i)} &\sim \mathrm{N}(0, \Psi). \end{aligned} \tag{2}$$

Note that we do not allow for dependence between parameters at the state level within a trial except through the intermediating cluster-level parameters, where the dependence arises through posterior sampling from the combination of change-point distributions, which may necessarily allocate observations from bordering states to a given state of interest, as well as through cluster membership distributions. While we do not offer explicit quantifications of theoretical biases induced by the likely forms of misspecification our models may suffer, we do attempt to verify they are not too deleterious via simulations in Section 5 and Appendix D.

3.1. Model fitting

The complexity of either model's full posterior distribution precludes analytically calculating any statistics of interest. We turn to reversible-jump Markov chain Monte Carlo, a stochastic method for sampling from an arbitrary distribution on a support that may change in dimension with the parameter values. Convergence of the chain's stationary distribution to the target distribution requires the construction of a reversible transition kernel for potential jumps between parameter sets, as [6] originally details (see also [15–17]). Transitions that retain the parameter space of the chain's current state may be attempted via the usual methodology, for example, vanilla Metropolis–Hastings or Gibbs sampling. However, as we treat the number of change-points in a model as random, we must specify a one-to-one transition kernel in transdimensional steps, allowing us to choose a positive dominating measure on an augmented parameter space (which will be determined in the process) to facilitate dimension matching.

The subtlety in *rjMCMC* algorithms often lies in the construction of jump proposals that can both maintain reversibility and explore the posterior in a computationally efficient manner. Various ways to track and reincorporate information about the progression of the chain and its successful jumps into new proposals have been explored by [18]. Alternatively, we proceed to tailor a transformation to the problem's structure, because gains to computational efficiency can be substantial with effective transitions. Let θ' represent a potential set of values for the parameter θ to take. For model (1), we need to be able to propose the following:

- $\beta_j^{(i)} \to \beta_j'^{(i)}$,
- $\beta_j^{(\mathcal{C}_h)} \to \beta_j'^{(\mathcal{C}_h)}$,
- $\gamma^{(i)} \to \gamma'^{(i)}$,
- $T^{(i)} \to T'^{(i)}$,
- $c_i \to c_i'$,
- $n_{\mathcal{C}_h} \to n'_{\mathcal{C}_h}$.

Proposals of all but the last two types are implemented through standard Metropolis–Hastings steps, as well as $c_i \to c_i'$, when $n_{c_i} = n_{c_i'}$. We need to implement reversible-jump methods in the transdimensional

cases: when proposing a change in the model generating a trial, if the current and proposed models exhibit different numbers of states, or when proposing a new number of change-points in a cluster. In either of these transdimensional cases, we must define an invertible transformation that takes the current state of the chain and a number of random samples equal to the difference in dimensionalities of original and proposed model spaces. (In the case of decreasing the number of states, this function will allow calculation of the draws that would have been necessary to generate the analogous reverse jump, i.e., that of the proposal of increasing the number of states; while both directions can be functions of random draws, it is often simpler to make one direction deterministic, if possible.) We then include the corresponding densities at these draws, relative forward and backward probabilities of these moves, and the Jacobian of the transformation as factors in the acceptance probability.

We provide detail about one proposal type in Appendix A, that of an increment to the number of states in a sub-model (the model specified by the parameters inherent to a cluster), and the similarly implemented transition $c_i \rightarrow c_i'$ in Appendix C. Transitions for model (2) are simplifications of those for model (1) and are clear upon disregarding the irrelevant parameters in the discussion of the Jacobian of the last enumerated transition type in Appendix B. Section 8 includes information on publicly available code for implementation.

3.2. *Initialization and estimation*

We can design simple heuristics for initialization of parameters near local maxima in the posterior. However, depending upon hyperparameter values, without proper tuning of the proposal distributions, the dimensionality of the data can be large enough that the information locally is sufficient to keep the chain near these local optima for many more iterations than is computationally reasonable. With sufficient tuning, acceptance probabilities can be large enough to allow arbitrary starting values for all parameters. Naturally, subsequent burn-in will be required.

When the chains have unambiguously reached their limiting distribution, we rely on exploring the parameter space through a sequence of conditioning steps. As $\mathcal{C}$-level estimates are not only subject to varying in dimension, but also theoretically identically distributed between clusters, attributing any definite psychological interpretation to an individual cluster may be misguided. To make assertions exclusive to a cluster of trials, one would first identify whether the group is unique among clusters locally in the chain, as the cluster may have other likely member trials. (For C large enough to computationally efficiently run the rjMCMC, it will often be the case that multiple clusters will have similar parameter estimates. With fewer clusters allowed, the necessarily increasing heterogeneity among trials in a cluster will eventually, effectively prohibit model-level transdimensional jumps.) Therefore, to find stable estimates of remaining parameters, we first maximize posterior probability over the marginals $p_i(c_i, n_{c_i})$, for each trial, to find jointly the estimates $\{\widehat{c_i}\}_i$ and $\{\widehat{n_{c_i}}\}_i$. This makes possible further inference (i.e., estimates of the remaining parameters are not well defined over differing numbers of states); we then estimate, via conditional maximum a posteriori (MAP) estimates, $T^{(i)} \mid \widehat{c_i}, \widehat{n_{c_i}}$ and via the state-specific conditional posterior means, $\vec{\beta}^{(i)}, \gamma^{(i)} \mid \widehat{T^{(i)}}, \widehat{c_i}, \widehat{n_{c_i}}$. Only the clusterings, number of states, and locations of change-points have a clear meaning, and only locally within the chain, rather than indefinitely once converged to the limiting distribution, as in many applications.

3.3. *Hyperparameters*

The model requires *a priori* specification of $\sigma, \boldsymbol{\alpha}, \phi, \chi, \psi, \lambda,$ and C. Using plausible estimates of these values in real data leads to much more stable fits and useful discriminability between overtly differing groups of trials. As will be seen via the constraints in 4, the nature of the data-reduction process results in input time series s.t. $\forall i,\ \int (X^{(i)})^2(t)dt = 1$. The heuristics used generalize well to any $X^{(i)}(t)$ that are standardized, unit length under the ℓ^2 norm, or have a similar magnitude scaling (and deteriorate especially, with the magnitudes of the differenced time series or probable ratio of change-points to time points). We let χ, ψ specify very uninformative hyperpriors (both 0.5 in our analyses), along with $\boldsymbol{\alpha}$, which we take to be the concentration vector of 1s. Estimates are not especially sensitive to λ; as results are largely indistinguishable over the range of $\lambda \in (1, 10)$, we set it to 2.5.

We take the number of clusters to be $C = 6$ for the simulated experiments and $C = 9$ for fitting real data. With too small a C, we are guaranteed overly heterogeneous cluster membership, and thus highly biased estimates of change-point location (in an unpredictable manner), as well as upward biased estimates of the number of states. Given sufficiently large C, similarity between trial collections will

fall out into overtly similar cluster-level parameter estimates, but neglecting to merge models does not adversely affect estimates of the number of change-points. We can use the algorithm to approximate the number of clusters in the most parsimonious set of distinct functional models as well, simply by repeatedly reducing C by 1 (a naturally agglomerative clustering method on the most similar collections of trials) until the algorithm converges to a stable set of $\widehat{c}_i$ (indicating that no two clusters are effectively modeling the same set of trials).

Estimates are most sensitive to the remaining hyperparameters, σ and ϕ. We use the sample variances of the series after removing conservative smoothing spline fits, averaged over all trials, for σ^2. Lastly, we will tend to overfit with very poor values of the hyperparameter ϕ; with limited C, underspecifying ϕ will greatly penalize trial-level deviations from model-level parameters and increase split probability, even with a good partitioning of trials; likewise, overstating ϕ will lead to otherwise marginal trials moving to the sub-model, also facilitating more splits. We find that ϕ an order of magnitude less than σ tends to perform well for similar time series (in length, and the smoothness and scaling inherent to the functional principal components analysis (fPCA) basis). In any case, erring on the side of caution (noninformativity) is wise, in application.

4. Reducing the data

Dimensionality reduction techniques rely upon the concept of the data being generated in some low-dimensional space of interest, before being projected to a higher-dimensional space in which we observe it. Redundancy in the information conveyed by the covariates measured in most complex systems makes these methods effective. The brain is no exception, and fMRI data, in particular, have been conducive to these attempts at making low-dimensional analysis tractable [19] while ameliorating the non-negligible multicollinearity problems inherent to modeling the unreduced data. We adopt one such approach, namely, a functional PCA-based representation. The large variability in the BOLD response of a single voxel, even across identical tasks and an ostensibly similar solution progression, is not likely due merely to additive noise and requires a method treating it as salient.

Functional principal component analysis is a natural extension of PCA to functional data. In our context, the imaging times would correspond to covariates, and voxels to samples or observations. The fPCA approach presumes images are taken from a smooth underlying function of blood oxygenation over time, and therefore, the limit as the number of covariates (times at which we interpolate) becomes dense is computationally tractable. This limit can be formalized as an eigendecomposition of the analog to the usual sample covariance matrix $\hat{\Sigma}$, the temporal covariance surface [20]:

$$\hat{\sigma}(s,t) = \sum_{v=1}^{V} (X_v(s) - \bar{X}(s))(X_v(t) - \bar{X}(t)),$$

where X_v represents the vth voxel's hypothetical continuous time BOLD response. Just as multivariate principal components, $\xi_j = (\xi_{j1}, \ldots, \xi_{jT})$ would maximize projected variability across voxels: $\max_{\xi_j} \mathbb{V}(\xi_j X')$, where X is the V by T data matrix, these functional principal components, $\xi_j(t)$, are similarly meant to explain variability across voxels over time:

$$\max_{\xi_j(t)} \mathbb{V}\left(\int \xi_j(t) X'(t) dt \right),$$

subject to the usual orthogonality and unit L^2-norm constraints, and are what will be used to summarize a trial. In doing so, the components capture what are generally conceptualized as temporal modes of variability, manifesting as the smooth trajectories of somewhat consistent deviations (in either direction) from mean activity, that upon projection, elicit (explain) the most variability across voxels. Again, by contrast to PCA, wherein researchers often interpret components by looking for relatively large loadings and characterize them as the low-dimensional, predominant feature groups explaining variability across observations, sustained departures from 0 in the fPCs would correspond to epochs of the trial duration during which perturbations from mean neural activity by some subset of regions explain substantial proportions of voxel-wise variability. Alternatively, consider a set of fPCs as a set of smooth, constrained functions which, when allowing the linear combinations thereof, optimally reconstruct the original data. In finding these functions, we automatically diminish the impact of voxels that generally vary little from

mean, or measure mostly noise (with no discernable temporal structure), and weight more heavily those with some apparent signal in determining a mode.

Dependence between trials is, fortunately, of no concern until change-point detection, as trials are reduced independently. Treating voxels as independent, however, is an assumption that should be made very tenuously. Some features of physiology, like low-frequency changes in heart-rate or respiration, will manifest obviously via the BOLD mean processes and would then be accounted for via the fPCA preprocessing. However, despite voxels in disparate regions of the brain representing processing performed in the service of transforming differing inputs in very different ways, substantial spatial correlations, usually attenuating with distance, underlie all recorded activity. Whether other subtler effects could be even more damaging to suitable low-dimensional representations is unclear, but the spatial correlation alone should be a primary consideration. Reference [21] even considers a change-point problem on the (one-dimensional) space on which functional observations are then sampled. Given the abrupt changes in function exhibited on the space of the brain (specific types of processing are known to be very localized), this would be an appropriate problem to consider in determining regions of interest. Generalizations to arbitrary higher-dimensional spaces are, however, non-trivial. Reference [22] studies the effects of non-linear temporal dependencies in great detail. Specifically, they consider the effects of reasonably general dependence structure on the estimation of functional principal components. In fact, one example in their examination is of change-point detection, and the biases potentially introduced by disregarding estimation of long-run covariance in place of the standard covariance. They also propose a test statistic for change-point detection in a process's expectation with appealing properties. However, they do not explicitly examine most results' performances (including that of the change-point test) under very short time series, where functional data analysis is less commonly employed. They do show that estimation of fPCs can be performed consistently and is somewhat robust to dependence. As expected, though, the fPC reduction cannot then capture all of the process's structure. For example, complex systems might admit effective low-dimensional representations that change over time. One appropriate step in modeling might be, therefore, adding a level of dynamicality to the basis. Reference [23] offers a powerful solution in the case of a one-dimensional underlying space, via dynamic functional principal component analysis. While perhaps spatially dynamic functional principal component representations would be more appropriate, these methods have not been explored nor would it be clear how to test for arbitrary classes of change-surfaces. To limit complexity of preprocessing, we proceed as if an fPCA representation is sufficient and examine reconstruction performance in our example incorporating some spatial dependence.

As this projection is performed for each trial individually (bearing in mind each component is unique only up to itself and its negation), trials with naturally similar sources of variability will be represented by similar basis functions. This is evidenced by the fact that identical projections of only voxels within areas such as the posterior superior parietal lobule, angular gyrus, and lateral inferior prefrontal cortex elicit similar basis functions, whereas subsets of the majority of the remainder of cortex provide more erratic projections that are dissimilar across trials.

It should be noted that further preprocessing prior to fPCA is often implemented in attempts to achieve stationarity. Model (2) does not allow for underlying trends; however, under model (1), a trial-wise linear background trend will be fit automatically. Nonetheless, for consistency, a mean process is always removed prior to eigendecomposition. Detrending could be crucial for resting state analyses in fMRI, or experiments in which trials are long. Likewise, in other neural recording modalities, like EEG or ECoG, MEG, or LFP, detrending will become necessary, as background or low-frequency non-stationarity could be irrelevant to the intent of the study. We, however, do not assume any temporal structure underlies the individual series (e.g., additive autoregressive noise processes), beyond the piecewise linear dependence, nor do we implement detrending schemes as complex as those considered by similar examinations, such as [5]. This is purely because it would be very difficult not to overfit, for example, a discrete cosine basis detrending, on such short time series.

We use the R package *fda* to perform the functional principal components analysis (functions are available online via [24]), as detailed in [20]. Calculating fPCs can be performed in a variety of ways, but the most intuitive is to interpolate on a much finer grid of time points than is actually sampled from, often through the use of some smooth basis, and then use standard PCA routines to proceed on the augmented data (where again, timepoints function as what would normally be the covariates). The smoothness of our function estimate (as represented by the integrated square of its second derivative) will be penalized in combination with the reconstruction error, in a proportion chosen by generalized cross-validation. Therefore, our assumption of smoothness over time in the original space will translate to smoothness over time in the bases.

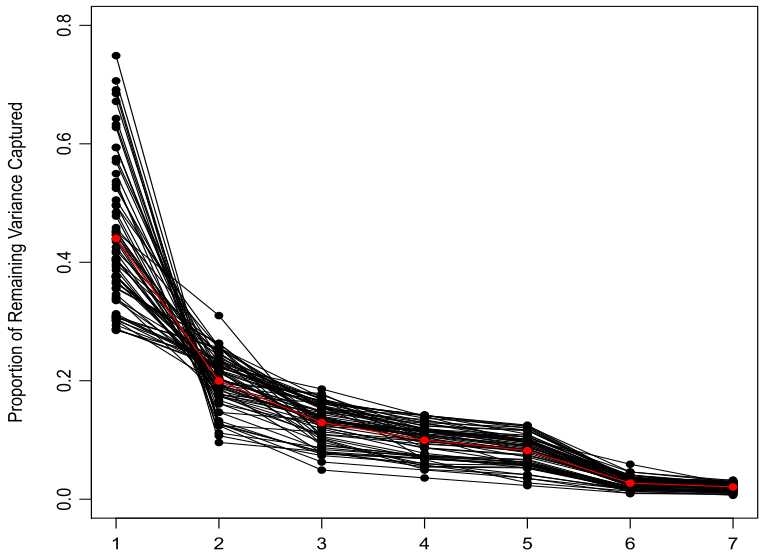

Figure 3. The variability captured by the next seven functional principal components (after the mean activation component) for all of one subject's trials. The red line is the average over trials.

In our analyses on fMRI data, we discard the first fPC, as it is almost exactly constant over time, capturing the widely varying mean BOLD signal across voxels. We then consider the non-trivial fPCs, that is, from the second component onward, evaluated at the time points of imaging. Figure 2 shows an example of such a projection, 14 trials in which the same subject correctly solved a *regular* problem. The trials in Figure 2 are the same as those in Figure 1. In both cases, we include four scans (8 s) before the problem-solving period, consisting of approximately 5 s of repetition detection followed by 3 s of fixation, as well as four scans after the problem-solving period, which encompass about 5 s of feedback and 3 s of repetition detection. The orthogonality constraint and near linearity of the first component result in fPCs resembling (approximately) polynomials of subsequently increasing order. In Figure 3, we see that the first non-trivial principal component (on average) captures about half of the variability remaining after accounting for the mean component. This component $X_2(t)$ in the projection of trial i will be the $X^{(i)}(t)$ upon which we assumed to operate in Section 3.

Returning to Figure 2, we see that modes of maximal variability of equal order are heterogeneous. Some exhibit very similar structure, up to what might be thought of as a time dilation (a linear scaling of time to stretch or compress the same effects to a different trial length). However, a constant location or number of change-points across all trials does not appear reasonable in a one-dimensional projection. We therefore use a model that clusters trials into groups that exhibit similar structure, that is, the same number of change-points, as well as similar temporal progressions given the number of change-points in that group. Trials that repeat the same problem, for example, are not required to be modeled identically and can therefore exhibit differing cluster membership.

5. Simulation study

To assess ground truth performance, we simulated a variety of datasets that were similar to projections of observed fMRI trajectories. We will display the most in-depth example, an instance of performance under an extreme misspecification of the piecewise linear modeling process (1). Three more studies exhibiting much better performance on more amenable simulated datasets can be found in Appendix D. Two of which (D.1 and D.2) were generated according to the piecewise linear model (1) and then fitted by the same model, while the other (D.3) is an example of data generated via the higher-dimensional piecewise constant model (2), specifically in five dimensions, and the corresponding model fit. The following simulation incorporates more features of real data, and of the full data analytic process. We take care to include features like varying trial lengths, and small numbers of aberrant trials in their number of states (relative to the majority of trials) prior to projection to high dimensions, where we enforce voxel-specific hemodynamic response functions, spatial correlations, and disparate scalings before using fPCA to extract trajectories of the form that are used in real data analysis.

5.1. Simulation 1

Seventy-five trials were simulated, each with probabilities of being generated by a zero, one, two, or three change-point models of 0.05, 0.1, 0.75, and 0.1, respectively. Trial lengths approximated the distribution later seen in the real fMRI data: uniform between 14 and 24, inclusive, and change-points were placed uniformly and rounded to the nearest half-integer (rejecting change-points directly adjacent to, or coincident with, trial end-points or other change-points). Within each trial, we took $\gamma^{(i)}$ and $\beta_1^{(i)}$ to be normally distributed (and $\left\{\beta_j^{(i)}\right\}_{j=2}^{k}$ specified conditionally upon the previous state parameter, β_{j-1}^{i}, otherwise identical to $\beta_1^{(i)}$, but such that all mass in $\left[\beta_{j-1}^{i}-0.3, \beta_{j-1}^{i}+0.3\right]$ is removed), with means zero and $\vec{\beta}^{(\mathcal{C}_h)}$, and standard deviations $\phi = 0.6$ and $\psi = 0.2$, respectively. IID Gaussian noise is added ($\sigma = 0.02$) before these series are normalized to have L^2 norms of 1. We then generate 100-dimensional series, more akin to the dimensionality encountered in megavoxel-based whole-brain analyses, with the given trial-wise mean process subjected to a random (Gaussian, mean 0 s, standard deviation 0.5 s) temporal shift for each dimension (interpolation is performed by simple two internal knot regression spline fits), and the multivariate normal correlation structure imposed by $\forall i, t, \sigma(X_i(t)) = 0.075$, and for $1 \leqslant i \leqslant 100, i-2 \leqslant j \leqslant i+2, i \neq j, \sigma(X_i(t), X_j(t)) = 0.015$, across the dimensions. This is analogous to a set of voxels lying on a line, and the structurally explicit spatial correlations only extending to the four other nearest voxels. Each dimension is then scaled via $AX_i + B$, where $A \sim \mathrm{N}(0, 0.3^2)$ and $B \sim \mathrm{Unif}(-50, 50)$, before finally being subjected to the fPCA preprocessing described in Section 4 to extract components, to which the piecewise linear model is fit. As detailed in 4, each trial is reduced independently and performed so agnostically to any information about its generation. This setup implies model misspecification on many levels, as well as serving to evaluate performance on fPC outputs, which are, by definition, smooth. The preprocessing method should also demonstrate robustness to spatial correlations, albeit those implemented here with subtlety only befitting a toy example. The most striking form of misspecification is in the nature of the clusters, which do not exist. Trials could be considered bound to one of four clusters, sharing only the number of change-points or each trial occupying its own cluster, although an obvious identifiability issue would arise if inferences on cluster membership were of interest to us. The modeling process will also inevitably be subjected to variable hemodynamic responses throughout the brain, also approximated here.

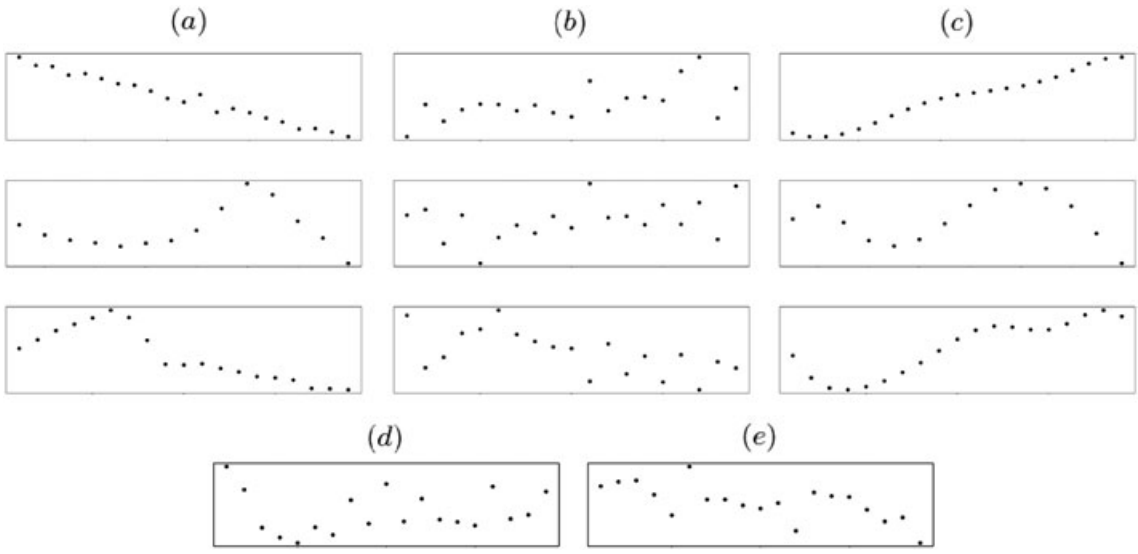

Figure 4. Simulated data at various stages. (a) Three trials of simulated mean functions, with zero, three, and two change-points, respectively. (b) Three of the one hundred projected dimensions corresponding to the third mean function in (a). Very little original structure appears to be maintained. (c) The fPCA reconstructed functions corresponding to the three functions in (a). Noting that these components are only unique up to a negation makes their similarity apparent. (d) The sample mean function across (all one hundred) dimensions in (b). The result is a function very dissimilar from the original. (e) The sample variance function across (all one hundred) dimensions in (b). Again, the function of interest is not reflected well in these simple summary statistics.

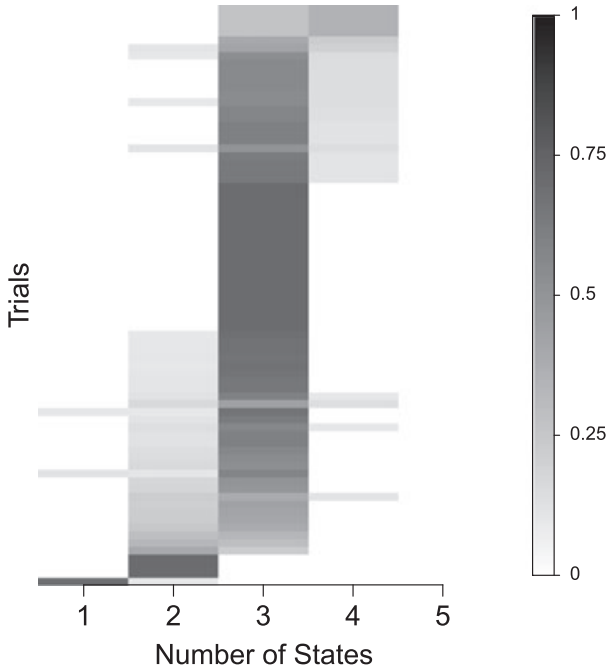

Figure 5. Posterior probability of the number of states by trial, for all 75 trials (one more than the number of change-points), sorted by posterior mean in the number of states.

5.2. Results

Naturally, the original mean functions are only identifiable, at best, up to a linear transformation. It would therefore be unwise to draw scientific inferences from parameter estimates other than the number and location of change-points. Figure 4 shows the estimated second fPCs down-sampled to the original data frequency (panel (c)) along with their corresponding original mean functions (panel (a)). Also shown are exemplar high-dimensional projections (panel (b)), corresponding to voxel-space data, from which very little about the underlying structure can be detected by eye. Very simple summary statistics, for example, the sample mean (panel (d)) and variance (panel (e)) across the full set of dimensions shown (in part) in panel (b), would not serve reconstruction purposes well. Figure 5 shows reasonable performance in the estimated posterior densities of the numbers of change-points by trial, where we find that $\approx 76\%$ of trials have posterior probability of greater than 0.75 of their being generated via exactly 2 change-points. While the smaller number of 'distractor' trials can technically influence the posterior densities for quantities of interest in the majority of three-state trials, the effect is minimal. Another statistic of fundamental interest is the estimated marginal (over trial) distribution of states, which takes mass $(0.02, 0.11, 0.81, 0.06)$ at zero, one, two, and three change-points, respectively.

5.3. Conclusions

With mixtures of trials generated from models of varying numbers of states, we find we can reliably recover clusterings of the functional forms used in simulation, as well as the nature of the effects governing the majority of trials, corresponding to the *regular* trials exhibiting usual behavior, under realistic noise conditions. Even with gross model misspecification, stronger covariability (homogeneity) within groups of trial-level parameters allows us to maintain the quality of change-point location estimates under increasing top-level additive noise conditions; in simulation D.2, the addition of trials with low signal-to-noise ratio (SNR) essentially inflates the variability in the estimates within the remainder of trials, although this is dependent on how similar the distractors are to the 'real' trials relative to how similar the informative trials are to each other (as with any discriminative task).

Extracting the greater information in multivariate series, as in simulation D.3, comes at the expense of computational time scaling approximately linearly with the number of dimensions. Even the correlated measurements of fMRI data would reinforce the ability to detect states, and without requiring assumptions on how the correlation structure might change from state to state, but the benefits would be increasingly marginal. Fitting model (2) to noisy data, such as that of D.3, required much greater than an order of mag-

nitude increase in iterations of Monte Carlo, for just five dimensions. Therefore, we have used piecewise linear models in the fMRI data analysis.

6. Data analysis

We analyzed the imaging data by first applying fPCA, as described in Section 4, and then fitting model (1) to the resulting univariate time series.

6.1. Whole-brain results

Figure 6 shows posterior probabilities on the number of states for *regular* and *exception* problems correctly solved by one subject. Models with three states (two change-points) are preferred for the majority

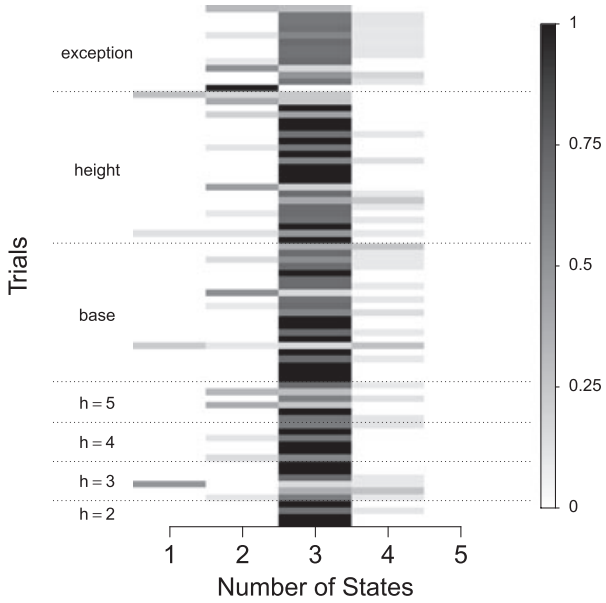

Figure 6. Posterior probability, with magnitude shown on the legend, of the number of states by trial, for all correctly solved problems by one subject. The three-state fit is most probable for the majority of trials.

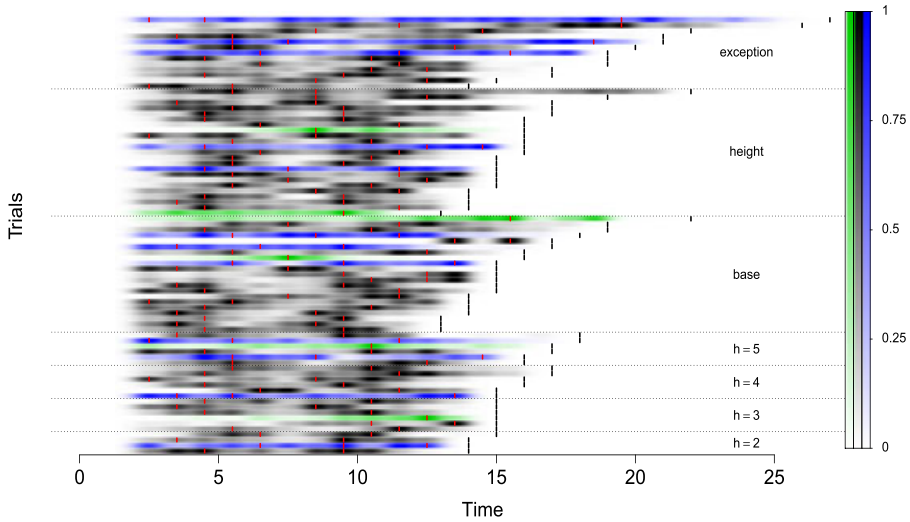

Figure 7. Posterior densities of change-point locations given $\widehat{c}_i$ and $\widehat{n_{c_i}}$, for all data from the subject in *Figure 6*. Red ticks are now MAP estimates of change-point locations. Trials for which two-state, three-state, and four-state models are posteriorly modal have densities in green, black and blue, respectively. Time is in scans.

of trials, although rarely with high probability. After conditioning on *regular* problem height (only problems with heights 2–5 were posed among the value-type *regular* problems), one might expect a shift in probability mass upward with the height of the problem being solved, but we did not observe this. Figure 7 shows the estimates of change-point locations by trial for the same set of problems shown in Figure 6. In both figures, trials are partitioned according to the type of problem being solved, where in the case of *value* problems, the height of problems for which value is being solved is specified. The densities of change-point locations of trials for which a two-state, three-state, and four-state solution was estimated as maximal *a posteriori* are shown in green, black, and blue, respectively, with the corresponding MAP estimates of the change-points shown in red.

The resulting state durations are similar to those found in [8]. The study inferred state durations that were consistent with change-points approximately coincident with the beginning and end of the problem-solving period. Our estimated states regularly segment the fixation and response/feedback phases in the problem (recall these are of a fixed, four-scan durations), as shown in Figure 8. The problem-solving phase itself varies linearly with trial length and, as such, is the most variable in duration. The trials with two, three, and four-state MAP estimates are shown in the same colors as in Figure 7. The lines represent the aggregated densities of the first change-point location for the set of trials with the given estimated number of states on the left, and the duration until end of trial from the last change-point for the same

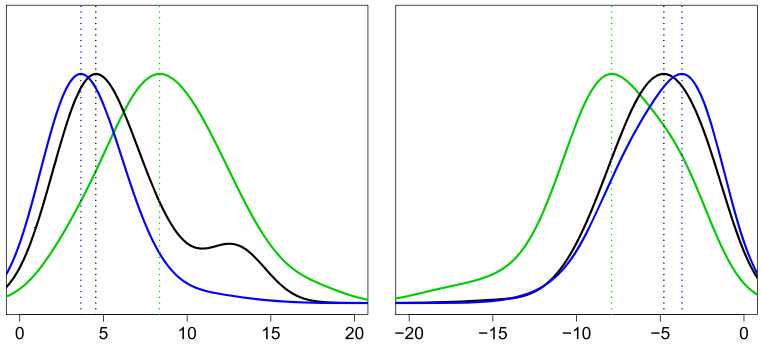

Figure 8. Density estimates of the locations of the first change-point (unaligned) and last change-point (relative to end of trial) for all data from the subject in Figures 6 and 7. Trials for which two-state, three-state, and four-state models are posteriorly modal have densities in green, black and blue, respectively. Lines correspond to posterior modes. Time is in scans.

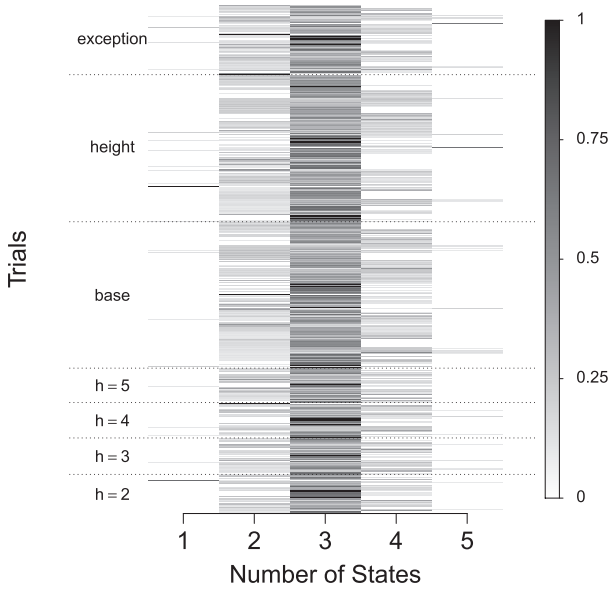

Figure 9. Posterior probability, with magnitude shown on the legend, of the number of states by trial, for all correctly solved problems by all 20 subjects. The three-state fit remains most probable for the majority of trials, across subjects and problem types.

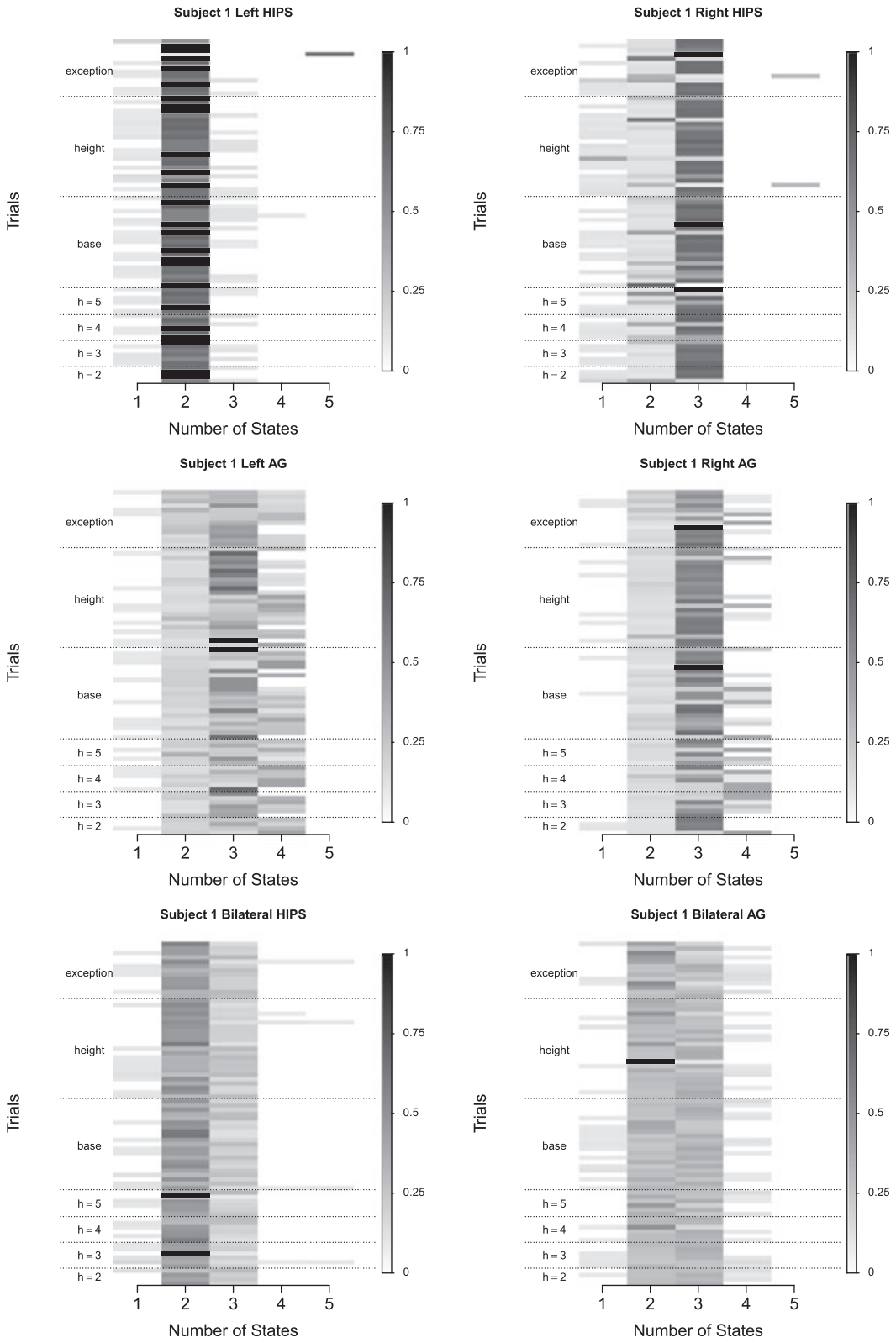

Figure 10. Likely regions of interest. Bilateral reductions may encompass two fairly similar areas, such as for angular gyrus, or blur two areas encoding different information across hemispheres, such as horizontal intraparietal sulcus.

sets of trials on the right. Models with more than two states generally pick out approximately where the problem-solving period itself starts and ends, whereas two-state models are posteriorly modal at a point less cognitively well-defined, toward the middle of the trial. Because the functional preprocessing tends to smooth over the neighborhood around any moderate deflection of the problem-solving phase, there may be slight overestimation of this phase (underestimation of the fixed length end phases). However, the problem-solving period in *exception* problems though is much more poorly delineated, which indicates that during the time the problem is presented, there are more scans in which the subject is performing some cognitive task other that which would be represented by the problem-solving period as segmented out of a *regular* problem.

In Figure 9, we see the posterior estimates of the number of change-points for all trials over all 20 subjects, where a three-state solution (2 change-points) is often preferred in *regular* problems albeit with less consistent results in *exception* problems. In fact, in approximately 45% of trials, we have a posterior probability of at least 0.75 of exactly three discernable states, given this temporal resolution. For approximately 74% of trials, again across all subjects, there is greater than or equal to 0.75 posterior probability of at least three segmentable states. As 0.75 probability corresponds to 3:1 posterior odds in favor of the tested hypothesis, this may be considered non-negligible evidence [25]. Only one of the 20 subjects had a posterior distribution on MAP number of states for *exception* trials with mean significantly greater than that of *regular*, when corrected for multiple comparisons under level $\alpha = \frac{0.05}{20}$ (permutation test; $p \approx 10^{-4}$). Note that while height is strongly correlated with trial length (Pearson's $\rho = 0.53,\ p \approx 10^{-16}$, on 429 *df*), neither height nor trial length is overwhelmingly covariable with posterior mean estimates of the number of states ($\rho = 0.04,\ -0.12,\ p \approx 0.39,\ 0.01$, respectively, on 429 *df*). This is evidence against both a consistent effect across *value* problems of different heights and artificial change-points due merely to longer trials. Additional manipulations due to greater problem height are manifested only through the problem-solving state duration (also similarly to [8]).

6.2. *Region of interest results*

All whole-brain analyses point to a majority of trials being segmentable up to the preparatory or fixating, problem solving, and feedback phases, using only a one-dimensional projection. However, the more complicated temporal progressions (those that exhibit less canonical activation) are not shared by many trials. They will therefore not be effectively estimated in this small subset. Within specific regions of interest however, using the same dimensionality reduction process as when addressing whole-brain activity, it is often the case that different phases are segmentable than those dominating the whole brain representation. Within angular gyrus, for example, we see estimated posterior distributions on the number of change-points in Figure 10 with moments similar to those of the whole brain for left, right, and bilateral reductions. In the horizontal intra-parietal sulcus though, while the right hemisphere exhibits a maximally probable three-state segmentation for most trials, left horizontal intraparietal sulcus (HIPS) only shows probable two-state solutions, possibly indicating a greater disparity in the tasks lateralized via HIPS. The horizontal intra-parietal sulcus, an area usually associated with visual processing, has also been tied to numerical processing [14]. It elicits a much more consistent response across the majority of trials and reaffirms the generally stronger (larger magnitude) responses often seen at the ACT-R-implicated areas of the left hemisphere, as has been observed previously in [7]. Interestingly, the lack of specifically lateralized responses in angular gyrus also reaffirms the results of [14].

7. Discussion

Replications across many related trials offer the substantial information needed for reliable change-point detection from short, noisy neuroimaging time series. However, even when the task remains fixed across trials, a subject's behavior and neural activity will change. Thus, brain activity across trials can be considered similar, but not the same. We therefore began this investigation with the assumption that the number of change-points (corresponding to brain states) would remain constant across trials, while their locations in time would vary across trials. We discovered that it was too restrictive to assume the number of change-points remained constant across all the trials and instead incorporated into the model the assumption that trials formed clusters sharing a given number of change-points. This complicated the implementation but made it much more flexible.

The models we developed provide probabilistic results on the number of states. Within single voxels, the piecewise constant activity model is plausible, but for analyses involving regions of interest or whole

brain, the piecewise linear model is preferable. We developed a reversible-jump sampler that correctly identified change-points in simulations, where the clustered trial structure was enforced. We then re-analyzed the arithmetic pyramid problem-solving data [8] and were able to provide substantial statistical support for conclusions drawn previously by other methods. By analyzing both one-dimensional projections of fMRI trajectories and trajectories within individual regions of interest, we were able to show that the data provide evidence for several states, corresponding to pre-problem solving, problem solving, and post-problem solving, but that some brain regions may participate only in a subset of these states.

While we have here applied the methodology to fMRI data, the combination of low signal-to-noise ratio, relatively short trial lengths, and high-dimensionality make change-point detection especially difficult in this context. In fact, the methods we developed could be applied to any data consisting of repeated multivariate time series, including those produced by magnetoencephalography, electroencephalography, electrocorticography, or multi-electrode recordings of local field potentials. The far superior time resolution of these modalities should provide much more information about change-point locations.

8. Software

Rj-MCMC-based algorithms for the implementation of posterior sampling in both the piecewise linear and piecewise multivariate constant models are available publicly on Bitbucket, along with all simulations and scripts to generate visualizations of posterior statistics, including (but not limited to) those displayed in the article. Implementations of (1), of the scale featured in Section 5, require on the order of 1 h per 20,000 samples, on a single 2.3 GHz core, but vary substantially, especially in proportion to the average estimated dimensionality. All code is written in R. Code can be accessed at: 'bitbucket.org/SKoerner1/change-point-detection'

Appendix A: Transitions in the number of states

Henceforth, θ' and θ'' will refer to the parameters of the states adjacent to the inserted change-point (before and after, respectively) of a proposed split, as all other state parameters remain unchanged (barring τ' itself, which is the only newly proposed change-point in the split). We use Θ_K to refer to a space at least encompassing that of both the current and proposed parameter spaces, Θ_k and Θ_{k+1}, respectively. In this case, the augmented parameter space Θ_K will simply be Θ_{k+1}. Analogously to the usual Metropolis–Hastings MCMC, we accept a proposal of $k = n_{\mathcal{C}_h} \to n_{\mathcal{C}_h} + 1 = k + 1$ where the set of trials currently generated by cluster h is denoted A_h, with probability $\max(1, \alpha)$, where

$$\alpha = \frac{\mathscr{L}\left(X; k+1, \vec{\theta}_{k+1}\right)}{\mathscr{L}\left(X; k, \vec{\theta}_k\right)} \cdot \frac{p\left(T_{k+1}^{(A_h)}, B_{k+1}^{(A_h)}, \Gamma_{k+1}^{(A_h)}, B_{k+1}^{(C_h)}, k+1\right)}{p\left(T_k^{(A_h)}, B_k^{(A_h)}, \Gamma_k^{(A_h)}, B_k^{(C_h)}, k\right)} \cdot \frac{p\left(\vec{u}_{\Theta_{k+1}\to\Theta_K}\right)}{p\left(\vec{u}_{\Theta_k\to\Theta_K}\right)} \cdot \frac{d_{k+1}}{b_k} \cdot |J|,$$

of which the second and third factors will be expanded as follows:

$$\begin{aligned}
&\frac{\mathscr{L}\left(X; k+1, \vec{\theta}_{k+1}\right)}{\mathscr{L}\left(X; k, \vec{\theta}_k\right)} \times \frac{p_{\text{Pois}(\lambda)}(k+1)}{p_{\text{Pois}(\lambda)}(k)} \times \frac{p_{\text{N}(0,\chi^2)}\left(\beta'^{(C_h)}\right) \cdot p_{\text{N}(0,\chi^2)}\left(\beta''^{(C_h)}\right)}{p_{\text{N}(0,\chi^2)}\left(\beta_j^{(C_h)}\right)} \\
&\quad \times \prod_{i \in A_h} \left(\frac{p_{\text{Dir}(\boldsymbol{\alpha})}\left(\frac{T'^{(i)}}{l^{(i)}}\right)}{p_{\text{Dir}(\boldsymbol{\alpha})}\left(\frac{T^{(i)}}{l^{(i)}}\right)} \right) \\
&\quad \times \prod_{i \in A_h} \left(\frac{p_{\text{N}\left(\beta'^{(C_h)}, \phi^2\right)}\left(\beta'^{(i)}\right) \cdot p_{\text{N}\left(\beta''^{(C_h)}, \phi^2\right)}\left(\beta''^{(i)}\right)}{p_{\text{N}\left(\beta_j^{(C_h)}, \phi^2\right)}\left(\beta_j^{(i)}\right)} \right) \\
&\quad \times \left(p_{\text{N}(0,\sigma_\theta^2)}\left(\arctan\left(\beta'^{(C_h)}\right) - \arctan\left(\beta_j^{(C_h)}\right)\right)\right)^{-1} \\
&\quad \times \prod_{i \in A_h} \left(p_{\text{N}\left(u_\theta, \sigma_{u_\theta}^2\right)}\left(\arctan\left(\beta'^{(i)}\right) - \arctan\left(\beta_j^{(i)}\right)\right)\right)^{-1} \\
&\quad \times \frac{d_{k+1}}{b_k} \times |J|,
\end{aligned} \tag{A.1}$$

where b_k and d_{k+1} represent the probabilities of birth and death of a new interval, respectively, and (omitting the densities at draws of change-point locations, which are uniform on the unit interval, and would thus be represented by a product of $\#(A_h)$ ones, prior to scaling; this scaling is then incorporated through the Jacobian factor) $|J|$ (as formulated in (B.1) within Appendix B) is the absolute value of the determinant of the Jacobian of the proposed transformation, which is defined by the deterministic parts of the sequence:

- Pick new change-points by drawing uniformly over the possible change-point locations in the chosen interval for each trial. Let this proposal for trial i be denoted $\tau'^{(i)}$. We enforce at least two images between potential subsequent change-points and/or the boundaries of the trial.
- Draw a normally distributed u_θ specifying an angle that defines the potential transformation of the underlying slopes (let $\beta_j^{(\mathcal{C}_h)} =: \tan(\theta^{(\mathcal{C}_h)})$, when splitting state j):

$$\beta'^{(\mathcal{C}_h)} = \tan\left(\theta^{(\mathcal{C}_h)} + u_\theta\right) \quad \text{and}$$

$$\beta''^{(\mathcal{C}_h)} = \frac{\beta_j^{(\mathcal{C}_h)} \sum_{i\in A_h}\left(\tau_j^{(i)} - \tau_{j-1}^{(i)}\right) - \beta'^{(\mathcal{C}_h)} \sum_{i\in A_h}\left(\tau'^{(i)} - \tau_{j-1}^{(i)}\right)}{\sum_{i\in A_h}\left(\tau_j^{(i)} - \tau'^{(i)}\right)}.$$

- Draw a set of normally distributed $u_{\theta^{(i)}}$, with mean u_θ to split the chosen $\beta_j^{(i)}$ $\left(=: \tan\left(\theta^{(i)}\right)\right)$ into the following:

$$\beta'^{(i)} = \tan\left(\theta^{(i)} + u_{\theta^{(i)}}\right) \quad \text{and}$$

$$\beta''^{(i)} = \frac{\beta_j^{(i)}\left(\tau_j^{(i)} - \tau_{j-1}^{(i)}\right) - \beta'^{(i)}\left(\tau'^{(i)} - \tau_{j-1}^{(i)}\right)}{\tau_j^{(i)} - \tau'^{(i)}}.$$

The mean angle of deviation, u_θ, is drawn from a $\mathrm{N}\left(0, \sigma_\theta^2\right)$. Similarly, $u_{\theta^{(i)}}$ is drawn from a $\mathrm{N}\left(u_\theta, \sigma_{\theta^{(i)}}^2\right)$ normal. This is a geometrically intuitive approach to constructing this type of jump, as visualized in Figure A.1. By automatically rejecting the proposal if $\left(\theta^{(\mathcal{C}_h)} + u_\theta\right) \notin \left(-\frac{\pi}{2} + \epsilon, \frac{\pi}{2} - \epsilon\right)$ or if any $\left(\theta^{(i)} + u_{\theta^{(i)}}\right) \notin \left(-\frac{\pi}{2} + \epsilon, \frac{\pi}{2} - \epsilon\right)$, we are essentially sampling from an appropriately truncated density to keep temporal derivatives of the projection from crossing a singularity in β. In all applications, the automatic rejection procedure is never employed, as tuning σ_θ and $\sigma_{\theta^{(i)}}$ is sufficient. This is, in part, because

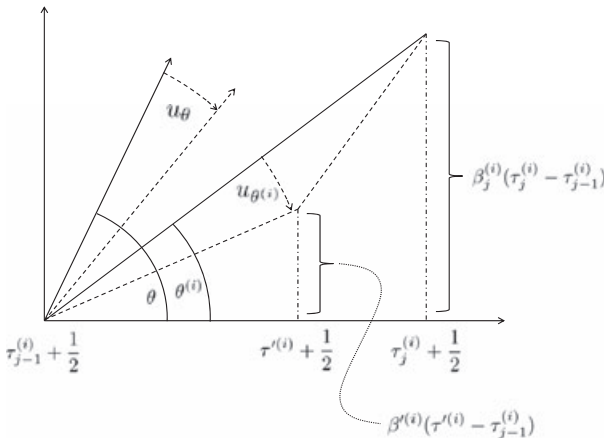

Figure A.1. Proposing a split of one state into two within a trial: u_θ, the perturbation to the original model-level slope (solid vector), is the new mean angle of deviation from trial-level slopes, to generate proposals. The current $\beta_j^{(\mathcal{C}_h)}$, the set of all trials' current and proposed change-points in the cluster, and the new draw, u_θ, determine both of the two proposed model-level slopes (the first given by the dashed vector), which have no geometric constraint, as there is no notion of continuity maintained by cluster-level parameters. The perturbation to the trial-level mean function, $u_{\theta^{(i)}}$, is distributed around u_θ and specifies the individual trial's deviant angle. Given the location of the proposed change-point within a trial, parameters of the second proposed state are then fully specified.

some form of regularization will be necessary prior to analyzing measurements on inherently different scales, which tends to bound absolutely the likely magnitude of β. In the case of merging, we invert the relevant ratios that would be generated by treating the current state as a hypothetical jump from the proposal, analogously to the split case. We would instead combine these parameters via the following:

$$\beta^* = \frac{1}{\#(A_h)} \sum_{i \in A_h} \frac{\beta_{j+1}^{(i)} \left(\tau_{j+1}^{(i)} - \tau_j^{(i)} \right) - \beta_j^{(i)} \left(\tau_j^{(i)} - \tau_{j-1}^{(i)} \right)}{\tau_{j+1}^{(i)} - \tau_{j-1}^{(i)}}$$

$$u_\theta^* = \arctan\left(\beta_j^{(\mathcal{C}_h)} \right) - \arctan(\beta^*)$$

$$\beta^{*(i)} = \frac{\beta_{j+1}^{(i)} \left(\tau_{j+1}^{(i)} - \tau_j^{(i)} \right) - \beta_j^{(i)} \left(\tau_j^{(i)} - \tau_{j-1}^{(i)} \right)}{\tau_{j+1}^{(i)} - \tau_{j-1}^{(i)}}$$

$$u_{\theta^{(i)}}^* = \arctan\left(\beta_j^{(i)} \right) - \arctan\left(\beta^{*(i)} \right).$$

Appendix B: Jacobian of the change to number of states

For the split/merge step in the piecewise linear case, as given by expression (A.1), the Jacobian J takes the form:

$$\begin{bmatrix}
\frac{\delta T}{\delta T} & \frac{\delta T}{\delta u_T} & \frac{\delta T}{\delta \beta} & \frac{\delta T}{\delta u_\theta} & \frac{\delta T}{\delta \gamma} & \frac{\delta T}{\delta u_\gamma} & \frac{\delta T}{\delta B} & \frac{\delta T}{\delta u_\Theta} \\
\frac{\delta T'}{\delta T} & \frac{\delta T'}{\delta u_T} & \frac{\delta T'}{\delta \beta} & \frac{\delta T'}{\delta u_\theta} & \frac{\delta T'}{\delta \gamma} & \frac{\delta T'}{\delta u_\gamma} & \frac{\delta T'}{\delta B} & \frac{\delta T'}{\delta u_\Theta} \\
\frac{\delta \beta}{\delta T} & \frac{\delta \beta}{\delta u_T} & \frac{\delta \beta}{\delta \beta} & \frac{\delta \beta}{\delta u_\theta} & \frac{\delta \beta}{\delta \gamma} & \frac{\delta \beta}{\delta u_\gamma} & \frac{\delta \beta}{\delta B} & \frac{\delta \beta}{\delta u_\Theta} \\
\frac{\delta \beta'}{\delta T} & \frac{\delta \beta'}{\delta u_T} & \frac{\delta \beta'}{\delta \beta} & \frac{\delta \beta'}{\delta u_\theta} & \frac{\delta \beta'}{\delta \gamma} & \frac{\delta \beta'}{\delta u_\gamma} & \frac{\delta \beta'}{\delta B} & \frac{\delta \beta'}{\delta u_\Theta} \\
\frac{\delta \gamma}{\delta T} & \frac{\delta \gamma}{\delta u_T} & \frac{\delta \gamma}{\delta \beta} & \frac{\delta \gamma}{\delta u_\theta} & \frac{\delta \gamma}{\delta \gamma} & \frac{\delta \gamma}{\delta u_\gamma} & \frac{\delta \gamma}{\delta B} & \frac{\delta \gamma}{\delta u_\Theta} \\
\frac{\delta \gamma'}{\delta T} & \frac{\delta \gamma'}{\delta u_T} & \frac{\delta \gamma'}{\delta \beta} & \frac{\delta \gamma'}{\delta u_\theta} & \frac{\delta \gamma'}{\delta \gamma} & \frac{\delta \gamma'}{\delta u_\gamma} & \frac{\delta \gamma'}{\delta B} & \frac{\delta \gamma'}{\delta u_\Theta} \\
\frac{\delta B}{\delta T} & \frac{\delta B}{\delta u_T} & \frac{\delta B}{\delta \beta} & \frac{\delta B}{\delta u_\theta} & \frac{\delta B}{\delta \gamma} & \frac{\delta B}{\delta u_\gamma} & \frac{\delta B}{\delta B} & \frac{\delta B}{\delta u_\Theta} \\
\frac{\delta B'}{\delta T} & \frac{\delta B'}{\delta u_T} & \frac{\delta B'}{\delta \beta} & \frac{\delta B'}{\delta u_\theta} & \frac{\delta B'}{\delta \gamma} & \frac{\delta B'}{\delta u_\gamma} & \frac{\delta B'}{\delta B} & \frac{\delta B'}{\delta u_\Theta}
\end{bmatrix},$$

which then reduces to a matrix of the following form, where $-$ indicates the entry does not exist (γ is unaffected) and X indicates a non-zero set of entries, which is not necessarily square.

$$\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & - & 0 & 0 \\
X & X & 0 & 0 & 0 & - & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & - & 0 & 0 \\
X & X & X & X & 0 & - & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & - & 0 & 0 \\
- & - & - & - & - & - & - & - \\
0 & 0 & 0 & 0 & 0 & - & 1 & 0 \\
X & X & X & X & 0 & - & X & X
\end{bmatrix}.$$

However, each sub-matrix on the diagonal of the form:

$$\begin{bmatrix} 1 & 0 \\ X & X \end{bmatrix}$$

is square, hence the simple expression for the full determinant:

$$\left| \frac{\delta T'}{\delta u_T} \right| \cdot \left| \frac{\delta \beta'}{\delta u_\theta} \frac{\delta \beta'}{\delta \beta} \right| \cdot \left| \frac{\delta B'}{\delta u_\Theta} \frac{\delta B'}{\delta B} \right|,$$

where each multiplicand is the determinant of a square matrix. This simplifies to the following:

$$\prod_{i\in A_h}\left(\tau_{j+1}^{(i)}-\tau_j^{(i)}-3\right)\times\prod_{i\in A_h}\frac{\tau_j^{(i)}-\tau_{j-1}^{(i)}}{t_j^{(i)}-\tau'^{(i)}}\sec^2\left(\arctan\left(\beta_j^{(i)}\right)+u_{\theta^{(i)}}\right)\times\frac{\sum_{i\in A_h}\left(\tau_j^{(i)}-\tau_{j-1}^{(i)}\right)}{\sum_{i\in A_h}\left(t_j^{(i)}-\tau'^{(i)}\right)}\sec^2\left(\arctan\left(\beta_j^{(\mathcal{C}_h)}\right)+u_\theta\right). \tag{B.1}$$

Note that we can observe from the positivity of $|J|$ that our transformation will be identifiably reversible from the inverse function theorem, under all possible draws not rejected given geometric infeasibility.

Appendix C: Cluster membership transitions

When fitting model either model (1) or (2), in order to propose a jump of the type $c_i \to c_i'$, we proceed similarly to the jump with acceptance probability detailed in (A.1), but in an iterated manner. A trial jumping from model to model is performed by proposing a sequence of split/combine moves on the individual trial, with the important distinction that original underlying $\beta^{(\mathcal{C}_h)}$ is not retained throughout the entire proposal nor are the transdimensional jumps targeted to $\beta^{(\mathcal{C}_h')}$. In the case of proposing a jump to a model of higher dimension, we define $g := n_{c_i'} - n_{c_i}$ and accept with probability $\min(1,\alpha)$, where

$$\alpha = p\left(c_i \to c_i'\right) = p\left(\beta^{\left(c_i^{(g)}\right)} \to \beta^{(c_i')} \mid c_i^{(g-1)} \to c_i^{(g)}\right)\times p\left(c_i^{(g-1)} \to c_i^{(g)} \mid c_i^{(g-2)} \to c_i^{(g-1)}\right)\times p\left(c_i^{(g-2)} \to c_i^{(g-1)} \mid c_i^{(g-3)} \to c_i^{(g-2)}\right)\times\cdots\times p\left(c_i^{(0)} \to c_i^{(1)} \mid c_i \to c_i^{(0)}\right)\times p\left(\beta^{(c_i)} \to \beta^{\left(c_i^{(0)}\right)}\right).$$

Here, the first proposal (last term) lets $\beta^{(c_i)}$ assume the values of $\beta^{(c_i^{(0)})} := \beta^{(i)}$, while the last proposal (the first term) is that the model-level slopes jump to those of the target model's $\beta^{(c_i')}$ (from $\beta^{(c_i^{(g)})}$). Each transdimensional jump is proposed as in the model split example, with $u_\theta^{(i)} = 0$. The $c_i^{(k)}$ represent temporary $n_{c_i}+k$ state intermediary models. The joint proposal is either accepted or rejected in full, and this simplifies ensuring reversibility, as none of these intermediaries can remain non-empty after a proposal, so the birth and death probabilities of these empty models, as are commonly made explicit in related treatments employing reversible-jump schemes, are required to cancel. Noting again that the intermediaries in propositions of these types are completely agnostic to the target model, we expect these jumps to have relatively low acceptance, although for pre-proposal sub-models with only one more or fewer change-point than the proposition, we find that aggregated acceptance rates are about 5% for our simulation, up to $\sim 15-25\%$ acceptance on recorded sets of fMRI trials, as often many iterations are spent with mostly similar sub-model dimensionalities, thus more relatively likely proposed sub-model parameter sets. The constraint to probable model dimensionalities via the trial lengths makes even the maximum number of intermediaries necessarily low (four at most), which keeps this sampling regime plausible computationally. For applications with a larger viable range of the number of change-points, leaving g unconstrained will quickly result in computational infeasibility. We reduce the number of sub-jumps proposed within a proposed cluster to cluster move by allowing a larger C and simply bounding g. For

$$n_{c_i'} - n_{c_i} = g = 1,$$

$$\mathbb{P}\left(c_i \to c_i'\right) = \mathbb{P}\left(\vec{\beta}^{(c_i)} \to \vec{\beta}^{\left(c_i^{(0)}\right)} = \vec{\beta}^{(i)}\right)\times\mathbb{P}\left(c_i^{(0)} \to c_i^{(1)} \mid \vec{\beta}^{(c_i)} \to \vec{\beta}^{\left(c_i^{(0)}\right)}\right)\times\mathbb{P}\left(\vec{\beta}^{\left(c_i^{(1)}\right)} \to \vec{\beta}^{(c_i')} \mid c_i^{(0)} \to c_i^{(1)}\right).$$

If deterministically specifying the intermediary cluster-level parameters for a proposal seems counterintuitive to the usual Metropolis architecture, recall steps of this type are constructed such that the stationary distribution is on the support of $\mathcal{C}_1, \ldots, \mathcal{C}_C$, for each trial (not, e.g., $\vec{\beta}^{(\mathcal{C}_h)}$). Our symmetric proposal distribution is as follows, in theory (that is, without constraining g):

$$c_i' \sim \text{DiscUnif}(\{\mathcal{C}_j\}_j \backslash c_i).$$

Appendix D: Additional simulation studies

To assess ground truth performance, we simulated a variety of datasets sharing similarities with projections of observed fMRI trajectories. We will display three examples: two of the piecewise linear modeling process (1) and one of the higher-dimensional piecewise constant model (2).

D.1. Simulation 1

Fifty trials were simulated, each with 0.5 probability of being generated by a one or two change-point sub-models, with normally distributed $\vec{\beta}^{(\mathcal{C}_h)}$ values. Within each trial, we took $\gamma^{(i)}$ and $\vec{\beta}^{(i)}$ to be normally distributed, with means zero and $\vec{\beta}^{(\mathcal{C}_h)}$, and variances $\phi = 0.5$ and $\psi = 0.5$, respectively. Change-points approximately followed a Dirichlet distribution law, with $\boldsymbol{\alpha} = \vec{1}$, up to the constraint that more than one scan must have separated them from another, or a trial start/end point. One change-point (randomly chosen) was moved an additional time point away when they occurred between adjacent scans. Before fitting, we standardized all time series to have mean 0 and sample variance 1. This creates model misspecification on the simulated data, because the trial-level slopes are no longer normally distributed nor centered at the original cluster-level parameters. The original parameters cannot easily be recovered, although nor can the scale of original BOLD data from the regularized basis functions.

D.2. Simulation 2

We constructed a simulation identically to D.1, but with four clusters of only 15 trials, generated by 0, 1, 2, and 2 change-points, respectively, fit simultaneously with 15 large magnitude white noise trials (not shown in figures of parameter estimates). In this case, the noise trials tested how strongly exclusive the signal clusters tended to be, and, by extension, how invariant the parameter estimates for trials of these clusters were to the distractors.

D.3. Simulation 3

A simulation was also constructed for employing and testing model (2). It was similar to D.1, except that five-dimensional multivariate normally distributed piecewise constant $\vec{\gamma}^{(i)}$ around $\vec{\gamma}^{(\mathcal{C}_h)}$ (also multivariate normal) were sampled, dimensions at all levels were independent (analogously to the univariate cases, $\phi = \psi = 0.5 \times \mathbf{I}_5$), and there were four entirely nested sub-models in the experiment, with parameters for additional states being appended to the beginning or end of the parameter set belonging to the model with one fewer state. We therefore had 15 trials of each 0, 1, 2, and 3 change-points.

D.4. Results

Figure D.1 is a visualization of the fitted estimates of (1) to D.1, indicated by solid lines, on four of the simulated time series having truly piecewise linear mean functions, with the locations of the true change-points indicated by dashed lines. Shrinkage toward sub-model-level parameters is evident in the first state of trials A and C, compared with the better-fit example trials B and D from the same true clusters, respectively. Otherwise, fits are as one would expect, with minimal misestimation of change-point locations. Figure D.2 shows posterior probabilities on the number of change-points (one less than the number of states) for all trials ordered such that true two change-point generated trials are above the dashed line. No probability mass exists outside the two-state or three-state solutions, and the four misclassified trials are the result of change-points occurring near the start or end of the series.

Figure D.3 shows the posterior densities of change-point locations given the estimates of $\widehat{c_i}$ and $\widehat{n_{c_i}}$ for all trials in D.1, with the true change-point locations shown as teal lines superimposed. Each strip represents one of the fifty trials, with the bottom 27 being generated by truly two-state sub-models; the top 23 have three states. Trials are ordered within cluster by the true mean change-point location and

normalized to show change-points and their estimates as proportions of trial length. Trials for which the number of change-points is misestimated have posterior densities shown in red.

Even in more complex simulations (not shown), with strong heteroskedasticity in top-level variability between clusters, we find that distractor trials (the noise trials similar in magnitude to the signal trials) have little influence on estimates, so long as clusters of trials with signal maintain their coherence. Figure D.4 displays estimates of change-point locations for the data simulated in D.2. Estimates of number of change-points for truly one-state trials were most noticeably affected by the noise trials. Four of

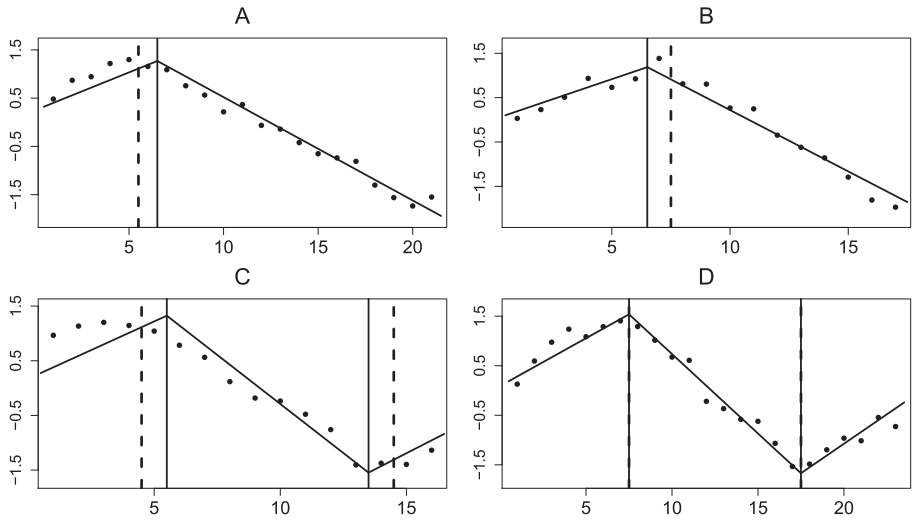

Figure D.1. Four trials of the simulated piecewise linear data with fitted parameters overlaid as solid lines; the true change-point locations are indicated by dashed lines. Trial A shows the influence of the cluster-level parameters, which behave more similarly to the trial-level parameters of trial B. The same effect is apparent in the three-state cluster containing trials C and D. Estimates of change-point locations are quite accurate; when the number of states is correctly estimated, they have a mean absolute error of approximately 0.42.

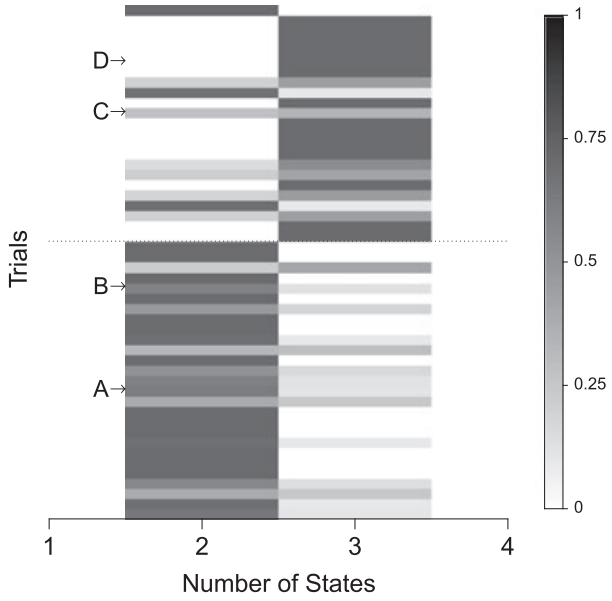

Figure D.2. Posterior probability of the number of states by trial, for all fifty trials (one more than the number of change-points). Trials truly generated from two states are below the dashed line; above it are three-state trials. The four trials shown in Figure D.1 are indicated. Misclassifications come from cases where a true change-point occurs toward the boundary of a trial, and top-level additive noise happens to continue the trend of the adjacent interval.

these trials were misclassified, as well as two truly two-state trials, for a total misclassification rate of 10%. Here, correctly segmented one-state trials are indicated by a dashed line through the duration.

In simulation D.3, we find that change-point estimates fitted from model (2), as shown in Figure D.5, are much better than the piecewise linear case with similar per-covariate variability, even under the generally difficult case of similar cluster-level parameters (in this case, clusters with larger number of states were generated to have sequential interior states with identical parameters to clusters with smaller numbers of states.) In this simulation, 6.7% of trials were misclassified (four out of 60), again with the least prototypical trials of the clusters suffering, as well as those with change-points very near trial end-points. Of course, methods of fitting are always tuned without knowledge of the generating parameters.

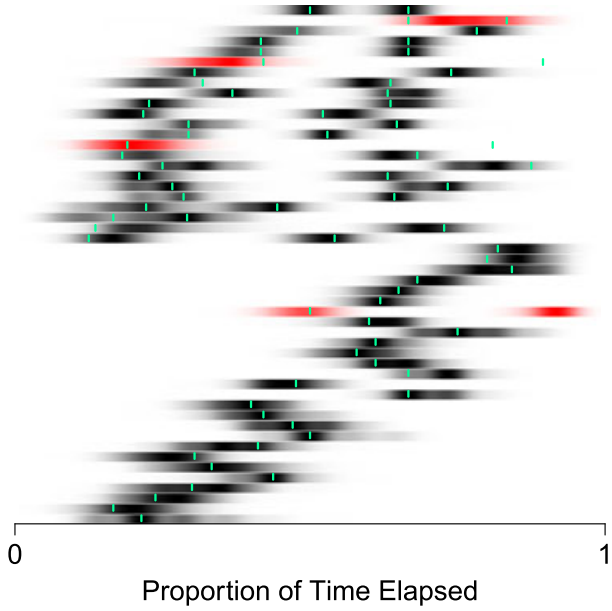

Figure D.3. Posterior densities of change-point locations given $\widehat{c_i}$ and $\widehat{n_{c_i}}$. Trials are ordered such that the first (bottom) 27 truly have two states, while the next (top) 23 are generated from three-state prototypes. The true change-point locations are marked in teal; trials with incorrectly estimated numbers of states have densities shown in red.

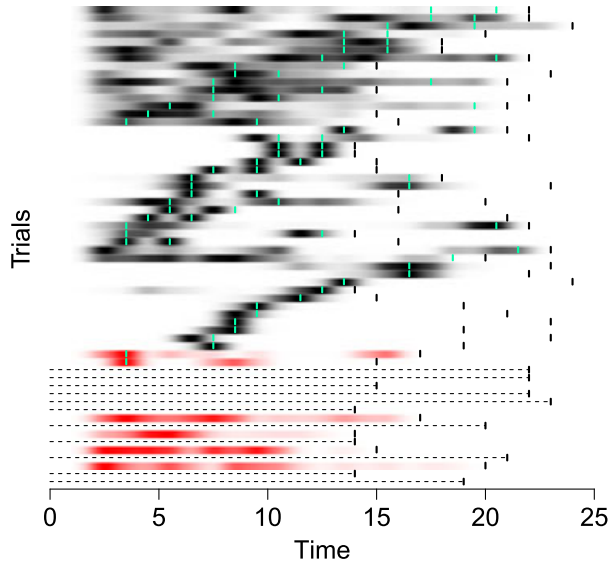

Figure D.4. Posterior densities of change-point locations given $\widehat{c_i}$ and $\widehat{n_{c_i}}$. Trials are ordered such that all within each block of 15 have the same true number of change-points: *0*, *1*, *2*, and *2*, by block. The true change-point locations are marked in teal; trials with incorrectly estimated numbers of states have densities shown in red.

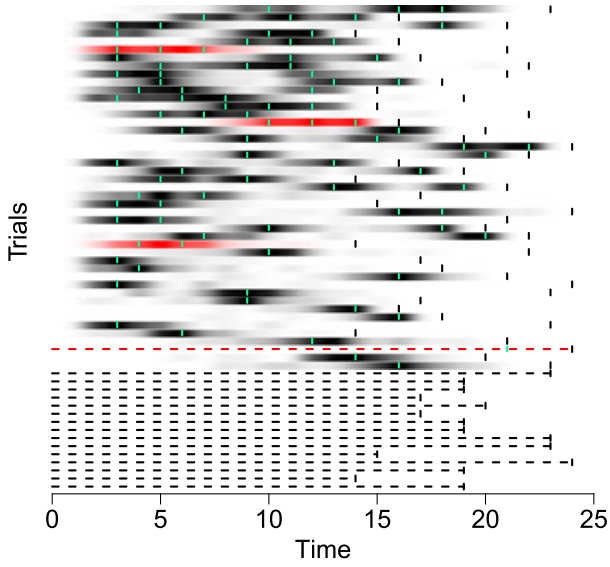

Figure D.5. Posterior densities of change-point locations in the five-dimensional regime, given $\widehat{c_i}$ and $\widehat{n_{c_i}}$. Trials are ordered such that each block of 15 have the same true number of change-points: *0*, *1*, *2*, and *3*, by block. The true change-point locations are marked in teal; trials with incorrectly estimated numbers of states have densities shown in red.

References

1. Matthews PM, Honey GD, Bullmore ET. Applications of fMRI in translational medicine and clinical practice. *Nature Reviews Neuroscience* 2006; **7**:732–744.
2. Basseville M, Nikiforov V. *Detection of Abrupt Changes: Theory and Application*. Prentice-Hall: Englewood Cliffs, NJ, 1993.
3. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 2004; **5**(4):557–572.
4. Aston JAD, Kirch C. Evaluating stationarity via change-point alternatives with applications to FMRI data. *The Annals of Applied Statistics* 2012; **6**(4):1906–1948.
5. Nam CFH, Aston JAD, Johansen AM. Quantifying the uncertainty in change points. *Journal of Time Series Analysis* 2012; **33**:807–823.
6. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**(4):711–732.
7. Anderson JR, Betts S, Ferris JL, Fincham JM. Cognitive and metacognitive activity in mathematical problem solving: prefrontal and parietal patterns. *Cognitive, Affective, and Behavioral Neuroscience* 2011; **11**(1):52–67.
8. Anderson JR, Fincham JM. Discovering the sequential structure of thought. *Cognitive Science* 2014; **38**(2):322–352.
9. Anderson JR, Lebiere C. *The Atomic Components of Thought*. Erlbaum: Mahwah, NJ, 1998.
10. Anderson JR. *How can the Human Mind Occur in the Physical Universe?* Oxford University Press: New York, 2007.
11. Anderson JR, Gluck K. What role do cognitive architectures play in intelligent tutoring systems? In *Cognition & Instruction: Twenty-five Years of Progress*, Klahr D, Carver Sm (eds). Erlbaum, 2001; 227–262.
12. Anderson JR, Betts S, Ferris JL, Fincham JM. Neural imaging to track mental states while using an intelligent tutoring system. *Proceedings of the National Academy of Sciences* 2010; **107**(15):7018–7023.
13. Woods RP, Grafton ST, Holmes CJ, Cherry SR, Mazziotta JC. Automated image registration: I. General methods and intrastudent intramodality validation. *Journal of Computer Assisted Tomography* 1998; **22**:139–152.
14. Rosenberg-Lee M, Lovett M, Anderson JR. Neural correlates of arithmetic calculation strategies. *Cognitive, Affective & Behavioral Neuroscience* 2009; **9**:270–285.
15. Robert CP, Rydén T, Titterington DM. Bayesian inference in hidden Markov models through jump Markov chain Monte Carlo. *Journal of the Royal Statistical Society: Series B* 2000; **62**(1):57–75.
16. Sisson S. Trans-dimensional Markov chains: a decade of progress and future perspectives. *Journal of the American Statistical Association* 2005; **100**:1077–1089.
17. Green PJ, Hastie DI. *Reversible jump MCMC*, 2009. https://people.maths.bris.ac.uk/~mapjg/papers/rjmcmc_20090613.pdf.
18. Brooks SP, Giudici P, Roberts GO. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B* 2003; **65**(1):3–55.
19. Viviani R, Grön G, Spitzer M. Functional principal component analysis of fMRI data. *Human Brain Mapping* 2005; **24**: 109–129.
20. Ramsay JO, Silverman BW. *Functional Data Analysis*. Springer: Berlin, 1997.
21. Berkes I, Gabrys R, Horváth L, Kokoszka P. Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society* 2009; **71**(5):927–946.

22. Hörmann S, Kokoszka P. Weakly dependent functional data. *The Annals of Statistics* 2010; **38**:1845–1884.
23. Hörmann S, Kidziński Ł, Hallin M. Dynamic functional principal components. *Journal of the Royal Statistical Society: Series B* 2015; **77**:319–348.
24. Ramsay JO, Silverman BW. *Functional data analysis software, R edition*. http://www.psych.mcgill.ca/misc/fda/downloads/FDAfuns/R/R.
25. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**:773–795.