nature methods

ARTICLES

https://doi.org/10.1038/s41592-020-0941-6

A temporal decomposition method for identifying venous effects in task-based fMRI

Kendrick Kay[1]✉, Keith W. Jamison[1,4], Ru-Yuan Zhang[1,2,3] and Kamil Uğurbil[1]

The spatial resolution of functional magnetic resonance imaging (fMRI) is fundamentally limited by effects from large draining veins. Here we describe an analysis method that provides data-driven estimates of these effects in task-based fMRI. The method involves fitting a one-dimensional manifold that characterizes variation in response timecourses observed in a given dataset, and then using identified early and late timecourses as basis functions for decomposing responses into components related to the microvasculature (capillaries and small venules) and the macrovasculature (large veins), respectively. We show the removal of late components substantially reduces the superficial cortical depth bias of fMRI responses and helps eliminate artifacts in cortical activity maps. This method provides insight into the origins of the fMRI signal and can be used to improve the spatial accuracy of fMRI.

Among the handful of noninvasive techniques that permit the study of human brain activity, fMRI based on blood oxygenation level dependent (BOLD) contrast has emerged as the most widely used approach in cognitive neuroscience. A primary advantage of fMRI over other measurement techniques is its spatial resolution. However, efforts to increase the spatial resolution of fMRI—especially to reach the submillimeter scale of mesoscopic brain organization—face a major challenge imposed by the 'draining vein' confound, first noted early in the history of fMRI[1]. Venous effects may appear as activation that is displaced from the original site of neural activity by as much as 4 mm (ref. [2]) and may reflect neural activity pooled over large spatial scales, thus degrading spatial specificity[3–6]. The field has long sought to measure BOLD responses from the microvasculature (capillaries and small venules) while avoiding BOLD responses from the macrovasculature (large veins)[7–9]. The confounding effects of the macrovasculature are especially critical to resolve given the growing popularity in the neuroscience community of using fMRI to probe laminar-specific responses[10–12].

To avoid the specificity loss caused by veins, the field has traditionally turned to the use of spin-echo acquisition at ultra-high magnetic fields[8,13] instead of conventional gradient-echo acquisition. However, spin-echo acquisition involves increased energy deposition, longer volume acquisition times and lower BOLD contrast-to-noise ratio compared to gradient-echo acquisition. Thus, to maintain measurement sensitivity, the experimenter is generally forced to reduce spatial coverage and/or substantially increase the amount of data collected per experimental condition. These trade-offs are often prohibitive for neuroscientists, given that measuring multiple brain regions is often critical, the sensitivity of fMRI is already relatively low to start with and increasing the duration of an experiment beyond more than a factor of two is often impractical.

Here we introduce an analysis method, called temporal decomposition through manifold fitting (TDM), that identifies and removes venous-related signals from task-based fMRI data. The TDM method is simple, principled and is compatible with a variety of experimental protocols including those based on gradient-echo acquisitions. We demonstrate TDM on visual experiments conducted in human participants (with BOLD fMRI at 7 T and 3 T), and show that TDM consistently removes venous effects while maintaining a reasonable level of sensitivity. The data used in this paper (both raw and preprocessed), code implementing TDM and a video tutorial are freely available at https://osf.io/j2wsc/.

Results

TDM provides a method for visualizing timecourse variation. In our experimental datasets involving brief presentations of visual stimuli (summary provided in Supplementary Fig. 1), we replicate previous observations[14–17] that BOLD timecourse amplitude, delay and width increase from inner to outer cortical depths and from lighter to darker voxel intensities (Supplementary Note 1 and Supplementary Figs. 2 and 3). These effects are consistent with the interpretation that large draining veins (which tend to reside near the pial surface and cause signal dephasing) lead to changes in BOLD timecourses, and they set the stage for the TDM method.

The first step in TDM is to take measured BOLD time-series data, extract response timecourses by fitting a finite impulse response (FIR) model and visualize distributions of these response timecourses in a low-dimensional space (Fig. 1). Specifically, we use principal components analysis (PCA) to determine the three orthogonal timecourse components that account for the most variance in the response timecourses. We then use these three timecourse components as axes of a three-dimensional (3D) space (PC1, PC2, PC3) in which each of the response timecourses corresponds to one point, and an associated vector, in this space. To visualize the results, we map the timecourse vectors to a unit sphere (by normalizing them to unit length) and use an orthographic projection to visualize the density of the timecourse vectors. Such a visualization reveals the shapes of commonly occurring response timecourses, independent of their amplitudes (density image in Fig. 1). We separately visualize the amplitudes of the response timecourses by com-

[1]Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Minneapolis, MN, USA. [2]Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai, China. [3]Institute of Psychology and Behavioral Science, Shanghai Jiao Tong University, Shanghai, China. [4]Present address: Department of Radiology, Weill Cornell Medicine, New York, NY, USA. ✉e-mail: kay@umn.edu

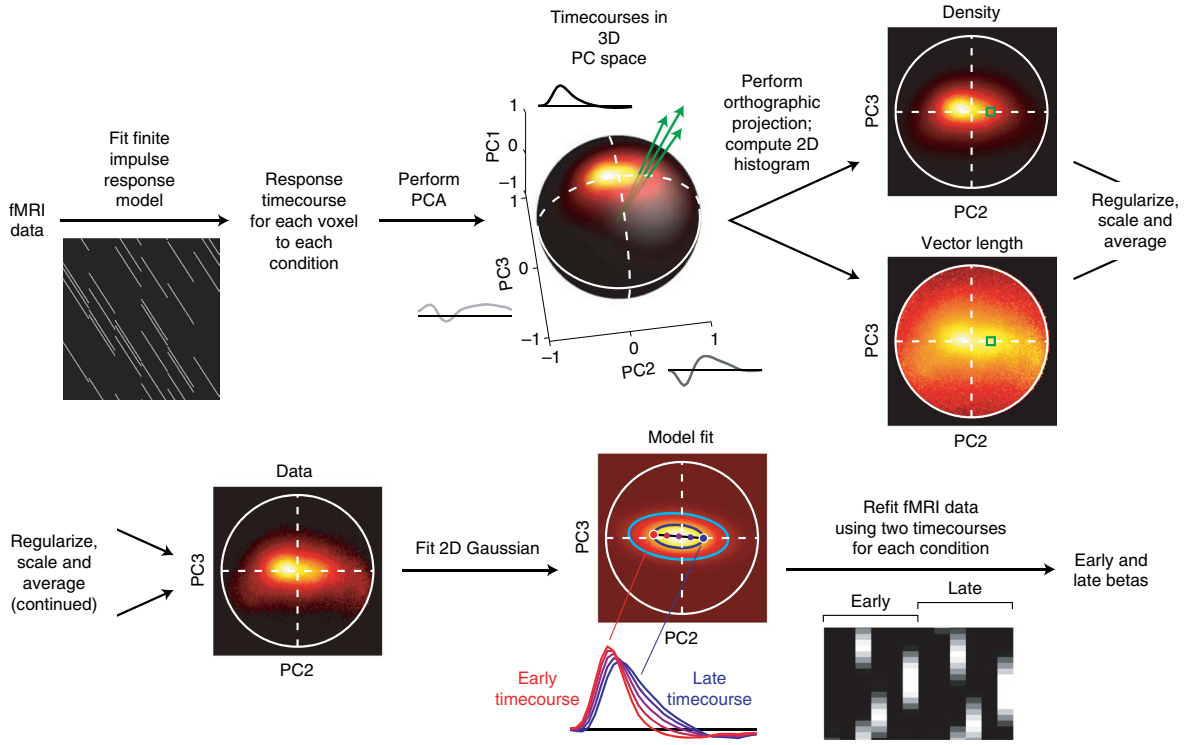


Fig. 1 | Schematic of the TDM method. Time-series data are fit with a FIR model to estimate response timecourses. PCA is performed on the timecourses to reduce their dimensionality to three. Using orthographic projection in the direction of the first PC, a 2D histogram image is calculated (density). The same projection and binning scheme is used to calculate an image representing timecourse amplitude (vector length). The two images are combined and fit with a 2D Gaussian to determine an early timecourse and a late timecourse that together summarize the principal axis of variation. Finally, the time-series data are refitted with a model incorporating the two timecourses.

puting the lengths of the original timecourse vectors and repeating the orthographic visualization (vector-length image in Fig. 1).

Applying these visualization procedures to a representative dataset (we obtained similar results in additional datasets, see Extended Data Figs. 1 and 2), we find that response timecourses typically reside near the pole of the unit circle where PC1 is maximal, with some variability around this pole (Fig. 2a). We find that the amplitudes of these timecourses are large in a similar portion of the space except for a small extension toward the lower left (Fig. 2a), thus highlighting the complementary information provided by these two plots. A separate plot shows the three principal component (PC) timecourse components that define the axes of the space (Fig. 2b). Response timecourses generally resemble a canonical hemodynamic response timecourse (PC1; Fig. 2b, black line), and vary primarily according to how strongly they load on a timecourse component (PC2; Fig. 2b, dark gray line) that shifts the peak either earlier or later in time.

TDM identifies an axis of timecourse variation. The second step in TDM is to identify an axis that captures the main variation in the observed response timecourses. As seen earlier (Fig. 2a), response timecourses empirically occupy a small portion of the 3D space. Furthermore, the timecourses can be approximated in the 3D space by a simple line segment defined on the sphere (arc) that characterizes variation with respect to timecourse delay; that is, early versus late. The interpretation we adopt here is that the two endpoints of the line segment correspond to latent hemodynamic timecourses associated with the microvasculature and the macrovasculature, respectively; and any single observed timecourse is simply a mixture of these two latent timecourses plus measurement noise (which causes deviation away from the line).

To calculate the axis of variation, TDM combines density and vector length, fits a two-dimensional (2D) Gaussian to the result, and extracts points positioned at ±1 s.d. along the main axis of the fitted Gaussian (Fig. 1). Examining results obtained on the representative dataset, we find that the combined image resembles both density and vector length, the fitted Gaussian approximates the data (Fig. 2c) and the extracted points reside in sensible locations (Fig. 2a). We reconstruct timecourses corresponding to the two extracted points (Fig. 2b, red and blue lines), and then label the timecourses 'early' and 'late' based on the time-to-peak of the reconstructed timecourses. Note that different mixtures of the early and late timecourses trace out an arc on the unit sphere (Fig. 2a, black line) and result in a continuum of timecourse shapes (Fig. 1).

We interpret the early and late timecourses as reflecting the microvasculature and macrovasculature, respectively. We offer several lines of reasoning that suggest the validity of this interpretation. First, we find that the late timecourse is consistently associated with large vector length (see Fig. 2a, center and Extended Data Fig. 1), indicating that response timecourses resembling the late timecourse tend to have large BOLD amplitudes. This is reasonable given that veins exhibit large percentage BOLD signal changes[1,16]. Second, if we use the same visualization methods (orthographic projection of the unit sphere) to examine the relationship between timecourse shape and bias-corrected echo planar imaging (EPI) intensity, we find that the late timecourse is consistently positioned in a zone of low EPI intensity (Fig. 2a, right plot and Extended Data Fig. 1). Since the TDM procedure does not make use of EPI intensities, this is an empirical finding that provides further evidence of validity, as it is known that veins cause static susceptibility effects in EPI images[1,4,18]. Third, the idea that veins exhibit delayed BOLD

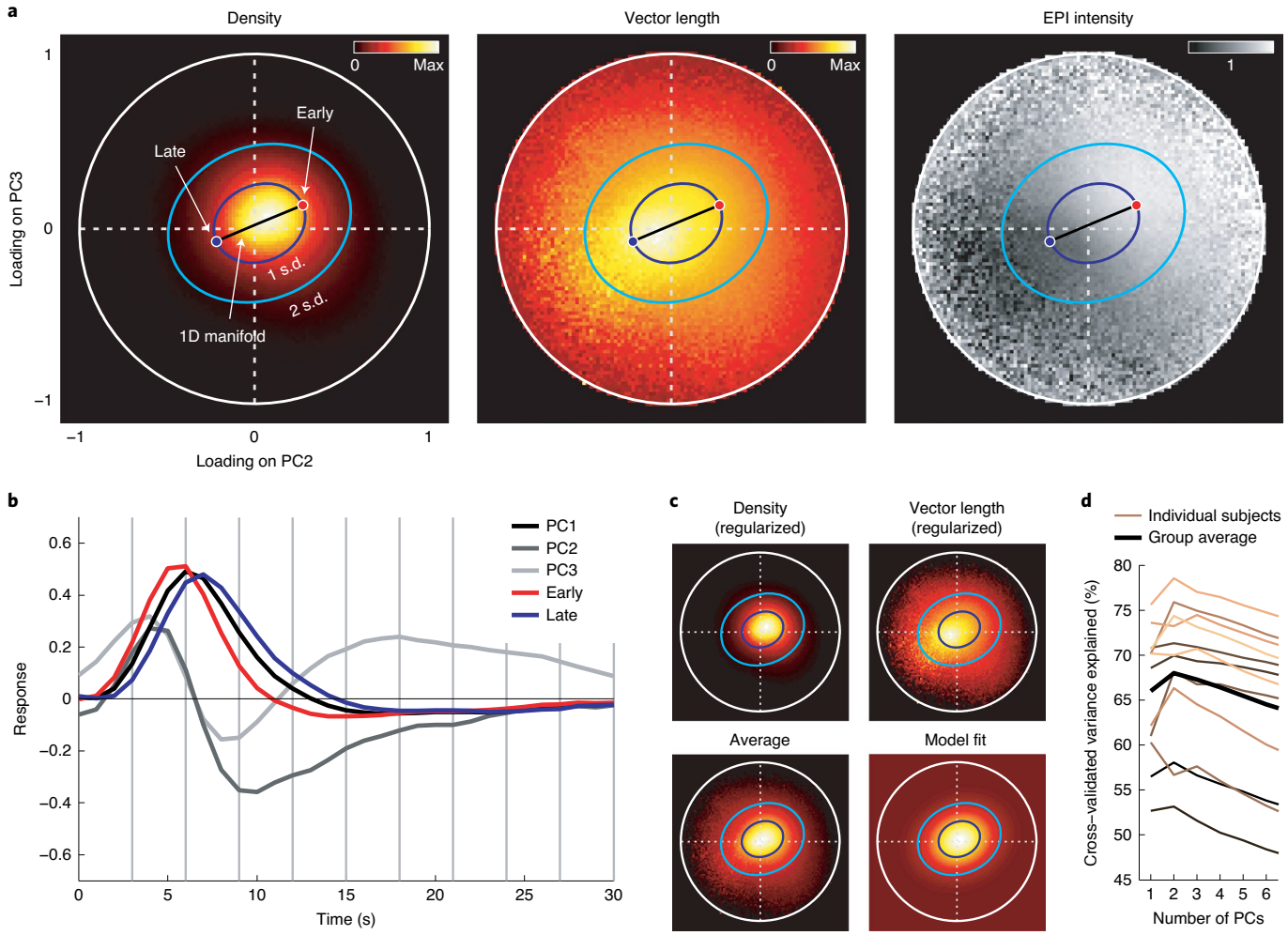


Fig. 2 | TDM captures timecourse variation along a one-dimensional manifold. a–c, Results for Dataset D9. TDM-calculated density image (**a**) showing shapes of commonly occurring timecourses (left), vector-length image showing timecourse amplitudes (middle), and an image showing bias-corrected EPI intensities (right). The black line indicates the identified one-dimensional (1D) arc that connects the early and late timecourses. Timecourses (**b**). All timecourses are unit-length vectors. Vertical gray lines mark 3-s intervals. Gaussian fitting procedure (**c**). Density and vector-length images are baseline-subtracted, scaled and truncated, producing regularized images (upper left, upper right). These images are then averaged (lower left) and fit with an oriented 2D Gaussian (lower right). **d**, Dimensionality of response timecourses for datasets D1–D12. We perform PCA on split halves of each dataset and assess how well a limited number of PC timecourses from one half will reconstruct timecourses measured in the other half.

responses is consistent with several previous experimental studies[14–17]. Finally, a biophysical model of vascular dynamics provides a potential explanation for why temporal delays occur in veins[19]. Additional control analyses further support the validity of TDM (Fig. 2d and Supplementary Note 2).

Summary of TDM timecourses across participants. For a comprehensive assessment of results, we plot the TDM-derived early and late timecourses obtained for each dataset (Fig. 3a). On the whole, the timecourses are stereotyped and largely consistent across datasets. However, there is substantial quantitative variability, consistent with the observation that BOLD timecourses vary across participants[20,21]. Computing the group-average early and late timecourses (Fig. 3a), we see that both timecourses are similar in overall shape and differ primarily in their delay and width. However, the timecourses also exhibit differences in the timing and magnitude of the post-stimulus undershoot; this undershoot influences the modeling of the time-series data and should not be overlooked. We next fit each group-average timecourse using a double-gamma function, and find that smooth parametric functions characterize the empirical results well (Fig. 3b). Finally, as a point of comparison, we plot the predicted timecourse for a hypothetical 4-s event using the default double-gamma hemodynamic response function (HRF) implemented in the commonly used software SPM. This timecourse coincides well with the group-average late timecourse (Fig. 3b). This makes sense, given that the default timecourse parameters in SPM were derived from fMRI measurements conducted at low (2 T) magnetic field strength[22] where the BOLD response is dominated by contributions from large vessels[23,24].

Decomposition using TDM timecourses removes artifacts from cortical maps. The final step of TDM involves analyzing the fMRI data using a general linear model (GLM) that incorporates the early and late timecourses time-locked to the onsets of each experimental condition. Fitting this GLM produces, for each vertex (or voxel) and condition, an estimate of the BOLD response amplitude from the microvasculature and an estimate of the BOLD response amplitude from the macrovasculature, respectively. These response

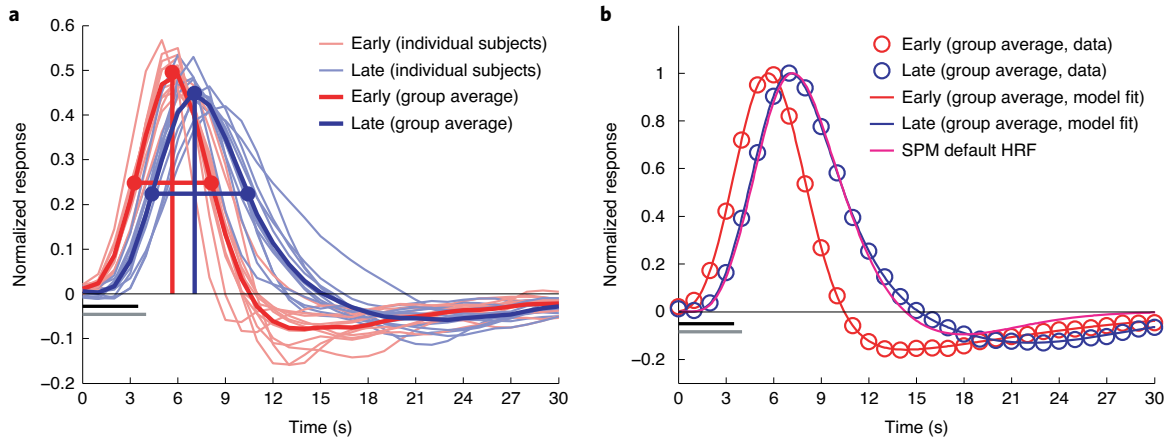


Fig. 3 | Early and late timecourses found by TDM. a, Summary of results. Unit-length-normalized early and late timecourses found in datasets D1–D12 (thin lines) and their average (thick lines) are plotted. Dots mark timecourse peaks and half-maximum response times. Horizontal black and gray bars below the *x* axis indicate the stimulus duration in the eccentricity and category experiments (3.5 and 4 s, respectively). **b**, Group-average results and parametric model fit. Group-average early and late timecourses are normalized to peak at 1 (red and blue circles), and are fit with a double-gamma function as implemented in SPM's spm_hrf.m (red and blue lines). The fitting was achieved by convolving a double-gamma function with a 4-s square wave and minimizing squared error with respect to the data. Estimated double-gamma parameters for the group-average early and late timecourses were (7.21 17.6 0.5 4.34 1.82 -3.09 50) and (5.76 21.6 1.11 1.72 3.34 0.193 50), respectively. As a comparison, we show the timecourse obtained using default SPM parameters (6 16 1 1 6 0 32) (magenta line).

amplitudes, or betas, can then be used in subsequent analyses according to the goals of the researcher. An important observation is that the early and late timecourses are often overlapping and correlated (Fig. 2a). In the GLM, the two timecourses are fit simultaneously to the data to optimally explain the measured BOLD time-series data (Fig. 1).

To assess the quality of the early and late betas, we generate cortical surface visualizations and compare these against visualizations of betas obtained using a standard GLM that incorporates a single canonical HRF time-locked to each condition (hereby referred to as standard betas). We focus specifically on visualizations for datasets D1–D5, which involved presentation of stimuli that vary in eccentricity (Supplementary Fig. 1), because studies of the visual system provide well-established 'ground-truth' expectations for neural activity patterns elicited by stimuli varying in eccentricity: in brief, neurons in early visual cortex respond selectively to stimuli at specific eccentricities, and the preferred eccentricity varies smoothly from the fovea to the periphery along the posterior-to-anterior direction[25].

Inspecting results for a representative dataset (Fig. 4, Extended Data Fig. 3 and Supplementary Note 3), we find that the early and late betas show different patterns. The early betas are relatively homogeneous across the cortical surface, relatively flat across cortical depth and are moderate in size at around 1–4% signal change. In contrast, the late betas are heterogeneous across the cortical surface (sparsely distributed), biased toward outer cortical depths and are sometimes large in size, reaching 10% or more signal change. The observation of additional activations arising in the late betas is consistent with the fact that the BOLD point-spread function appears larger when sampling late in the BOLD response[6]. Furthermore, comparing the spatial pattern of the late betas against the spatial pattern of bias-corrected EPI intensities (Fig. 4), we see a general correspondence between the sparsely distributed locations where large BOLD responses are observed and regions with dark EPI intensities.

In maps of peak eccentricity tuning, all three versions of the betas (standard, early, late) exhibit the expected smooth large-scale progression from foveal (blue) to peripheral (red) eccentricities from posterior (left) to anterior (right) in early visual cortex (Fig. 4). However, the quality or robustness of the eccentricity map is highest for the standard betas, moderately high for the early betas and low for the late betas. Moreover, for the late betas, there is a substantial decrease in quality moving from outer to inner cortical depths; this is consistent with the sharp fall-off in the magnitude of betas moving from outer to inner depths (Fig. 4). We also observe that although large-scale eccentricity patterns are similar across the three versions of the betas, the maps show divergence at a fine scale. In particular, there are artifacts in eccentricity tuning that are present in the standard and late betas but absent in the early betas.

To further clarify these results, we examine activity patterns elicited by a single experimental condition (Fig. 4). Based on known tuning properties of early visual cortex[25,26], we expect a compact 'stripe' of positive activity extending along the superior–inferior direction. All versions of the activity pattern (standard, early, late) indeed show evidence of a stripe. However, only the early version exhibits a well-behaved stripe that is relatively homogeneous within its spatial extent and relatively flat across cortical depth. Detailed illustrations of how TDM demixes early and late timecourses are shown in Extended Data Figs. 4 and 5.

Line profiles and comparison to simpler methods. For quantitative and more comprehensive assessment of single-condition activity patterns, we extract line profiles along iso-angle contours in primary visual cortex (V1) from all participants (Fig. 5a). This analysis also provides an opportunity to directly compare against results of an alternative approach for avoiding responses from the macrovasculature, namely, sampling early timepoints in evoked BOLD responses[6,27]. The results show that activity patterns constructed from standard and late betas exhibit large and idiosyncratic responses that are variable across participants and biased toward outer cortical depths. In contrast, activity patterns constructed from early betas exhibit more focal activations that are consistent across participants and relatively homogeneous across depth. For the timepoint-based approach, we use timecourse estimates provided by a FIR model and examine responses at 2, 3, 4 and 5 s after trial onset. We find that at 2 s, responses are weak. At 5 s, responses are strong but closely resemble the large and idiosyncratic responses observed in the standard and late analyses. Intermediate timepoints

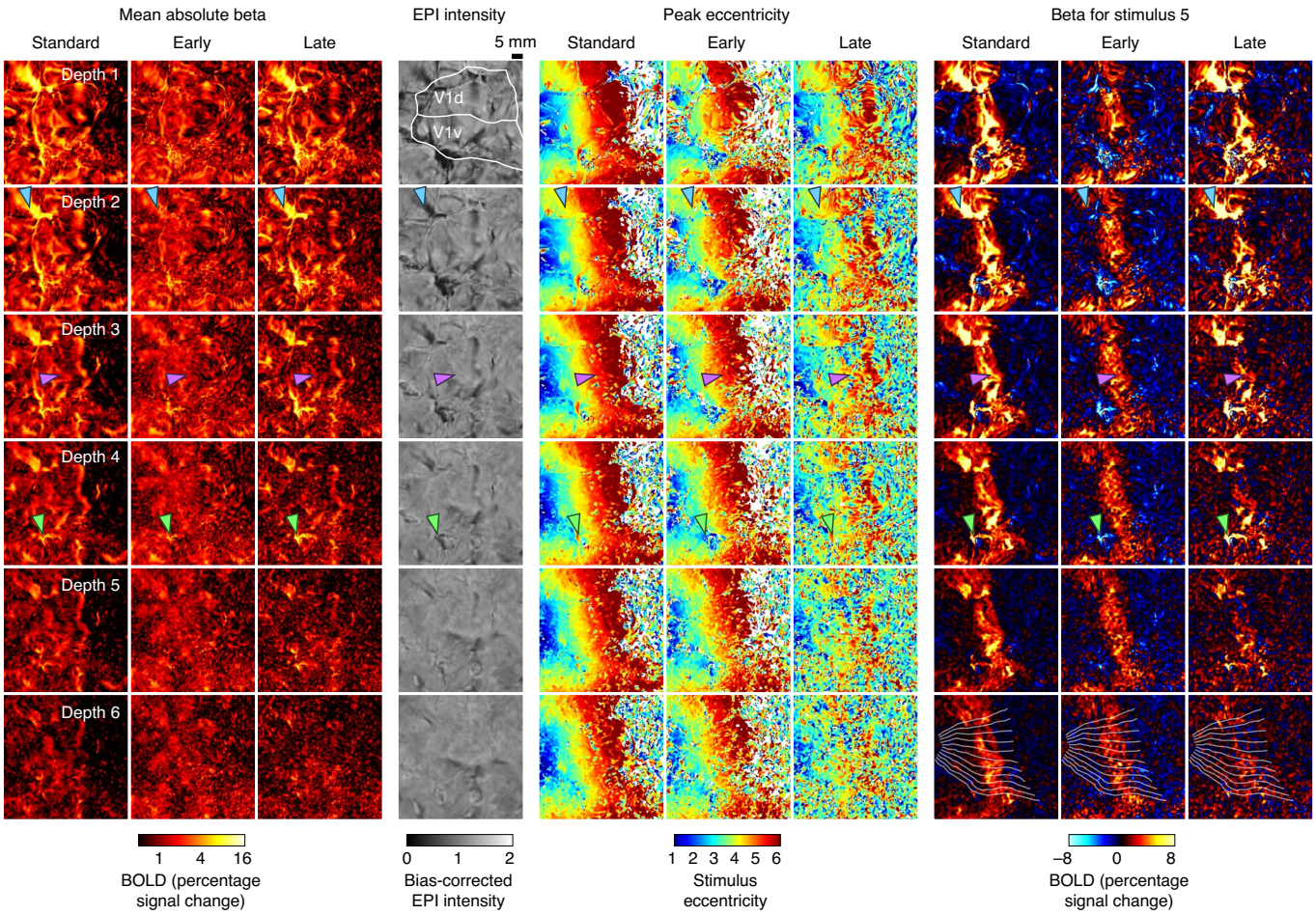


Fig. 4 | TDM decomposes brain activity patterns into early and late components. Results from dataset D1. Rows correspond to different cortical depths (depth 1 is most superficial) for a small patch covering primary visual cortex (flattened surface, left hemisphere). Three versions of results are shown: the first (standard) reflects betas from a GLM incorporating a single canonical HRF, while the next two (early, late) reflect betas from a GLM incorporating the timecourses found by TDM. Four types of image are shown: (1) absolute value of betas averaged across conditions, (2) bias-corrected EPI intensities, (3) peak eccentricity quantified as the center of mass of the six betas corresponding to different eccentricities and (4) single-condition activity patterns for the fifth stimulus. Blue, purple and green arrows mark vertices illustrated in greater detail in Extended Data Fig. 4. Thin gray lines are an example of the lines plotted in Fig. 5.

3 and 4 s perform better at avoiding the idiosyncratic responses, but are weaker in magnitude.

To summarize performance of the different methods, we perform a split-half analysis in which we calculate the similarity (Pearson's correlation) between all pairs of activity profiles produced by the various analysis approaches as well as additional analyses that either use a single HRF matched to the TDM-derived early timecourse or a single HRF matched to the TDM-derived late timecourse (Fig. 5b). This analysis reveals that early responses (location i, $r=0.76$) have higher split-half reliability than FIR (3 s) responses (location ii, $r=0.63$). Although early responses (location i, $r=0.76$) have lower split-half reliability than FIR (5 s) responses (location iii, $r=0.88$), the FIR (5 s) responses resemble both early and late responses (location v; $r=0.61$, $r=0.57$) and thus cannot be interpreted as avoiding effects from the macrovasculature. Even FIR (3 s) responses have equitable balance between early and late responses (location iv; $r=0.55$, $r=0.31$). This is not surprising given that macrovascular timecourses already exhibit substantial rise 3 s after trial onset (Fig. 3). We also observe that the analyses involving a single HRF produce responses that are similar to the standard responses (location vi; $r=0.95$, $r=0.97$), and that using the TDM-derived early timecourse as the single HRF still produces responses that reflect substantial influence of both early and late responses (location vii; $r=0.58$, $r=0.66$).

We acknowledge that the split-half analysis does not quantify the accuracy of the activity profiles generated by the different methods, as a quantitative ground-truth measure is not available. Nevertheless, the results indicate that although sampling early timepoints can help avoid macrovascular responses, TDM outperforms the timepoint-based analysis in terms of specificity (avoidance of the macrovasculature) and sensitivity (reliability of response estimates). In addition, the results indicate that the benefits of TDM come from simultaneously including both early and late timecourses in the model, and are not obtained by simply modifying the HRF used in a conventional GLM.

Discussion

Our analysis approach can be viewed as comprising two distinct components. The first is our data-driven method for deriving early and late timecourses. This method is fully automated and principled (as opposed to a heuristic method that might attempt to quickly find a few 'fast' voxels), and might be valuable in and of itself for extracting hemodynamic timecourses for comparison across brain regions, individuals and/or groups. The second is the application

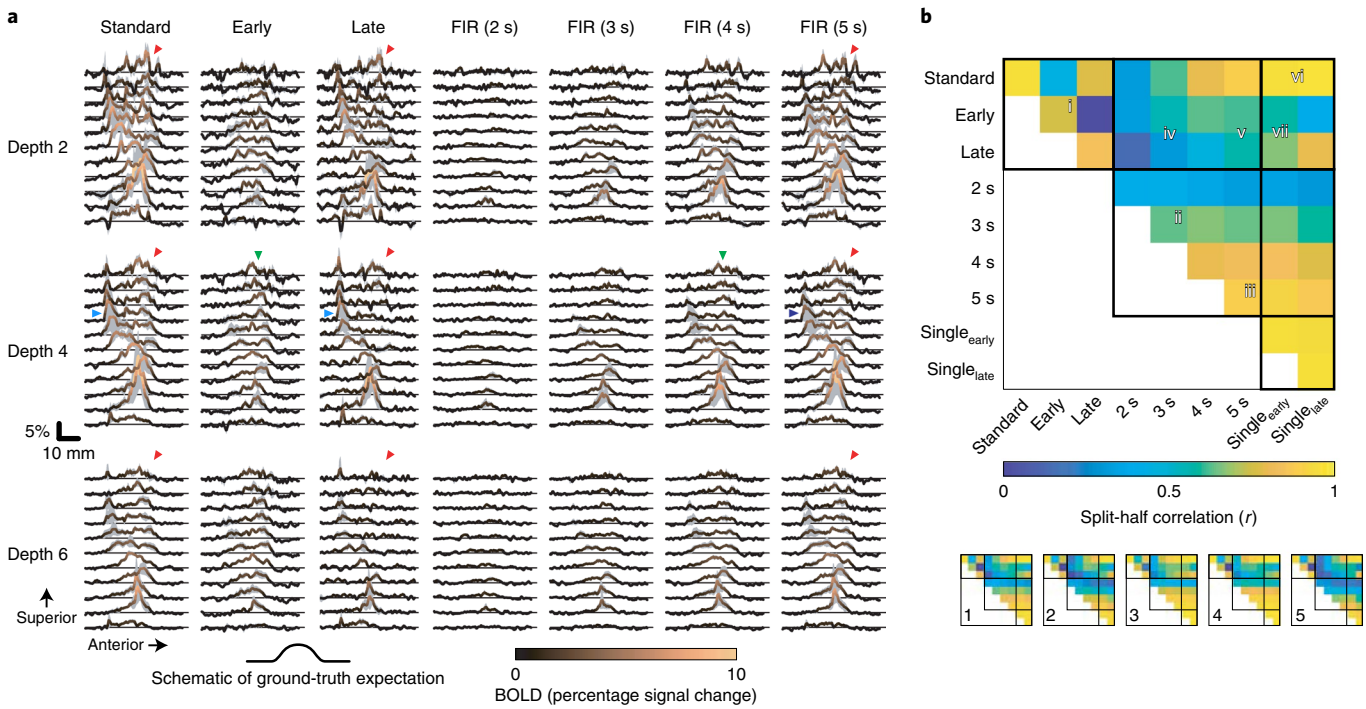


Fig. 5 | TDM outperforms a simple timepoint-based analysis. a, Line profiles of BOLD activity evoked by the fourth eccentricity stimulus for 11 iso-angle lines in right hemisphere V1 (see thin gray lines in Fig. 4). Each line indicates the mean across datasets (D1–D5); gray shaded regions indicate standard error across datasets. Red arrows show large responses at outer cortical depths. Blue arrows show artifacts. Green arrows show activity profiles that are homogeneous across depth and compact in shape. **b**, Similarity analysis. Split-half correlations between activity profiles produced by different analysis approaches. The main plot shows the group-average result; the smaller plots show results for individual datasets.

we demonstrate for using these timecourses to estimate and reduce venous effects in task-based fMRI.

The value of the work presented here does not lie in the discovery of a new phenomenon: the idea that veins carry delayed responses is not new[14,16,17], nor is the idea that early responses have the potential to be more spatially specific[6,16,28–30]. However, making an observation regarding a phenomenon and having a robust method that can exploit that observation in practice to generate biologically informative results are two different contributions. Our work introducing the design and validation of analysis methodology primarily falls in the latter category.

It is important to note that TDM incurs some loss in sensitivity due to correlation between early and late timecourses (Supplementary Note 3 and Extended Data Fig. 6). Nonetheless, we believe TDM has distinct advantages over other methods that have been proposed for dealing with venous effects in fMRI[5,6,13,27,31–40] (Supplementary Discussion).

Algorithmically, TDM uses a manifold-fitting method to characterize latent structure in timecourse variations. There are other methods that can characterize latent structure; two widely used methods are PCA and independent components analysis (ICA). However, due to the orthogonality constraint in PCA, it is necessarily the case that the PC timecourses returned by PCA are orthogonal. Although TDM does make use of PCA to determine the 3D space within which to perform further analyses, the PCA timecourses themselves do not constitute good candidates for latent timecourses because there is no reason to expect hemodynamic timecourses in the brain to be orthogonal.

In contrast to PCA, ICA does not impose the constraint of orthogonality. Instead, ICA optimizes timecourses with respect to statistical independence, often through some measure of non-Gaussianity (for example, kurtosis). Although derived timecourses from an ICA-based procedure are sometimes similar to those produced by TDM (Extended Data Fig. 1), the TDM method has advantages. First, the data visualization and explicit modeling performed by TDM allow the user to evaluate and confirm the data features that give rise to the derived timecourses. ICA, without further analysis, remains a 'black box' and it is difficult to understand the specific features of the data that give rise to its results. Second, there is no a priori reason to think that loadings on early and late hemodynamic timecourses must necessarily conform to statistical independence. Third, ICA alone does not identify the early and late timecourses; rather, we found it necessary to couple the results of ICA with several post hoc procedures that are heuristic in nature. On the whole, we suggest that TDM is more explicit, more direct and more interpretable than ICA.

GLM-based analyses of fMRI time-series data sometimes allow flexible modeling of timecourse shape through the inclusion of a canonical hemodynamic response timecourse and its temporal derivative[22] or some other basis function decomposition such as PCA[41]. While TDM shares the common feature of providing a means to capture timecourse variation, the key difference with respect to these alternative approaches lies in the specific timecourses that are chosen by TDM. The early and late timecourses found by TDM are often correlated (unlike a timecourse and its derivative or those returned by PCA). Moreover, the early and late timecourses have specific biological meanings, and so the beta loadings found for these timecourses have specific value. It is possible that alternative timecourse models can yield fits to a set of data that are as good as the fit achieved by TDM, but the beta loadings associated with these models cannot be interpreted in terms of the microvasculature and macrovasculature.

In this study, we have demonstrated that TDM delivers robust and meaningful results in each of our fMRI datasets, which include not only gradient-echo but also spin-echo and low-resolution acquisitions (Supplementary Note 4). We have also made freely

available data and analysis code to ensure that the TDM method is transparent and reproducible[42]. However, efforts to further assess and validate the TDM method would nonetheless be useful. Further work could be directed at assessing and optimizing the technique with respect to experimental design characteristics such as the duration of experimental conditions, the spatial and temporal resolution of the acquisition and the amount of data acquired. In addition, it would be worthwhile to test the technique on other types of experiment (other sensory, cognitive and/or motor experiments) and other brain areas. It would be interesting to assess how well TDM can resolve fine-scale variation in neural representations, such as ocular dominance columns[43]. Since TDM makes no restrictions on the spatial loadings of the early and late timecourses, the technique should in principle be applicable not only to large-scale neural representations such eccentricity, but also fine-scale representations such as ocular dominance.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-020-0941-6.

Received: 6 December 2019; Accepted: 3 August 2020;
Published online: 07 September 2020

References

1. Menon, R. S., Ogawa, S., Tank, D. W. & Ugurbil, K. Tesla gradient recalled echo characteristics of photic stimulation-induced signal changes in the human primary visual cortex. *Magn. Reson. Med.* **30**, 380–386 (1993).
2. Turner, R. How much cortex can a vein drain? Downstream dilution of activation-related cerebral blood oxygenation changes. *NeuroImage* **16**, 1062–1067 (2002).
3. Bianciardi, M., Fukunaga, M., van Gelderen, P., de Zwart, J. A. & Duyn, J. H. Negative BOLD-fMRI signals in large cerebral veins. *J. Cereb. Blood Flow Metab.* **31**, 401–412 (2011).
4. Kay, K. et al. A critical assessment of data quality and venous effects in sub-millimeter fMRI. *NeuroImage* **189**, 847–869 (2019).
5. Olman, C. A., Inati, S. & Heeger, D. J. The effect of large veins on spatial localization with GE BOLD at 3 T: displacement, not blurring. *NeuroImage* **34**, 1126–1135 (2007).
6. Shmuel, A., Yacoub, E., Chaimow, D., Logothetis, N. K. & Ugurbil, K. Spatio-temporal point-spread function of fMRI signal in human gray matter at 7 Tesla. *NeuroImage* **35**, 539–552 (2007).
7. Cheng, K. Exploration of human visual cortex using high spatial resolution functional magnetic resonance imaging. *NeuroImage* **164**, 4–9 (2018).
8. Ugurbil, K. What is feasible with imaging human brain function and connectivity using functional magnetic resonance imaging. *Phil. Trans. R. Soc.* **371**, 20150361 (2016).
9. Yacoub, E. & Wald, L. L. Pushing the spatio-temporal limits of MRI and fMRI. *NeuroImage* **164**, 1–3 (2018).
10. De Martino, F. et al. The impact of ultra-high field MRI on cognitive and computational neuroimaging. *NeuroImage* **168**, 366–382 (2018).
11. Dumoulin, S. O., Fracasso, A., van der Zwaag, W., Siero, J. C. W. & Petridou, N. Ultra-high field MRI: advancing systems neuroscience towards mesoscopic human brain function. *NeuroImage* **168**, 345–357 (2018).
12. Lawrence, S. J. D., Formisano, E., Muckli, L. & de Lange, F. P. Laminar fMRI: applications for cognitive neuroscience. *NeuroImage* **197**, 785–791 (2017).
13. Yacoub, E., Harel, N. & Ugurbil, K. High-field fMRI unveils orientation columns in humans. *Proc. Natl Acad. Sci. USA* **105**, 10607–10612 (2008).
14. de Zwart, J. A. et al. Temporal dynamics of the BOLD fMRI impulse response. *NeuroImage* **24**, 667–677 (2005).
15. Kim, J. H. & Ress, D. Reliability of the depth-dependent high-resolution BOLD hemodynamic response in human visual cortex and vicinity. *Magn. Reson. Imaging* **39**, 53–63 (2017).
16. Lee, A. T., Glover, G. H. & Meyer, C. H. Discrimination of large venous vessels in time-course spiral blood-oxygen-level-dependent magnetic-resonance functional neuroimaging. *Magn. Reson. Med.* **33**, 745–754 (1995).
17. Siero, J. C. W., Petridou, N., Hoogduin, H., Luijten, P. R. & Ramsey, N. F. Cortical depth-dependent temporal dynamics of the BOLD response in the human brain. *J. Cereb. Blood Flow Metab.* **31**, 1999–2008 (2011).
18. Ogawa, S., Lee, T. M., Nayak, A. S. & Glynn, P. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn. Reson. Med.* **14**, 68–78 (1990).
19. Havlicek, M. & Uludağ, K. A dynamical model of the laminar BOLD response. *NeuroImage* **204**, 116209 (2020).
20. Handwerker, D. A., Gonzalez-Castillo, J., D'Esposito, M. & Bandettini, P. A. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *NeuroImage* **62**, 1017–1023 (2012).
21. Taylor, A. J., Kim, J. H. & Ress, D. Characterization of the hemodynamic response function across the majority of human cerebral cortex. *NeuroImage* **173**, 322–331 (2018).
22. Friston, K. J. et al. Event-related fMRI: characterizing differential responses. *NeuroImage* **7**, 30–40 (1998).
23. Haacke, E. M. et al. 2D and 3D high resolution gradient echo functional imaging of the brain: venous contributions to signal in motor cortex studies. *NMR Biomed.* **7**, 54–62 (1994).
24. Uludağ, K., Müller-Bierl, B. & Uğurbil, K. An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging. *Neuroimage* **48**, 150–165 (2009).
25. Wandell, B. & Winawer, J. Imaging retinotopic maps in the human brain. *Vision Res.* **51**, 718–737 (2011).
26. Wandell, B. & Winawer, J. Computational neuroimaging and population receptive fields. *Trends Cog. Sci.* **19**, 349–357 (2015).
27. Goodyear, B. G. & Menon, R. S. Brief visual stimulation allows mapping of ocular dominance in visual cortex using fMRI. *Hum. Brain Mapp.* **14**, 210–217 (2001).
28. Menon, R. S. & Goodyear, B. G. Submillimeter functional localization in human striate cortex using BOLD contrast at 4 Tesla: implications for the vascular point-spread function. *Magn. Reson. Med.* **41**, 230–235 (1999).
29. Yu, X., Qian, C., Chen, D., Dodd, S. J. & Koretsky, A. P. Deciphering laminar-specific neural inputs with line-scanning fMRI. *Nat. Methods* **11**, 55–58 (2014).
30. Yu, X. et al. Direct imaging of macrovascular and microvascular contributions to BOLD fMRI in layers IV-V of the rat whisker-barrel cortex. *NeuroImage* **59**, 1451–1460 (2012).
31. De Martino, F. et al. Cortical depth dependent functional responses in humans at 7T: improved specificity with 3D GRASE. *PLoS ONE* **8**, e60514 (2013).
32. Fracasso, A., Luijten, P. R., Dumoulin, S. O. & Petridou, N. Laminar imaging of positive and negative BOLD in human visual cortex at 7T. *NeuroImage* **164**, 100–111 (2018).
33. Heinzle, J., Koopmans, P. J., den Ouden, H. E. M., Raman, S. & Stephan, K. E. A hemodynamic model for layered BOLD signals. *NeuroImage* **125**, 556–570 (2016).
34. Huber, L. et al. High-resolution CBV-fMRI allows mapping of laminar activity and connectivity of cortical input and output in human M1. *Neuron* **96**, 1253–1263.e7 (2017).
35. Lu, H., Golay, X., Pekar, J. J. & Van Zijl, P. C. M. Functional magnetic resonance imaging based on changes in vascular space occupancy. *Magn. Reson. Med.* **50**, 263–274 (2003).
36. Markuerkiaga, I., Barth, M. & Norris, D. G. A cortical vascular model for examining the specificity of the laminar BOLD signal. *NeuroImage* **132**, 491–498 (2016).
37. Marquardt, I., Schneider, M., Gulban, O. F., Ivanov, D. & Uludağ, K. Cortical depth profiles of luminance contrast responses in human V1 and V2 using 7 T fMRI. *Hum. Brain Mapp.* **464**, 1155 (2018).
38. Moerel, M. et al. Sensitivity and specificity considerations for fMRI encoding, decoding, and mapping of auditory cortex at ultra-high field. *NeuroImage* **164**, 18–31 (2018).
39. Olman, C. A. et al. Layer-specific fMRI reflects different neuronal computations at different depths in human V1. *PLoS ONE* **7**, e32536 (2012).
40. Polimeni, J. R., Fischl, B., Greve, D. N. & Wald, L. L. Laminar analysis of 7T BOLD using an imposed spatial activation pattern in human V1. *NeuroImage* **52**, 1334–1346 (2010).
41. Woolrich, M. W., Behrens, T. E. J. & Smith, S. M. Constrained linear basis sets for HRF modelling using variational Bayes. *NeuroImage* **21**, 1748–1761 (2004).
42. Poline, J. B. & Poldrack, R. A. Frontiers in brain imaging methods grand challenge. *Front. Neurosci.* **6**, 96 (2012).
43. Cheng, K., Waggoner, R. A. & Tanaka, K. Human ocular dominance columns as revealed by high-field functional magnetic resonance imaging. *Neuron* **32**, 359–374 (2001).

Methods

Participants. Eleven participants (five male, six female; age range 19–37; participant S1 was an author (K.K.)) took part in the experiments described in this study. All participants had normal or corrected-to-normal visual acuity. Informed written consent was obtained from all participants and the experimental protocol was approved by the University of Minnesota Institutional Review Board.

We conducted four experiments. Experiment E1 measured responses to eccentricity stimuli using a high-resolution (7 T, 0.8 mm) gradient-echo protocol. Experiment E2 measured responses to category stimuli also using the high-resolution gradient-echo protocol; data from this experiment are the same as described in a previous publication[4]. Experiment E3 measured responses to the same eccentricity stimuli in E1 but used a spin-echo protocol (7 T, 1.05 mm). Experiment E4 measured responses to the same eccentricity stimuli in E1 but used a low-resolution (3 T, 2.4 mm) gradient-echo protocol.

A total of 16 datasets (scan sessions) were collected: five corresponding to experiment E1; seven corresponding to experiment E2; two corresponding to experiment E3 and two corresponding to experiment E4. To facilitate direct comparison, experiments E3 and E4 were conducted in participants who also participated in experiment E1. A breakdown of participants, experiments and datasets is provided in Supplementary Fig. 1.

Stimulus presentation. For the 7-T datasets, stimuli were presented using a Cambridge Research Systems BOLDscreen 32 LCD monitor positioned at the head of the scanner bed (resolution 1,920 × 1,080 at 120 Hz, viewing distance 189.5 cm). For the 3-T datasets, stimuli were presented using a NEC NP4100 DLP projector that was focused onto a backprojection screen positioned at the head of the scanner bore (resolution 1,024 × 768 at 60 Hz, viewing distance 102 cm). Participants viewed the monitor or backprojection screen via a mirror mounted on the radio frequency coil. A Mac Pro (7 T) or iMac (3 T) computer-controlled stimulus presentation using code based on Psychophysics Toolbox[44,45]. Behavioral responses were recorded using a button box.

Experimental design. In the eccentricity experiment (experiments E1, E3, E4), stimuli consisted of rings positioned at six different eccentricities (that is, distances from the center of gaze), and were confined to a circular region with diameter 11°. Each ring was filled with a black and white contrast pattern that updated at 10 Hz. Ring size scaled with eccentricity, and rings were presented on a neutral gray background (Supplementary Fig. 1). Stimuli were presented in 4-s trials. In a trial, one of the six rings was presented for 3.5 s (35 images presented sequentially, each with duration 0.1 s) and was followed by a brief gap of 0.5 s. Each run lasted 368.116 s and included 12 presentations of each of the six rings as well as blank trials (also of 4-s duration). Throughout stimulus presentation, a small semi-transparent dot (50% opacity) was present at the center of the stimulus. The color of the dot switched between red, white and black every 1–5 s, and participants were instructed to maintain fixation on the dot and to press a button whenever the color changed. A total of nine runs were collected in each 7-T scan session, and a total of six runs were collected in each 3-T scan session. (The eccentricity experiment was time-locked to the refresh rate of the LCD monitor, which caused the additional 116 ms in the total run duration. To compensate for this slight offset, we preprocessed the fMRI data for the eccentricity experiment at a sampling rate of 1.000316 s and then, for simplicity, treated the data in subsequent analyses as if the sampling rate was exactly 1.0 s.)

The category experiment (experiment E2) was the same as the 'functional localizer' experiment conducted in a previous paper[4]. This experiment (http://vpnl.stanford.edu/fLoc/) was developed by the Grill-Spector laboratory[46]. Stimuli consisted of grayscale images of different semantically meaningful categories. There were ten categories, grouped into five stimulus domains: characters (word, number), body parts (body, limb), faces (adult, child), places (corridor, house) and objects (car, instrument). Each stimulus was presented on a scrambled background and occupied a square region with dimensions 10° × 10°. Stimuli were presented in 4-s trials. In a trial, eight images from a given category were presented sequentially, each with duration 0.5 s. Each run lasted 312.0 s and included six presentations of each of the ten categories as well as blank trials (also of 4-s duration). Throughout stimulus presentation, a small red fixation dot was present at the center of the stimulus. Participants were instructed to maintain fixation on the dot and to press a button whenever they noticed an image in which only the background was present ('oddball' task). A total of 10–12 runs were collected in each scan session.

MRI data acquisition and preprocessing. Acquisition and preprocessing procedures are the same as described in a previous paper[4], except for the addition of a spin-echo acquisition protocol. A summary of all procedures is provided next, and we refer the reader to the previous paper for details.

Acquisition. MRI data were collected at the Center for Magnetic Resonance Research at the University of Minnesota. Some data were collected using a 7-T Siemens Magnetom scanner equipped with SC72 body gradients and a custom four-channel-transmit, 32-channel-receive radio frequency head coil. Other data were collected using a 3-T Siemens Prisma scanner and a standard Siemens 32-channel radio frequency head coil. Head motion was mitigated using standard foam padding.

Anatomical data were collected at 3 T at 0.8-mm isotropic resolution. We used a whole-brain T1-weighted MPRAGE sequence (repetition time (TR), 2,400 ms: echo time (TE), 2.22 ms; inversion time (TI) 1,000 ms; flip angle, 8°; bandwidth, 220 Hz per pixel; no partial Fourier, in-plane acceleration factor (iPAT) 2; acquisition time (TA), 6.6 min per scan) and a whole-brain T2-weighted SPACE sequence (TR 3,200 ms, TE 563 ms, bandwidth 744 Hz per pixel, no partial Fourier, in-plane acceleration factor (iPAT) 2, TA 6.0 min per scan). Several T1 and T2 scans were acquired for each participant to increase signal-to-noise ratio.

Functional data for experiments E1 and E2 were collected at 7 T using gradient-echo EPI at 0.8-mm isotropic resolution with partial-brain coverage (84 oblique slices covering occipitotemporal cortex, slice thickness 0.8 mm, slice gap 0 mm, field-of-view 160 mm (frequency encoding (FE)) × 129.6 mm (phase encoding (PE)), phase-encode direction inferior–superior (F ≫ H in Siemens' notation), matrix size 200 × 162, TR 2.2 s, TE 22.4 ms, flip angle 80°, echo spacing 1 ms, bandwidth 1,136 Hz per pixel, partial Fourier 6/8, in-plane acceleration factor (iPAT) 3, multiband slice acceleration factor 2). Gradient-echo fieldmaps were also acquired for post hoc correction of EPI spatial distortion (same slice slab as the EPI data, resolution 2 × 2 × 2.4 mm^3, TR 391 ms, TE1 4.59 ms, TE2 5.61 ms, flip angle 40°, bandwidth 260 Hz per pixel, no partial Fourier, TA 1.3 min). Fieldmaps were periodically acquired over the course of each scan session to track changes in the magnetic field.

Functional data for experiment E3 were collected at 7 T using spin-echo EPI at 1.05-mm isotropic resolution with partial-brain coverage (64 (or 48 for dataset D14) slices, slice thickness 1.05 mm, slice gap 0 mm, field-of-view 128 mm (FE) × 111.2 mm (PE), phase-encode direction inferior–superior (F ≫ H in Siemens' notation; dataset D13 was reversed H ≪ F), matrix size 122 × 106, TR 2.2 s, TE 39 ms, flip angle 90°, echo spacing 1 ms, bandwidth 1,138 Hz per pixel, partial Fourier 6/8, in-plane acceleration factor (iPAT) 2, multiband slice acceleration factor 2). Corresponding gradient-echo fieldmaps were also acquired.

Functional data for experiment E4 were collected at 3 T using gradient-echo EPI at 2.4-mm isotropic resolution with partial-brain coverage (30 slices, slice thickness 2.4 mm, slice gap 0 mm, field-of-view 192 mm (FE) × 192 mm (PE), phase-encode direction anterior–posterior (A ≫ P in Siemens' notation), matrix size 80 × 80, TR 1.1 s, TE 30 ms, flip angle 62°, echo spacing 0.55 ms, bandwidth 2,232 Hz per pixel, no partial Fourier, no in-plane acceleration, multiband slice acceleration factor 2). Corresponding gradient-echo fieldmaps were also acquired.

Preprocessing. T1- and T2-weighted anatomical volumes were corrected for gradient nonlinearities, coregistered and averaged (within modality). The averaged T1 volume (0.8-mm resolution) was processed using FreeSurfer[47] v.6 beta (build-stamp 20161007) with the -hires option. We generated six cortical surfaces spaced equally between 10 and 90% of the distance between the pial surface and the boundary between gray and white matter, increased the density of surface vertices by bisecting each edge and truncated the surfaces retaining only posterior cortex to reduce memory requirements. The resulting surfaces are termed 'depth 1' to 'depth 6' where 1 corresponds to the outermost surface and 6 to the innermost surface. Cortical surface visualizations were generated using nearest-neighbor interpolation of surface vertices onto image pixels.

Functional data were preprocessed by performing one temporal resampling and one spatial resampling. The temporal resampling consisted of one cubic interpolation of each voxel's time-series data; this interpolation corrected differences in slice acquisition times and also upsampled the data to 1.0 s (Supplementary Fig. 1). Data were prepared such that the first time-series data point coincides with the acquisition time of the first slice acquired in the first EPI volume. The motivation for upsampling is to exploit the intrinsic jitter between the data acquisition and the experimental paradigm. The spatial resampling consisted of one cubic interpolation of each volume; this interpolation corrected head motion (rigid-body transformation) and EPI distortion (determined by regularizing the fieldmaps and interpolating them over time) and also mapped the functional volumes onto the cortical surface representations (affine transformation between the EPI data and the averaged T2 volume). Note that the temporal correction is applied first and, then based on the resulting temporally corrected volumes, the spatial correction is applied. Although the topic of optimal ordering is out of scope of the current paper, we believe that performing temporal correction first is most appropriate in datasets where head motion is relatively low.

After preprocessing, the data consisted of EPI time series sampled every 1.0 s at the vertices of the depth-dependent cortical surfaces (depth 1–6). As a final preprocessing step, for the purposes of identifying vertices affected by venous susceptibility effects, we computed the mean of the EPI time-series data obtained for each vertex and divided the EPI intensities by a fitted 3D polynomial (up to degree 4); this produced bias-corrected EPI intensities that can be interpreted as percentages (for example, 0.8 means 80% of the brightness of typical EPI intensities).

GLM analysis. We analyzed the preprocessed time-series data using three different GLM models (FIR, standard, TDM).

The first GLM model, termed FIR, is a GLM in which separate regressors are used to model each timepoint in the response to each experimental condition[48]. Results from this model are used as inputs to the TDM method. The FIR model characterized the response from 0 to 30 s after condition onset, yielding a total of 31 regressors for each condition. (Modeling the response to 30 s was sufficient to capture most of the hemodynamic responses, see Fig. 3 and Supplementary Fig. 3.) We divided the trials for each experimental condition into two groups using a 'condition-split' strategy[4], thereby producing two estimates for each response timecourse. Fitting the FIR model produced BOLD response timecourses (timecourses of betas) with dimensionality N vertices × 6 depths × M conditions × 31 timepoints × 2 condition splits where N is the number of surface vertices for a given participant and M is the number of conditions in the experiment.

The second GLM model, termed standard, is a GLM in which a canonical HRF (getcanonicalhrf.m) is convolved with condition onsets to create a regressor for each experimental condition. We used six condition splits, thereby producing six response estimates for each condition. Fitting the standard model produced BOLD response amplitudes (betas) with dimensionality N vertices × 6 depths × M conditions × 6 condition splits.

The third GLM model, termed TDM, is a GLM in which two hemodynamic timecourses (early, late) are separately convolved with condition onsets to create two regressors for each experimental condition. The exact nature of these timecourses is determined by the TDM method as described next. We used six condition splits, thereby producing six response estimates for each combination of condition and timecourse. Fitting the TDM model produced BOLD response amplitudes (betas) with dimensionality N vertices × 6 depths × M conditions × 2 timecourses × 6 condition splits.

GLMs were prepared and fit to the data using GLMdenoise[49,50]. In GLMdenoise, the GLM consists of experimental regressors (which may take on different forms, as described previously), polynomial regressors that characterize the baseline signal level in each run, and data-derived nuisance regressors. In the case of the FIR model, experimental regressors consisted of binary values (0 and 1 s). In the case of the standard and TDM models, experimental regressors consisted of the convolution of condition onsets (1 s) with hemodynamic timecourses that are normalized to peak at 1 (for example, Supplementary Fig. 2). After fitting the GLMs, estimated betas were converted from raw scanner units to units of percentage BOLD signal change by dividing by the mean signal intensity observed at each vertex and multiplying by 100.

Betas were further analyzed using simple summary metrics. To quantify overall BOLD activity at a given vertex, we calculated mean absolute beta (for example, Fig. 4, left) by averaging betas across condition splits, taking the absolute value of the results, and then averaging across conditions. To summarize responses to the eccentricity stimuli, we calculated peak eccentricity (for example, Fig. 4, middle) by averaging betas across condition splits, performing positive half-wave rectification (that is, setting negative values to zero) and then calculating center of mass[51]. Specifically, center of mass was calculated as the weighted average of the integers 1–6 (corresponding to the six ring eccentricities from fovea to periphery) using the rectified betas as weights.

Timecourse quantification and metrics. In some analyses (Supplementary Figs. 2 and 3), we summarize the typical timecourse shape observed in a set of timecourses. This was accomplished using a PCA-based procedure (derivehrf.m). In the procedure, we first subtract the mean of each timecourse. We then perform PCA on the entire set of timecourses and extract the first PC (this is the timecourse vector along which variance in the total set of timecourses is maximized). Next, we add a constant offset to the first PC such that the first timepoint equals 0, and, if necessary, flip the sign of the PC such that the mean over the range 0–10 s is positive. Finally, we calculate the weight that minimizes squared reconstruction error for each timecourse, compute the absolute value of these weights and then scale the PC by the average weight. The motivation for de-meaning the timecourses before PCA is to suppress low-frequency noise present in weak BOLD responses. In general, the use of PCA for summarizing timecourse shape[52] has advantages over simply computing the mean timecourse: PCA handles negative BOLD timecourses, and PCA allows timecourses with larger BOLD responses to have greater influence on the resulting timecourse shape (thereby producing more robust results).

Several timecourse metrics were computed (Supplementary Figs. 2 and 3 and Fig. 3). Given a timecourse, we upsampled the timecourse to a sampling rate of 0.01 s using sinc interpolation. We then identified the maximum of the resulting timecourse (peak amplitude) and its associated time (time-to-peak). We used linear interpolation to calculate the time at which the timecourse rises to half of the maximum value (rise time) and the time at which the timecourse falls to half of the maximum value (fall time). Finally, we computed the time elapsed between the rise time and the fall time (full-width-at-half-maximum).

TDM method. *Theory.* TDM is a data-driven technique that identifies a principal axis of timecourse variation present in a set of experimentally measured timecourses. It does this by examining timecourses projected into a low-dimensional space defined by the first three PCs of the timecourses and extracting a one-dimensional manifold—specifically, an arc on the unit sphere—that captures the variation of interest. The procedure can be viewed as a powerful method for summarizing and extracting the signal present in timecourses that, considered individually (that is, one response timecourse at a time), would likely be insufficiently reliable. In our fMRI measurements of responses to 4-s visual stimuli, we consistently find that one endpoint of the line corresponds to an early timecourse peaking at around 5–7 s and the other endpoint of the line corresponds to a late timecourse peaking at around 6–9 s (Fig. 3a and Supplementary Figs. 2 and 3). These timecourses are interpreted as reflecting hemodynamic responses from the microvasculature (capillaries and venules) and hemodynamic responses from the macrovasculature (veins), respectively. TDM then uses the identified timecourses in a regression model to decompose observed hemodynamic responses into early and late components. The researcher can choose to analyze further the early component, the late component or both.

There are three main quantities involved in the TDM technique: density, referring to the timecourse shapes that tend to be present in the data; vector length, referring to the amplitudes of the timecourses in the data and EPI intensity, referring to the bias-corrected EPI intensity of the vertex (or voxel) to which each timecourse belongs. (Note that these three quantities are distinct from the three PC dimensions.) TDM combines density and vector length and fits an oriented 2D Gaussian to the result to identify the one-dimensional arc. The motivation for incorporating vector length beyond density alone is to ensure that veins—which generate BOLD responses with large amplitudes but constitute only a fraction of the total set of responses—have sufficient influence on the determination of the arc. Note that EPI intensity does not directly participate in the determination of the arc, and can therefore provide useful validation of the results (Fig. 2a and Extended Data Figs. 1 and 2).

There are a few important conceptual points regarding the nature of the TDM method. For any given voxel (or vertex), the BOLD response to an experimental event is expected to reflect a mixture of early (microvasculature) and late (macrovasculature) timecourses. The specific proportion of these timecourses is expected to vary from voxel to voxel simply due to heterogeneity in the spatial structure of the vasculature (for example, one voxel might be centered on a large vein, whereas another voxel may only partially overlap the vein). Different proportions of the timecourses manifest in TDM as different points, and these points collectively trace out an arc on the unit sphere (Fig. 1). Empirically, we confirm that a diversity of proportions are observed in response timecourses (Extended Data Fig. 5).

Another important point is that TDM is not equivalent to estimating a different HRF for each voxel[52–54]. Using a single hemodynamic timecourse for different experimental conditions (time-condition separability) yields at most one amplitude estimate (beta) for each condition. In contrast, TDM allows experimental conditions to have different loadings on the early and late timecourses and yields two amplitude estimates (betas) for each condition (this critical feature is elaborated in Extended Data Fig. 4).

Finally, note that the GLM analyses performed in TDM rely on the assumption of linear summation of BOLD responses over time. Our experiments, like many used in cognitive neuroscience, involve a large number of trials that are presented fairly rapidly to maximize statistical power (for example, less than 10 s of rest in between trials). In such experiments, it is a practical necessity to assume temporal linearity. Moreover, nonlinear effects are likely to average out when using randomized experimental designs. Nevertheless, the accuracy with which TDM identifies timecourses may be limited if there exist nonlinear effects[55] and especially if these nonlinearities vary for different types of vasculature[27,56,57].

Prerequisites. The TDM method requires a task-based experiment in which neural events occur at prescribed times, and is therefore inapplicable to resting-state paradigms, at least in its current form. Despite this, our approach is still valuable to the wide array of task-based fMRI studies being conducted in the field of cognitive neuroscience. We suspect that TDM will be most effective for event-related paradigms where experimental events are short (for example, 4 s or less). Block designs involving prolonged events (for example, 16–30 s) or designs involving continuously changing experimental parameters (for example, sinusoidal variation of a stimulus property) are likely to generate microvasculature-related and macrovascular-related timecourses that are more similar and therefore harder to disambiguate.

In terms of data acquisition, the TDM method is likely compatible with a broad range of acquisition styles. As we show, TDM can be applied to data from standard spatial resolutions (2–3 mm, Extended Data Figs. 2 and 3) or data from high spatial resolutions (<1 mm, Fig. 4). Presumably, the spatial resolution needs only to be high enough to allow diverse sampling of vasculature in the brain. With respect to temporal resolution, the temporal requirements of TDM are not stringent: we show that TDM can be successfully applied to data acquired at even fairly slow rates, such as the 2.2-s sampling rate used in datasets D1–D14. This is likely aided by the fact that we jittered the acquired timepoints with respect to the experimental conditions (Supplementary Fig. 1).

Because TDM is a data-driven technique, it is necessary to acquire sufficient data to support the method. For example, there must be sufficient data to estimate response timecourses from the voxels in a dataset. If a given dataset is overly noisy or if not enough data are collected, timecourse estimates may be noisy and nearly

isotropic in their distributions in the 3D PCA space (for example, dataset D6 in Extended Data Fig. 1), making it difficult to extract the underlying structure of the data.

Algorithm. The TDM algorithm starts with a set of timecourses and determines a pair of timecourses that characterize the overall variation in the timecourses. The following are the steps in the TDM algorithm (extracthrfmanifold.m):

1. Perform PCA on the timecourses. We collect timecourses into a 2D matrix of dimensionality L timecourses × T timepoints, and then perform singular value decomposition. This produces a matrix with dimensionality T timepoints × T singular vectors where columns correspond to PC timecourses in decreasing order of variance explained. Since the sign of the returned singular vectors is arbitrary, we flip the sign of the first PC, if necessary, to ensure that the mean of the timecourse is positive over the range 0–10 s. Note that previous studies have applied PCA to fMRI response timecourses but in different contexts[41,58].
2. Use PC1–PC3 to define a 3D space for further analysis. Our convention for visualization is that PC1 points out of the page (positive z axis), PC2 points to the right (positive x axis) and PC3 points to the top (positive y axis).
3. Map timecourses onto the unit sphere. We project each timecourse onto PC1, PC2 and PC3. One complication here is the possibility of negative BOLD timecourses. To first approximation, such timecourses can be treated as a sign-flipped version of positive hemodynamic responses[59]. Thus, we take the coordinates of each timecourse and mirror these coordinates across the origin, if necessary, to ensure that the loading on PC1 is positive. In this way, negative BOLD timecourses are flipped and treated in the same way as positive BOLD timecourses. (If the user wishes to simply discard negative BOLD timecourses, this can be achieved by setting opt.ignorenegative to 1.) After projection, each timecourse is represented by a set of coordinates (loadings) and can be interpreted as a 3D vector. We normalize each vector to unit length (thereby placing the vector on the unit sphere), and also save the original vector length for later use.
4. Calculate a 2D image that represents density. We orthographically project timecourses onto the xy plane, and then calculate a 2D histogram. This produces a 2D image where pixel values represent frequency counts. This image indicates typical timecourse shapes found in the data.
5. Calculate 2D images that represent vector length and EPI intensity. Using the same orthographic projection and binning scheme of step 4, we calculate the median vector length of the timecourses found in each bin. We also calculate the median bias-corrected EPI intensity of the voxels (vertices) associated with the timecourses in each bin. This produces 2D images where pixel values represent vector lengths (indicating timecourse amplitudes) and EPI intensities (indicating static susceptibility effects caused by veins), respectively.
6. Regularize the density image by subtracting a baseline bias. We distribute a collection of particles on the unit sphere (S2 Sampling Toolbox, https://www.github.com/AntonSemechko/S2-Sampling-Toolbox/, assign timecourses to their nearest particles and count the number of timecourses associated with each particle. We then calculate a histogram of these counts and determine the bin B with the highest frequency. Finally, we stochastically subsample the timecourses such that the number of timecourses associated with each particle is reduced by the middle value of bin B. The resulting subsampled timecourses are used to generate a new density image.
7. Scale the density image. The regularized density image from step 6 is scaled such that 0 maps to 0 and the maximum value maps to 1. Values are then truncated to the range [0,1].
8. Regularize the vector-length image by subtracting a baseline bias. This is accomplished in a similar manner as step 6: we calculate a histogram of the values in the vector-length image, determine the bin B with the highest frequency, and subtract the middle value of bin B from all image pixels.
9. Scale the vector-length image. The regularized vector-length image from step 8 is scaled such that 0 maps to 0 and the maximum value maps to 1. Values are then truncated to the range [0,1].
10. Average the density and vector-length images. Although the default behavior is to simply average the density and vector-length images, if the user desires a different weighting (for example, giving more weight to the vector-length image), a flag can be used (opt.vlengthweight).
11. Fit 2D Gaussian. The image resulting from step 10 is fit with a 2D Gaussian. This choice of model is simplistic but sufficient; for more complex distributions, one might consider the use of principal curves[60]. The Gaussian is controlled by two parameters specifying the center, two parameters specifying the spreads along the major and minor axes, a rotation parameter, a gain parameter and an offset parameter. In model fitting, the error metric is set up such that the image is interpreted as a probability distribution (pixels with larger values reflect higher density and thus contribute more heavily to the error metric).
12. Extract two points along the major axis of the Gaussian. We determine points corresponding to the mean plus or minus one standard deviation along the major axis of the Gaussian. The choice of one standard deviation is arbitrary but appears to produce satisfactory results.
13. Reconstruct timecourses corresponding to the identified points. We place the points determined in step 12 on the unit sphere, and use their associated coordinates to weight and sum the PC1, PC2 and PC3 timecourses. This yields two reconstructed timecourses. Based on time-to-peak, we label one timecourse as 'early' and the other timecourse as 'late'.

While the algorithm described here has several parameters, default parameter values were used for all datasets in this paper. Some of the parameters are relatively minor and likely do not need adjustment. These include parameters related to constructing histograms and regularizing the images (opt.numspherehistbins, opt.bins, opt.spheresparticles) and a parameter for the temporal range over which to quantify timecourse sign (opt.rng). Other parameters are more important, and may warrant adjustment. These include the relative weighting of the density and vector-length images (opt.vlengthweight) and the choice of one standard deviation for early and late timecourses (results.fullarc can be used to choose other points along the arc for the timecourses).

Application of algorithm. In our datasets, we obtained timecourses by fitting an FIR model to the fMRI time-series data. We then calculated the amount of variance explained (R^2) by the FIR model. To focus the TDM algorithm on cortical locations with BOLD responses, we selected all vertices within a given region of interest (ROI) that exceeded an automatically determined threshold (specifically, the value at which the posterior probability switches between two Gaussian distributions fitted as a mixture model to the data; see findtailthreshold.m). This produced a set of timecourses with dimensionality P vertices × M conditions × 31 timepoints × 2 condition splits. We applied the TDM algorithm to the timecourses averaged across the condition splits (thus, reflecting the entire dataset) and also to the timecourses from each condition split separately to assess reliability. In both cases, the number of timecourses given to the TDM algorithm is $L = P \times M$ and the number of timepoints is $T = 31$. After completion of the TDM algorithm, the identified timecourses (early, late) were incorporated into a GLM model to decompose the fMRI time-series data into early and late components (see GLM analysis).

ROI definition. We used two ROIs. For the eccentricity experiment, we used the union of visual areas V1, V2 and V3 from a publicly available atlas of visual topography[61]. For the category experiment, we used a manually defined region in occipital, parietal and temporal cortex that covers visually responsive vertices (same region used in ref. [4]). Both ROIs were defined in FreeSurfer's fsaverage space and backprojected to individual participants for the purposes of vertex selection.

Line-profile analysis. A line-profile analysis[4] was used to quantify and summarize activity patterns observed in the eccentricity experiment. To define a set of lines, we downloaded the publicly available HCP 7 T Retinotopy Dataset[62], took the group-average results for population receptive field angle and eccentricity, and visualized these results on FreeSurfer's fsaverage sphere surface. We then manually selected on each hemisphere 5 eccentricity × 3 angle = 15 vertices in primary visual cortex (V1) corresponding to the intersection of the vertical and horizontal meridians and eccentricity values 0.3°, 0.8°, 1.8°, 3.3° and 6.4° (these values are approximately equally spaced along the cortical surface). These fsaverage vertices were mapped to participant-native surfaces via nearest-neighbor interpolation, and used to draw lines corresponding to iso-angle contours (lines are drawn on an orthographic projection of the sphere surface). Each line consists of four line segments (one for each successive pair of eccentricity values) and is represented as a sequence of surface vertices. To sample the full extent of V1, we used linear interpolation to create four lines evenly spaced between the vertical and horizontal meridians, yielding a total of 1 (lower vertical meridian) + 4 + 1 (horizontal meridian) + 4 + 1 (upper vertical meridian) = 11 lines representing iso-angle contours in each hemisphere of each participant (Fig. 4). Cortical distance was quantified using Euclidean distance between pairs of vertices on participant-native white surfaces.

We used the defined lines to generate group-level profiles of BOLD activity (Fig. 5). The core idea is to treat the vertices corresponding to the five selected eccentricity values as waypoints for the purposes of interparticipant alignment. First, to determine approximate physical units, we computed the distance between successive pairs of waypoints and averaged the resulting values across lines and participants. This yielded a sequence of averaged distances in millimeter units. Then, for each participant, the sequence of vertices between successive pairs of waypoints was linearly rescaled to match the averaged distances. Finally, beta weights from a given analysis of interest (for example, standard) were extracted for each vertex and regridded onto an evenly spaced 0.3-mm grid using linear interpolation. This process ultimately produced sets of activity profiles in V1 that extend from 0.3° to 6.4° eccentricity and that are directly comparable across lines, participants and analyses.

To quantify the similarity of different analysis approaches (Fig. 5b), we computed, for split halves of each dataset, a full set of activity profiles for all lines, stimulus conditions, depths and hemispheres. These activity profiles were then correlated across each pair of analysis approaches, averaging across resampling cases. (For example, activity profiles from standard on split 1 were correlated with activity profiles from late on split 2, activity profiles from late on split 1 were correlated with activity profiles from standard on split 2 and the two resulting

correlation values were averaged.) The use of split halves for the similarity analysis is valuable as it provides a measure of reliability for each analysis approach.

Statistics. Reliability of results was assessed using both within-session and across-session analyses, and is depicted by error bars in the various figures. In the within-session case, reliability was assessed by splitting experimental trials into nonoverlapping groups (condition splits), analyzing the groups separately, and then quantifying variability of results across the groups. For the FIR GLM model, two condition splits were used; for the standard and TDM GLM models, six condition splits were used. In the across-session case, reliability was assessed by simply quantifying variability of results across sessions. TDM results were replicated in 16 scan sessions conducted in 11 distinct participants. For some analyses, robustness of BOLD response amplitudes was quantified using t values (mean divided by standard error across condition splits). Effect sizes are expressed using percentage BOLD signal change and Pearson's correlation. Cross-validation was used to assess accuracy of modeling procedures (Fig. 2d and Supplementary Fig. 6c). Monte Carlo simulations were used to formally assess the robustness of analysis procedures (Supplementary Figs. 5 and 7).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Materials related to this paper, including all datasets used, are available at https://osf.io/j2wsc/. Raw data in BIDS format[63] are hosted at OpenNeuro at https://doi.org/10.18112/openneuro.ds002702.v1.0.1, whereas preprocessed data (that is, temporally and spatially corrected fMRI time-series data in surface format) are provided on the OSF site.

Code availability

Data were primarily analyzed using custom code written in MATLAB R2018a. The OSF site (https://osf.io/j2wsc/) includes an archive of the code used in this paper, sample data and scripts demonstrating the TDM method, and a link to a detailed video tutorial demonstrating the scripts and discussing the methodology and rationale therein. TDM source code is licensed under the BSD 3-Clause License, and is available at https://github.com/kendrickkay/TDM/ and on the executable platform Code Ocean (https://doi.org/10.24433/CO.4779366.v1)[64].

References

44. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
45. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).
46. Stigliani, A., Weiner, K. S. & Grill-Spector, K. Temporal processing capacity in high-level visual cortex is domain specific. *J. Neurosci.* **35**, 12412–12424 (2015).
47. Fischl, B. FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
48. Dale, A. M. Optimal experimental design for event-related fMRI. *Hum. Brain Mapp.* **8**, 109–114 (1999).
49. Charest, I., Kriegeskorte, N. & Kay, K. N. GLMdenoise improves multivariate pattern analysis of fMRI data. *NeuroImage* **183**, 606–616 (2018).
50. Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F. & Wandell, B. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* **7**, 247 (2013).
51. Hansen, K. A., Kay, K. N. & Gallant, J. L. Topographic organization in and near human visual area V4. *J. Neurosci.* **27**, 11896–11911 (2007).
52. Kay, K. N., David, S. V., Prenger, R. J., Hansen, K. A. & Gallant, J. L. Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. *Hum. Brain Mapp.* **29**, 142–156 (2008).
53. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
54. Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B. & Gramfort, A. Data-driven HRF estimation for encoding and decoding models. *NeuroImage* **104**, 209–220 (2015).
55. Friston, K. J., Josephs, O., Rees, G. & Turner, R. Nonlinear event-related responses in fMRI. *Magn. Reson. Med.* **39**, 41–52 (1998).
56. Thompson, S. K., Engel, S. A. & Olman, C. A. Larger neural responses produce BOLD signals that begin earlier in time. *Front. Neurosci.* **8**, 159 (2014).
57. Zhang, N., Yacoub, E., Zhu, X.-H., Ugurbil, K. & Chen, W. Linearity of blood-oxygenation-level dependent signal at microvasculature. *NeuroImage* **48**, 313–318 (2009).
58. d'Avossa, G., Shulman, G. L. & Corbetta, M. Identification of cerebral networks by classification of the shape of BOLD responses. *J. Neurophysiol.* **90**, 360–371 (2003).
59. Shmuel, A., Augath, M., Oeltermann, A. & Logothetis, N. K. Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1. *Nat. Neurosci.* **9**, 569–577 (2006).
60. Hastie, T. & Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **84**, 502–516 (1989).
61. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic maps of visual topography in human cortex. *Cereb. Cortex* **25**, 3911–3931 (2015).
62. Benson, N. C. et al. The human connectome project 7 tesla retinotopy dataset: description and population receptive field analysis. *J. Vis.* 18, https://doi.org/10.1167/18.13.23 (2018).
63. Gorgolewski, K. J. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3**, 1–9 (2016).
64. Kay, K. Temporal decomposition through manifold fitting. *Code Ocean* https://doi.org/10.24433/CO.4779366.v1 (2020).

Acknowledgements

We thank E. Margalit and N. Petridou for helpful discussions and L. Dowdle for assistance with preparing data in BIDS format. This work was supported by National Institutes of Health grant nos. P41 EB015894 (K.U.), P41 EB027061 (K.U.), P30 NS076408 (K.U.), S10 RR026783 (K.U.), S10 OD017974-01 (K.U.) and U01 EB025144 (K.U.), and the W. M. Keck Foundation (K.U.).

Author contributions

K.K. designed the experiment. R.-Y.Z. collected the data. K.K. and K.W.J. developed techniques and analyzed the data. K.K. wrote the paper. K.K., R.-Y.Z., K.W.J. and K.U. discussed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

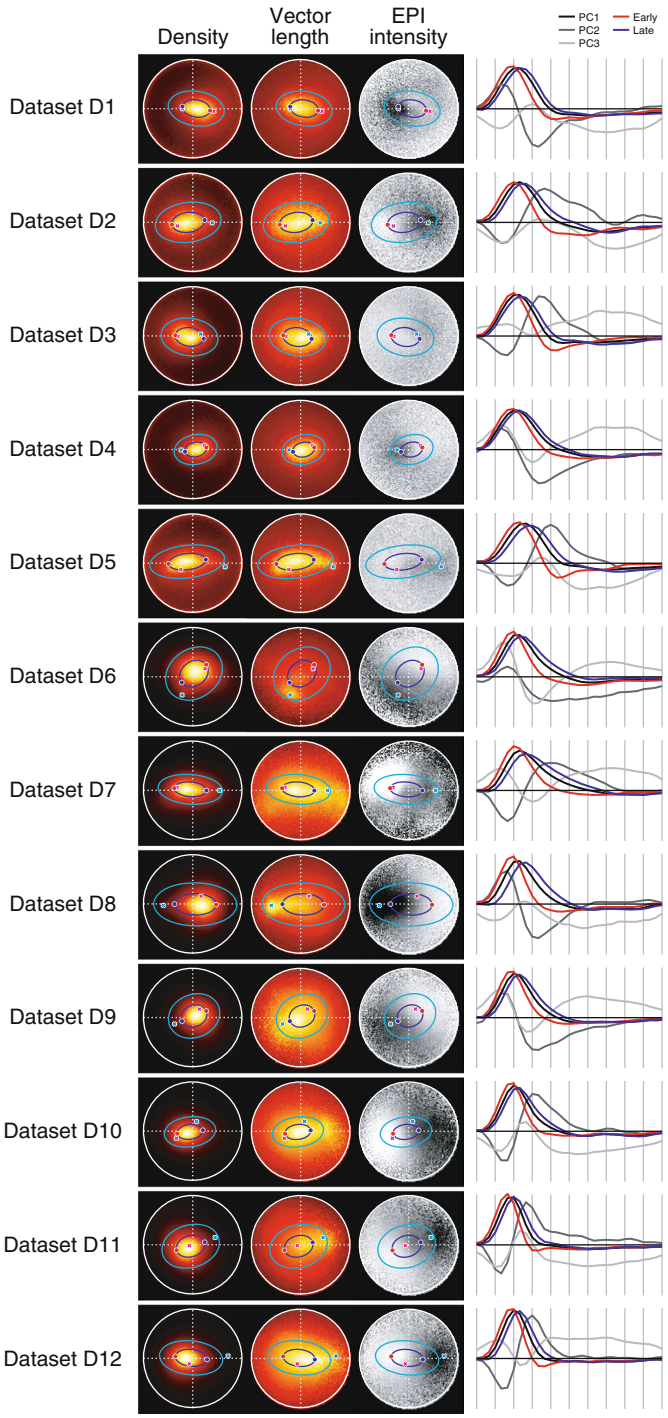
Extended data is available for this paper at https://doi.org/10.1038/s41592-020-0941-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-020-0941-6.

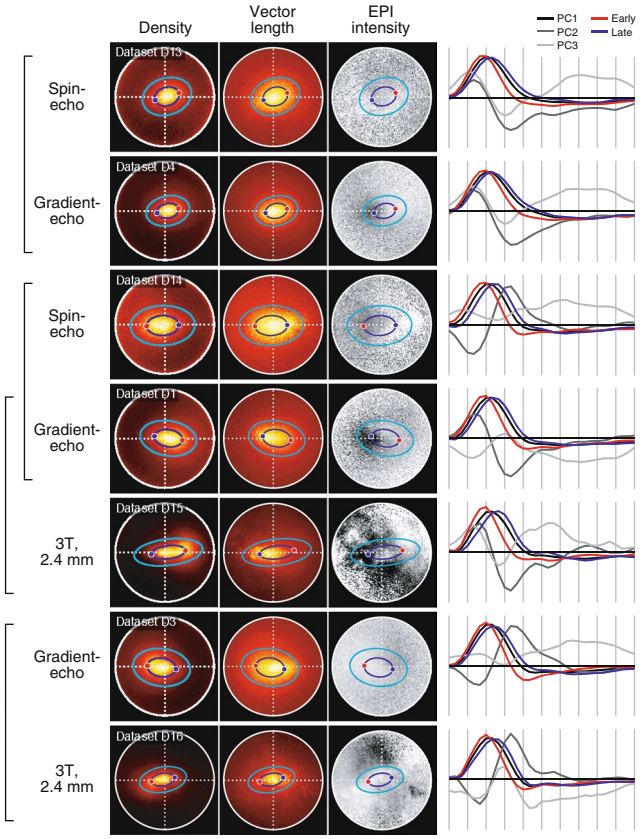
Correspondence and requests for materials should be addressed to K.K.

Peer review information Nina Vogt was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

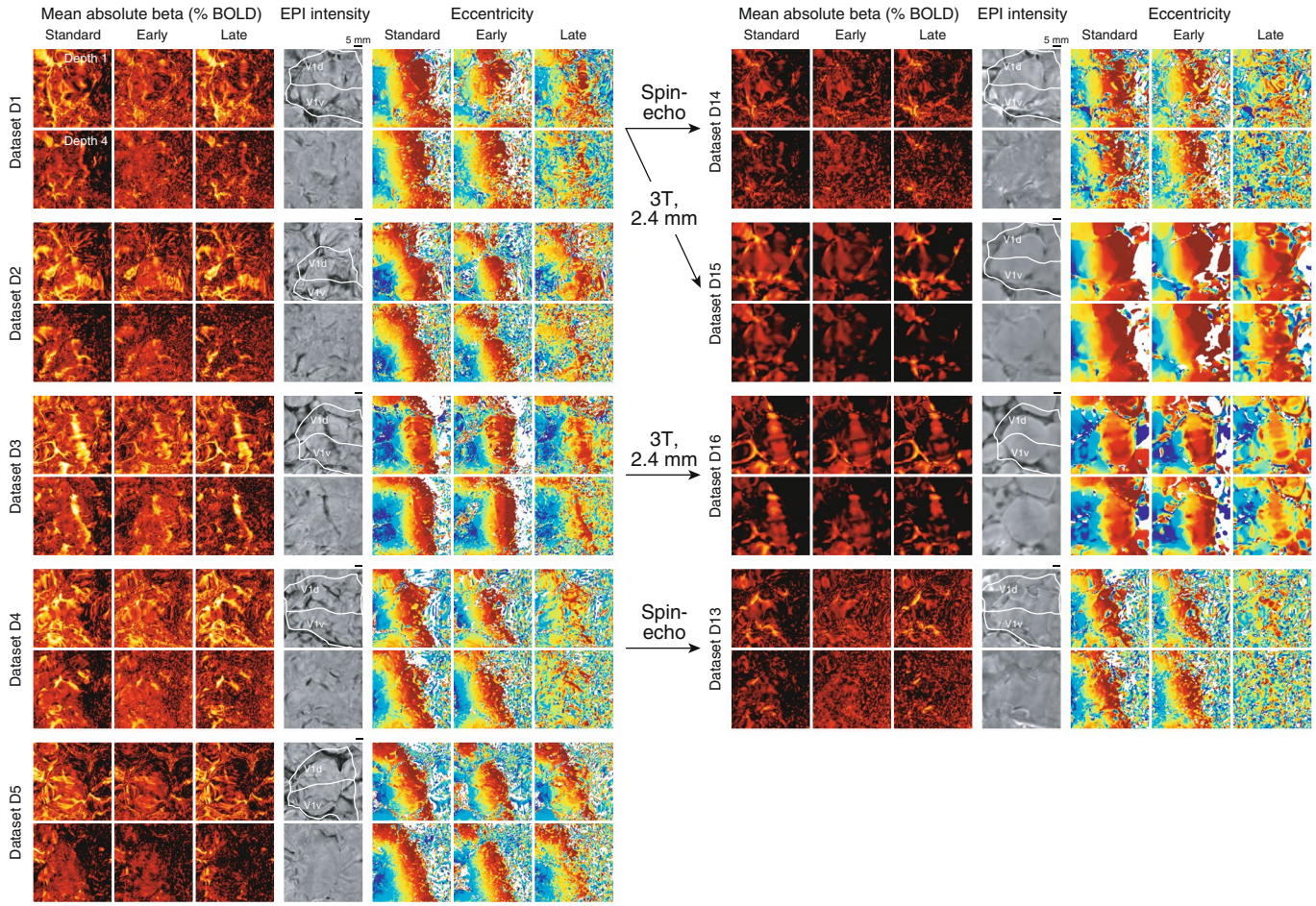
Reprints and permissions information is available at www.nature.com/reprints.



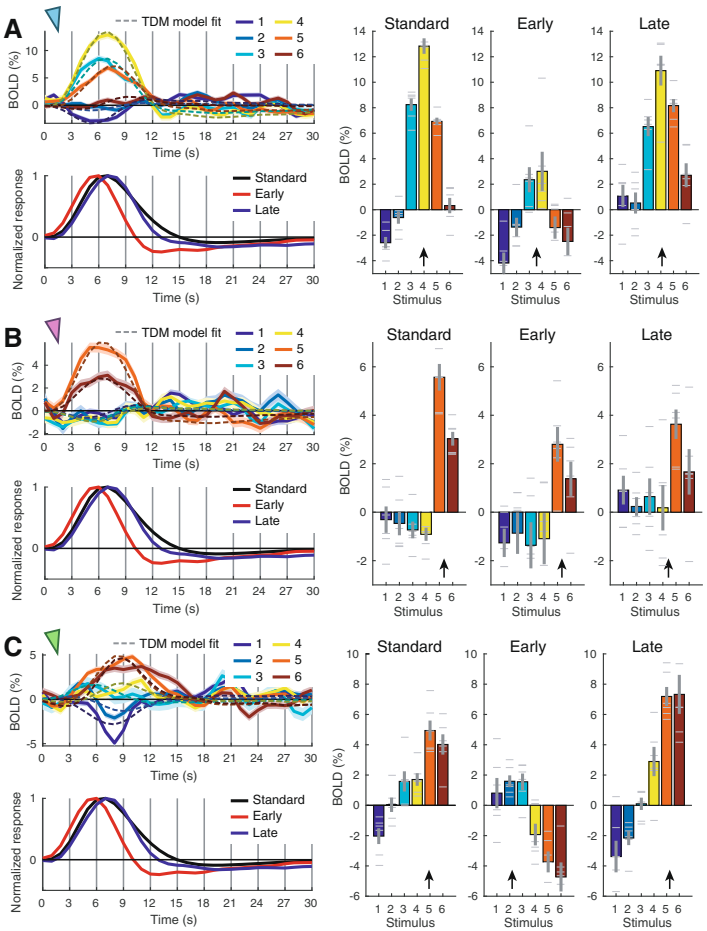
Extended Data Fig. 1 | TDM results for the high-resolution gradient-echo datasets (D1–D12). Magenta and cyan crosses indicate the early and late timecourses derived from the ICA-based procedure. The ICA-based procedure yields timecourses similar to TDM in some datasets (for example D4), but diverges substantially in others (for example D8).



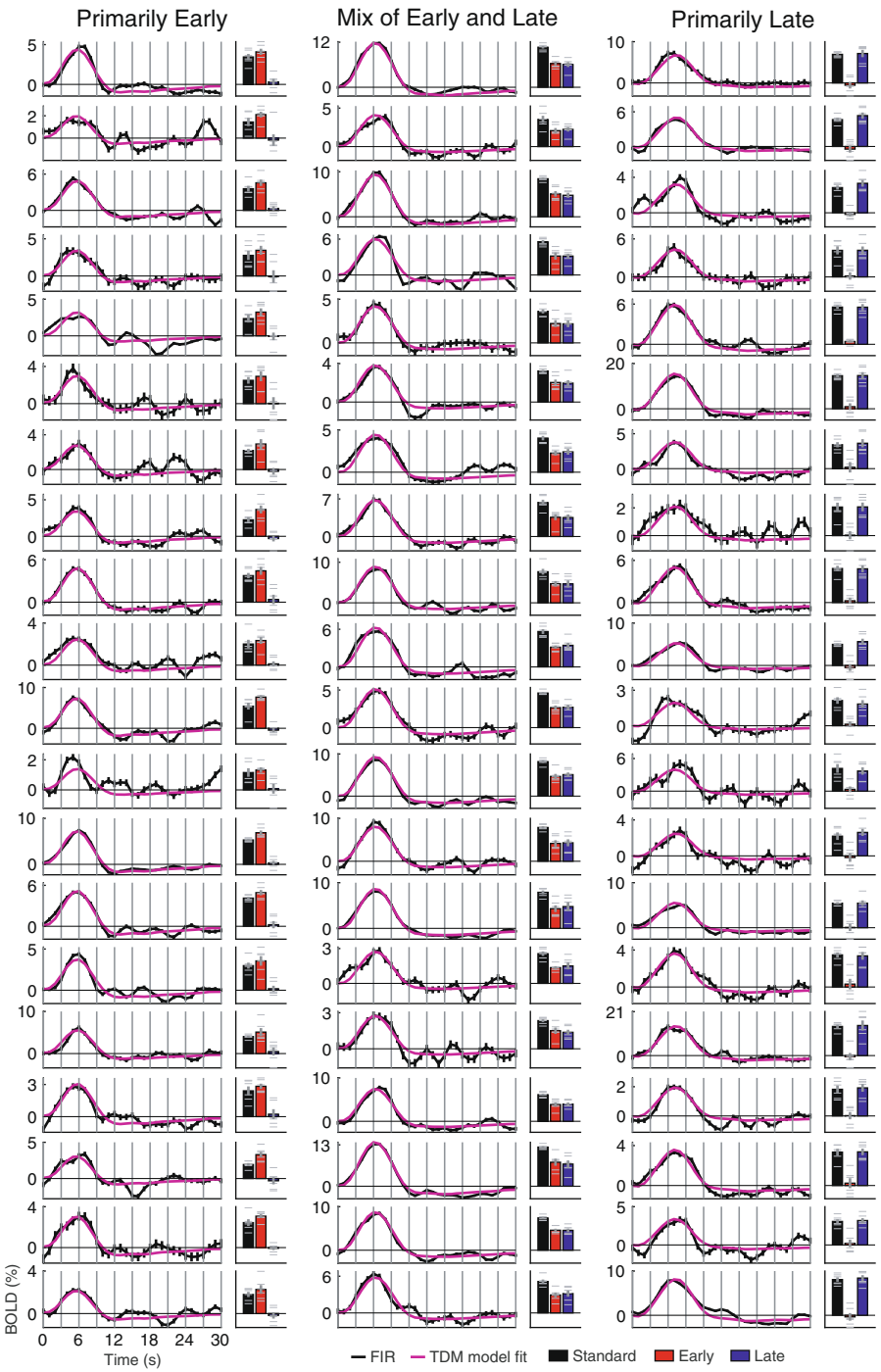
Extended Data Fig. 2 | TDM results for the alternative acquisition protocols (D13–D16). To facilitate comparison, results obtained using the spin-echo and low-resolution protocols are placed next to results obtained using the high-resolution gradient-echo protocol.



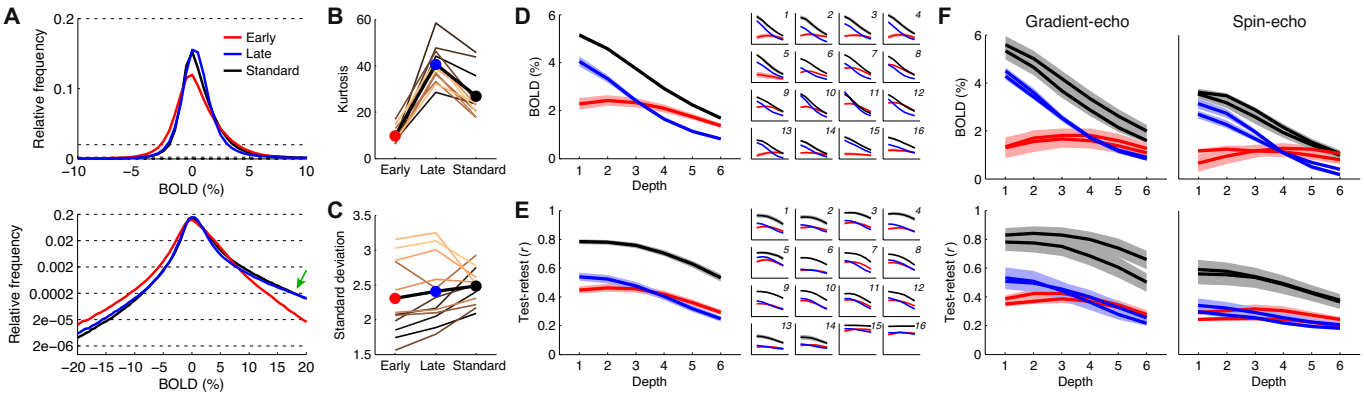
Extended Data Fig. 3 | Decomposition of brain activity patterns across datasets and acquisition protocols. On the left are results obtained using high-resolution (0.8-mm) 7T gradient-echo (Datasets D1–D5). On the right are results obtained using high-resolution (1.05-mm) 7T spin-echo (Datasets D13–D14) and low-resolution (2.4-mm) 3T gradient-echo (Dataset D15–D16). These alternative acquisition protocols were conducted in the same subjects as the high-resolution gradient-echo protocol (correspondence indicated by arrows).



Extended Data Fig. 4 | Detailed inspection of example vertices. ***a–c***, Results for three surface vertices marked by arrows in Fig. 4. At the upper left are FIR timecourses with ribbon center and width indicating mean and standard error across two condition-splits. Dotted lines indicate the overall fit of the TDM model for each condition (reflecting a weighted sum of the Early and Late timecourses). At the lower left are canonical and TDM-derived timecourses. On the right are the three versions of the betas with bars and error bars indicating mean and standard error across six condition-splits and black arrows indicating peak eccentricity. Rainbow colors indicate stimulus eccentricity (1 = most foveal, 6 = most peripheral).



Extended Data Fig. 5 | Response timecourses exhibit diverse proportions of early and late timecourses. Each subplot depicts results for a single condition at a single vertex (Dataset D1). The left shows FIR timecourses (black, with lines and error bars indicating mean and standard error across two condition-splits) and the overall fit of the TDM model (purple). The right shows beta estimates (bars and error bars indicate mean and standard error across six condition-splits). To select which cases to show, we first identified vertices whose R^2 under the TDM GLM is greater than 10%. We then examined the estimated betas and calculated their *t*-values (beta divided by standard error across condition-splits). We determined (i) all cases with a robust Early beta ($t > 5$) and a weak Late beta (absolute value less than 1/10 of the Early beta), (ii) all cases with robust Early and Late betas ($t > 5$) and where each beta is at least 9/10 of the other beta, and (iii) all cases with a robust Late beta ($t > 5$) and a weak Early beta (absolute value less than 1/10 of the Late beta). Finally, we randomly selected 20 cases from each of the three groups.



Extended Data Fig. 6 | Quantitative assessment of BOLD amplitude estimates provided by TDM. a, Histogram. The top plot shows distributions of BOLD amplitudes aggregated across Datasets D1–D12; the bottom plot shows results on a log scale and with a wider *x*-axis range. **b**, Kurtosis. Results are shown for individual datasets (thin lines, D1–D12) and the group average (thick black line). **c**, Standard deviation. **d**, Cortical depth profiles. The main plot shows the average depth profile observed in Datasets D1–D12, with ribbons indicating standard error across datasets; the inset plots show results for individual datasets (D1–D16), with ribbons indicating standard error across conditions. **e**, Reliability. Average correlation of betas across 6 splits of each dataset. **f**, Gradient-echo versus spin-echo. We re-plot results from panels D and E, directly comparing the gradient-echo and spin-echo datasets.

nature research

Corresponding author(s): Kendrick Kay

Last updated by author(s): Jun 24, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Psychophysics Toolbox 3.0.12, MATLAB R2018a
Data analysis	MATLAB R2018a, FreeSurfer 6, GLMdenoise 1.4, TDM 1.0, FastICA Toolbox 2.5, S2 Sampling Toolbox 1.0, SPM5, FSL 5.0.7. Analysis code has been made freely available at https://osf.io/j2wsc/ and https://github.com/kendrickkay/TDM/.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw and pre-processed data from the paper have been made freely available at OpenNeuro (https://openneuro.org/datasets/ds002702/versions/1.0.1) and an OSF site (https://osf.io/j2wsc).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All analyses are conducted at the within-subject level in order to demonstrate the precision and robustness of the demonstrated method. (Additional analyses that summarize across subjects are also performed.) No formal method was used to predetermine sample size. Sample size was chosen (12 main datasets conducted in 11 distinct subjects) such that researchers can be confident that the method works reliably in every individual subject.
Data exclusions	No data were excluded from the analyses, except for a few fMRI runs (~5 min in length) in which subjects were not performing (asleep).
Replication	Experimental findings were demonstrated (and therefore replicated) in 12 individual datasets (11 unique subjects).
Randomization	Different experiments were run on different subjects. No rationale was used to assign subjects to experiments, other than convenience, availability, and consent of subjects. We have no reason to believe that the low-level effects studied in this manuscript (fMRI responses, vasculature) would be idiosyncratic to healthy adult subjects, aside from quantitative variation in the spatial and temporal characteristics of the fMRI responses.
Blinding	Blinding is not relevant to this study, as the subjects did not participate in different groupings. Each individual subject experienced the relevant manipulation (presentation of stimuli of different types); thus, there is little that the investigators could have done to bias the nature of the recorded data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	A mixture of males and females were used (5 males, 6 females; age 19–37). All findings are shown in individual subjects; thus, population variability is not a concern.
Recruitment	Participants were recruited through word of mouth and local advertisements. Given the local population of academic researchers, there is a potential bias for recruited educated subjects interested in cognitive research. This is unlikely to affect the phenomenon being investigated, namely, vascular effects in basic sensory brain responses.
Ethics oversight	University of Minnesota Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	Task, event-related
Design specifications	Each of the main scan sessions involved 4-s trials, between 6-10 distinct conditions, and between 60-108 trials for each experimental condition. Trials were presented with no gaps, aside from occasional blank trials (of length 4 s).
Behavioral performance measures	Behavior was not analyzed in this study.

Acquisition

Imaging type(s)	Functional, structural
Field strength	3T and 7T
Sequence & imaging parameters	The primary sequence was gradient-echo EPI, FOV 160 mm x 129.6 mm, matrix size 200 x 162, slice thickness 0.8 mm, orientation oblique, TR 2.2 s, TE 22.4 ms, flip angle 80°.
Area of acquisition	Partial brain scan focused on visual cortex

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	A combination of custom software and FreeSurfer 6
Normalization	Most analyses used no normalization: effects are studied at the level of the individual. In one analysis (line-profile analysis), we use FreeSurfer's fsaverage space to summarize/combine data across subjects in V1.
Normalization template	FreeSurfer's fsaverage space was used; this was used for the purposes of defining a generic region-of-interest and for the line-profile analysis. All other analyses operated in subject-native space.
Noise and artifact removal	The data-driven analysis method GLMdenoise was used; this method can account for a variety of sources of noise (e.g. physiological, motion, scanner artifacts).
Volume censoring	No censoring was performed.

Statistical modeling & inference

Model type and settings	fMRI response amplitudes are estimated for individual subjects, and error is assessed across trials (within subjects). In some analyses, group-level summaries are created, with error assessed across subjects.
Effect(s) tested	Maps of visual organization are generated based on measured fMRI signals. Such maps typically interpreted via visualization.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	All voxels were analyzed. No statistical thresholding is used.
Correction	Not applicable.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis

nature research

https://doi.org/10.1038/s41592-020-0941-6

Supplementary information

A temporal decomposition method for identifying venous effects in task-based fMRI

In the format provided by the authors and unedited

Supplementary Note 1: Systematic variation in timecourse amplitude, delay, and width

To investigate BOLD timecourse characteristics, we analyzed the time-series data in each dataset using a finite impulse response model (0–30 s, 31 time points). This produced, for each surface vertex, an estimate of the BOLD response timecourse to each experimental condition. We then binned these timecourses either with respect to cortical depth or with respect to EPI intensity. For each bin, we summarized the timecourses found in that bin using a PCA-based procedure (see Methods for details).

Results for a representative dataset are shown in **Supplementary Figure 2**. Proceeding from inner cortical depths (Depth 6) to outer cortical depths (Depth 1), we observe an increase in timecourse amplitude and an increase in timecourse delay (panel A). Proceeding from darker EPI intensities (0–0.5) to lighter EPI intensities (>1), we again observe increases in amplitude and delay but also a small increase in timecourse width (panel B). Summarizing results across different datasets (reflecting different subjects and experiments), we see that these effects are consistently observed (panel D). Notice that although different datasets show similar patterns in relative timing (e.g., time-to-peak is longer for low EPI intensity than for high EPI intensity), there is large variance in absolute timing across datasets (e.g., time-to-peak is approximately 8 s in one dataset but 6 s in another dataset). This is consistent with well-established observations of variability of hemodynamic response functions across subjects[20,21].

Notice that the variations in amplitude (panel D, top row) and the variations in timecourse delays (panel D, middle rows) are more pronounced when binning by EPI intensity (panel D, middle column) than when binning by cortical depth (panel D, left column). We suggest that the underlying source of these effects consists in the high-amplitude, delayed BOLD responses carried by macroscopic veins. Binning by EPI intensity provides a relatively direct proxy for where these venous effects occur[1,4,18]. In contrast, binning by cortical depth provides a less direct proxy for these effects (e.g., pial veins affect outer cortical depths more than inner cortical depths). Thus, binning by EPI intensity will tend to accentuate and highlight the amplitude and delay effects.

The fact that veins carry delayed BOLD responses has been previously shown[14–17]. Prior studies have also demonstrated increased temporal delays at superficial cortical depths[15,17]. Thus, the observations we make here are not novel, but provide reassurance that these effects can be reproduced in our data and establish a starting point for the development of the TDM method.

The Early and Late timecourses identified by TDM can be compared to the timecourse results described above. We see that the Early and Late timecourses resemble the ones found through binning by cortical depth and binning by EPI intensity, but are somewhat more extreme in nature (**Supplementary Figure 2**). For example, time-to-peak is earlier for the Early timecourse and later for the Late timecourse compared to the binning-based timecourses. These effects can be seen more clearly in the quantitative summary (right column). Thus, TDM appears to be extracting sensible timecourses (see also **Supplementary Figure 3**).

Supplementary Note 2: Control analyses for TDM

To further support the validity of TDM, we perform several control analyses:

1. Given that there are timecourse latency differences across cortical depth (see **Supplementary Figure 2**), one might wonder what would happen if TDM were applied to timecourses at different depths. Results show that timecourse latency does in fact increase at superficial depths (**Supplementary Figure 4**); however, identified Early and Late timecourses are, on the whole, quite similar across depth, consistent with the observation that venous effects are not isolated to superficial depths but can extend throughout the thickness of cortex[4].
2. One might wonder how well TDM performs on data of lesser quality or quantity. Through simulations in which we impose different levels of noise and different levels of subsampling, we show that the performance of TDM is quite robust, but eventually breaks down as would be expected for a data-driven technique (**Supplementary Figure 5**).
3. Independent components analysis (ICA) is a commonly used statistical method for deriving latent structure in a set of data, and is commonly applied in fMRI. We show that applying an ICA-based procedure yields results similar to those obtained from TDM in some datasets, but divergent and unsatisfactory results in other datasets (**Extended Data Figure 1**).
4. The observed variability in timecourses across subjects (see **Figure 3**) is not due to measurement noise: split-half analyses show that the TDM-derived timecourses are highly stable across splits of each dataset (**Supplementary Figure 3**).
5. TDM is a data-driven technique that attempts to flexibly estimate timecourses in each given dataset. We show that accuracy of timecourse estimation is impaired if one uses canonical “off-the-shelf” basis functions from commonly used software packages SPM and FSL (**Supplementary Figure 6**). Thus, to achieve the most accurate characterization of responses, one should tailor timecourses to what is empirically observed in individual datasets.
6. Although the use of three dimensions in the TDM visualizations is appealing because it is practical to create pictorial representations of a small number of dimensions, visualizing timecourses in only three dimensions might provide an incomplete characterization of the full diversity of timecourses. To investigate this issue, we perform a cross-validation analysis to determine the number of timecourse dimensions necessary to capture signals of interest in the response timecourses (**Figure 2D**). We find that in every dataset, cross-validation performance is maximized using three or fewer PCs. Thus, it appears that using three PC dimensions is sufficient and that additional dimensions would likely be dominated by measurement noise.

Supplementary Note 3: Quantitative results for TDM

We calculate several quantitative metrics of the performance of the TDM method. First, we construct histograms of the distributions of betas obtained under the different GLMs (**Extended Data Figure 6A**). These distributions have long tails (**Extended Data Figure 6A, bottom**) that appear to correspond to very large BOLD responses from the macrovasculature (see **Figure 4**). We quantify the magnitude of these tails using kurtosis, a metric that is large for heavy-tailed distributions. We find that Late betas have very high kurtosis, whereas Early betas have relatively low kurtosis (**Extended Data Figure 6B**). Standard betas have an intermediate level of kurtosis, consistent with the interpretation that Standard betas behave essentially like an average or mixture of the Early and Late betas. In addition to quantifying the tails of the distributions, we quantify the overall magnitude of the betas by calculating the standard deviation of each distribution. We find that unlike kurtosis, the standard deviations of the beta distributions are similar across the three versions of the betas (**Extended Data Figure 6C**). These results confirm that Early and Late betas are generally large in magnitude (indicating that BOLD responses are observed for both Early and Late timecourses), but Late betas have the unique feature of sometimes reaching extremely high values. This is consistent with the interpretation of Late betas as reflecting macrovascular responses.

Next, we construct depth profiles in order to understand how the magnitudes of BOLD responses change with cortical depth. For these analyses, we restrict our quantification of BOLD responses to portions of cortex that have at least some substantive BOLD response for each given experimental condition. Specifically, for each condition, we select vertices whose Standard beta exceeds 3% signal change at any cortical depth. We then average the BOLD response across these vertices at each depth (thus, a common set of vertices is used across depths). Consistent with inspections of surface maps, we find that the magnitudes of both Standard betas and Late betas exhibit strong depth dependence (**Extended Data Figure 6D**). For example, the BOLD signal change for Late betas is about 4 times greater at Depth 1 (outer) than at Depth 6 (inner). In contrast, the BOLD signal change for Early betas is fairly constant across depth. This effect is observed both in the group average (**Extended Data Figure 6D, main plot**) as well as in each individual dataset (**Extended Data Figure 6D, inset plots**).

An important aspect of betas is their reliability, i.e. robustness or consistency across repeated measurements. We find that the Standard betas are the most reliable across splits of the data, and there is a drop in reliability for the Early and Late betas (**Extended Data Figure 6E**). The decrease in reliability for the Early and Late betas is unfortunate but not surprising: Early and Late timecourses are typically highly correlated, and the accuracy of regression estimates in the case of correlated regressors is expected to be somewhat degraded.

Supplementary Note 4: TDM is compatible with other acquisition methods

To gain further insight into the nature of TDM and its generalizability, we repeated the eccentricity experiment using low-resolution fMRI (3T, 2.4-mm; Datasets D15–D16) as well as spin-echo fMRI (7T, 1.05-mm; Datasets D13–D14). The acquisition was performed in subjects who also participated in the high-resolution gradient-echo acquisition, enabling direct comparison of results.

We find that the TDM method successfully applies to both acquisition styles. In the low-resolution data, we see that the distribution of timecourse shapes tightens (**Extended Data Figure 2, fifth and seventh rows**), which likely reflects a combination of averaging diverse timecourses within individual voxels and reduction of thermal noise. Moreover, we find that TDM derives Early and Late timecourses that closely resemble those found in the high-resolution data (**Extended Data Figure 2, fourth column**). This implies that even at a resolution of 2.4 mm, there is sufficient diversity of timecourses to support data-driven discovery of latent timecourses. Examining the surface visualizations (**Extended Data Figure 3**), we observe results for the low-resolution measurements that are consistent with the high-resolution measurements. However, the differences between the spatial patterns of the Early and Late betas are reduced, indicating that the differences between microvascular and macrovascular effects tend to wash out at low spatial resolutions. In other words, the primary benefits of TDM may be more for high-resolution fMRI studies where idiosyncrasies of the vasculature are more prominent, and less for fMRI studies in which acquisition resolution is low or heavy spatial smoothing is used.

In the spin-echo data, we find patterns of results that look remarkably similar to the gradient-echo data. Early and Late timecourses are identified (**Extended Data Figure 2**), and large betas are found for Late timecourses (**Extended Data Figure 3**). Standard GLM analysis of spin-echo data yields betas that exhibit depth-dependent bias in BOLD signal change (**Extended Data Figure 6F, upper right, black lines**), and this bias is largely eliminated after applying TDM (**Extended Data Figure 6F, upper right, red lines**). Importantly, the spin-echo measurements suffer from a decrease in sensitivity compared to the gradient-echo measurements (**Extended Data Figure 6F, bottom plots, black lines**), even though larger voxels were used for the spin-echo measurements. Overall, these results indicate that spin-echo measurements still contain substantial contributions from the macrovasculature and that TDM is able to identify and remove the macrovasculature-related effects.

Supplementary Note 5: Description of alternative ICA-based procedure

Independent components analysis (ICA) is a widely used technique in fMRI, and is a potential alternative method for determining latent timecourses. We designed an ICA-based procedure (*icadecomposehrf.m*) that serves as a drop-in replacement for the TDM algorithm. In the procedure, we take the two condition-split versions of the FIR-derived timecourses and perform ICA on each set of timecourses (FastICA Toolbox, https://research.ics.aalto.fi/ica/fastica/, default nonlinearity). This produces 31 independent component (IC) timecourses for each split of the data. Because ICA does not provide a natural ordering or grouping of the ICs, the challenge is to determine which specific pair of ICs to use as the latent timecourses in TDM.

To determine the pair of ICs to use, we devised the following heuristic procedure. We first greedily order the ICs within each split to maximize variance explained. That is, we choose the IC that maximizes variance explained in the timecourses, choose a second IC that, when combined with the first, maximizes variance explained, and so on. For normalization purposes, we flip each IC if necessary to ensure that it is positive over the range 0–10 s and normalize it to unit length. Next, we perform greedy matching in order to match the ICs in the second split of the data to the ICs in the first split. Specifically, we choose the IC in the second split that is most similar in a squared-error sense to the first IC in the first split, choose the remaining IC in the second split that is most similar to the second IC in the first split, and so on. To reduce the number of ICs under consideration, we select ICs that both exhibit high consistency across the two splits of the data (R^2 > 50%) and are within the top 50% of the ICs in the first split (with respect to the ordering based on variance explained). Finally, from the ICs that meet these selection criteria, we perform an exhaustive search to determine the unique pair of ICs that explain the most variance in the original timecourses. Based on time-to-peak, we label one IC as ‘Early’ and one IC as ‘Late’.

Supplementary Discussion: Comparison of TDM to other methods

Strengths of the TDM method include the following:

- *Flexibility*. TDM has fairly minimal prerequisites. It does not require auxiliary data, and can be applied to different types of acquisitions, including spin-echo and low-resolution acquisitions (**Supplementary Note 4**). As a data-driven method, TDM makes no assumptions about shapes of hemodynamic timecourses that might be found in a dataset and naturally adapts to different brain areas, subjects, and/or datasets. Furthermore, since the response decomposition is performed independently for each voxel, the technique makes no assumptions regarding the spatial distribution of BOLD responses in a given experiment. In other words, the loadings on the Early and Late timecourses can vary from voxel to voxel, and the technique can, in theory, capture these variations. Finally, TDM applies to single-condition activity maps, and therefore avoids the assumptions required in differential paradigms where unwanted non-specific effects are assumed to be removed through subtraction.
- *Interpretable visualizations*. An integral aspect of TDM is direct visualization of distributions of timecourses. Thus, it is easy for the user to understand the nature of the data and whether the derived timecourses are meaningful. This stands in contrast to 'black box' methods that might produce unusual results without explanation.
- *Simplicity and robustness*. TDM is simple in its design, is fully automated, and does not, as far as we have seen, require fine-tuning of parameters to be effective.
- *Analysis not acquisition*. Since TDM is an analysis method, it can be retrospectively applied to datasets that have already been acquired. Moreover, since TDM does not place major constraints on acquisition, the user is not burdened with making difficult decisions regarding optimal acquisition parameters (e.g., choosing between a standard acquisition scheme that is guaranteed to produce reasonably strong signals versus a specialized acquisition scheme that might suffer from low sensitivity).

Limitations of the TDM method include the following:

- *Sensitivity loss*. TDM involves decomposing fMRI responses using two timecourses that are often highly correlated. Thus, from a regression perspective, one expects to incur a penalty in terms of high variance in beta estimates, and the correction provided by TDM may suffer from instability. Nonetheless, simulations confirm that model estimates are expected to recover, on average, the correct values (**Supplementary Figure 7**). If sensitivity (i.e. reliability of beta estimates) is the sole priority and specificity (i.e. accurate estimates of local neural activity) is not critical, the TDM method is not recommended. If, on the other hand, specificity is of utmost importance, TDM is likely to be a valuable method. In short, TDM does not deliver more robust fMRI maps (e.g. large blobs of statistically significant activations), but aims to deliver more spatially accurate and neurally meaningful maps.
- *Intrinsic physiological limitations*. The TDM method attempts to disambiguate BOLD contributions from the microvasculature and macrovasculature based on their respective associated timecourses. The more similar these timecourses, the more difficult it will be to estimate the distinct contributions of the timecourses. Thus, the intrinsic physiology of the subject places limits on the overall effectiveness of the TDM method. For example, in our data (see **Extended Data Figure 1**), we find that Subject S5 exhibits Early and Late timecourses that are widely separated in the 3-dimensional PCA space and are highly distinct, whereas Subject S4 exhibits Early and Late timecourses that are close together in the 3-dimensional PCA space and are fairly similar (note that Subject S4's responses do not simply suffer from poor data quality but are highly reliable, as shown in the figure). This may be the reason why some datasets experience a larger reduction in reliability when using TDM compared to other datasets (see **Extended Data Figure 6E**). One possible approach to achieve optimal results is to screen subjects according to the temporal separability of their microvasculature- and macrovasculature-related timecourses.
- *Complexity of the vasculature.* TDM proposes a simple two-component model to decompose BOLD timecourses. The vasculature is certainly more complex than this simple characterization[64]. Thus, the

correction provided by TDM may be only a first-pass approximation to remove venous effects, and it may be fruitful to develop a more nuanced characterization of vasculature types and their dynamics.

- *Potential interactions with other analysis steps.* Functional MRI experiments come in many varieties and are analyzed in many different ways. We caution that subtle differences in experimental design and pre-processing pipelines (e.g., exactly how slice time differences are handled) may substantially impact the performance of TDM. When approaching new datasets, the stability, accuracy, and validity of TDM should not be taken for granted, but should be carefully assessed (e.g., by inspecting visualizations of timecourse characteristics such as those depicted in **Figure 2**).

How does TDM compare to other approaches for removing or avoiding venous effects in fMRI? Some researchers have proposed simple heuristic selection methods. For example, sampling BOLD responses at only deep cortical depths[40] can help avoid the influence of large draining veins near the pial surface. However, this comes at the cost of not being able to infer response properties in superficial cortical depths; moreover, it is still possible for veins to penetrate deep into cortex (see **Figure 4**). Another example is masking out voxels with very high percent BOLD signal change[6], low EPI intensity[5], and/or low temporal signal-to-noise ratio[32]. While these are reasonable heuristics for removing voxels that are most egregiously affected by large veins, it is not clear what principle can be used to set the threshold to be used. Moreover, similar to the approach of sampling only deep cortical depths, this approach fails to recover usable signals from the removed voxels.

One suggestion found in older work[6,27] and more recent work (Blazejewska, Nasr, Polimeni, ISMRM 2018 abstract) is to sample early time points in the BOLD response. This is certainly consistent with the spirit of TDM and may produce a response snapshot that is more weighted towards the microvasculature. However, our results indicate that Early and Late timecourses are highly overlapping (see **Figure 3**). If the chosen time point is not sufficiently early, this incurs the risk of the late component "bleeding" into the analysis results, and choosing only one or a few time points does not make efficient use of all of the available data. These shortcomings are indeed borne out in our evaluation of a timepoint-based analysis (see **Figure 5**). Compared to these various heuristic selection methods, we believe that TDM has substantial appeal: *TDM can recover signals at all depths and even in voxels that have substantial venous influence; it makes efficient use of all of the fMRI data collected; and, as a data-driven technique, it naturally identifies appropriate timecourse parameters for each dataset.*

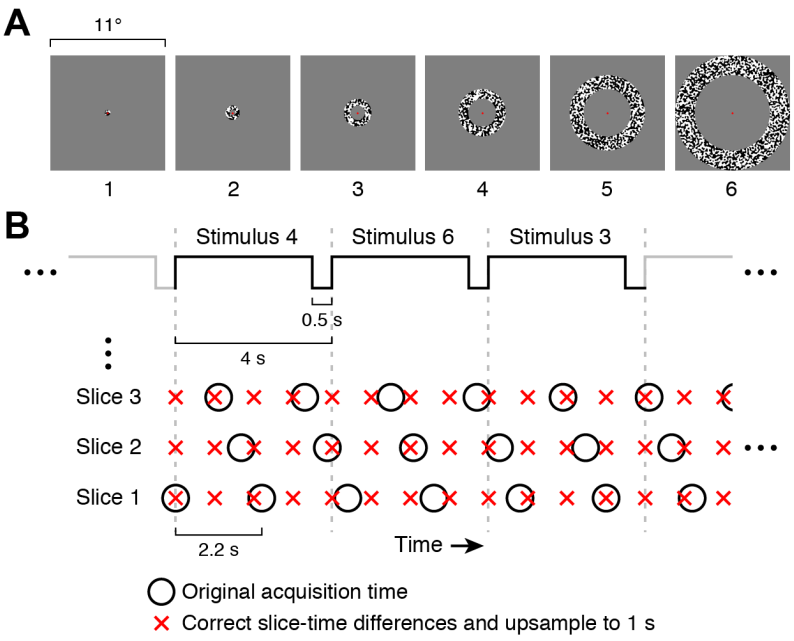
Recently, a method has been proposed that first constructs a forward model characterizing the mixing of hemodynamic signals from different cortical depths due to blood drainage towards the pial surface and then uses this model to invert observed BOLD depth profiles[33,36,37]. This method bears a parallel to TDM in the sense that it is a spatial deconvolution approach whereas TDM is a temporal deconvolution approach. However, the accuracy of the spatial deconvolution approach may be dependent on the correctness of the model parameters, which might vary across brain regions and/or subjects. In addition, the approach deals only with vascular effects that vary across depths. In contrast, TDM is a data-driven technique that adapts to each given dataset and compensates for vascular effects present at every voxel.

Besides analysis methods, one can consider using acquisition methods to avoid venous effects. Switching from conventional gradient-echo pulse sequences to spin-echo pulse sequences (or related techniques such as GRASE[31,38,39]) has traditionally been considered the standard approach for mitigating venous effects in fMRI[13]. While the refocusing of T2* effects by the 180° RF pulse in spin-echo eliminates sensitivity to extravascular effects around large veins, it is important to note this holds only for a specific point in time (typically the center of the readout window). The remainder of the image acquisition incurs T2* effects[65]. Furthermore, spin-echo does not eliminate intravascular effects in large vessels[66–68]. Thus, spin-echo does not provide full elimination of venous effects. In addition, spin-echo acquisitions suffer from increased energy deposition, limits on spatial coverage, lower temporal resolution, and loss of signal-to-noise ratio.

In this study, we have provided a direct comparison of TDM and spin-echo. In order to maintain sensitivity, we acquired spin-echo data at a lower resolution (1.05-mm vs. 0.8-mm) but maintained the same TR and the same

overall experiment duration as the gradient-echo data. Our results show that the spin-echo data analyzed using a standard GLM (single canonical hemodynamic timecourse) is more robust than gradient-echo data decomposed using TDM (see **Extended Data Figure 6F**). One reason for the increased robustness of the spin-echo data is its lower spatial resolution, providing the data with some advantage over the gradient-echo data. However, even if the two types of data were matched in resolution, spin-echo should not be viewed as a complete solution since it does not fully suppress venous effects. Indeed, we demonstrate that TDM can be applied to the spin-echo data in order to remove venous influences present in those data (see **Supplementary Note 4**). When comparing gradient-echo data and spin-echo data that have both been decomposed using TDM, gradient-echo has greater robustness (see **Extended Data Figure 6F**). Thus, we suggest that if removal of venous effects is top priority and one plans to use TDM, there is little benefit to spin-echo acquisition over conventional gradient-echo acquisition.

A promising alternative to spin-echo is vascular space occupancy (VASO), a pulse sequence that indexes changes in cerebral blood volume[35]. Although VASO must be corrected for the BOLD effect through a division operation, after correction the approach appears to generate highly specific measures of task-driven hemodynamic responses[34,69]. A promising direction for future work is to perform direct comparison of an gradient-echo acquisition optimized for use with TDM against an optimized VASO acquisition. Ideally, results would be rigorously validated against direct neural activity measurements, such as electrophysiological recordings with laminar resolution[70,71].



C

Subjects S1–S11	E1	E2	E3	E4
S1	D1	D8	D14	D15
S2	D2			
S3	D3			D16
S4	D4		D13	
S5	D5			
S6		D6		
S7		D7		
S8		D9		
S9		D10		
S10		D11		
S11		D12		

Experiment E1: Eccentricity stimuli, 7T, 0.8 mm, 2.2 s, gradient-echo

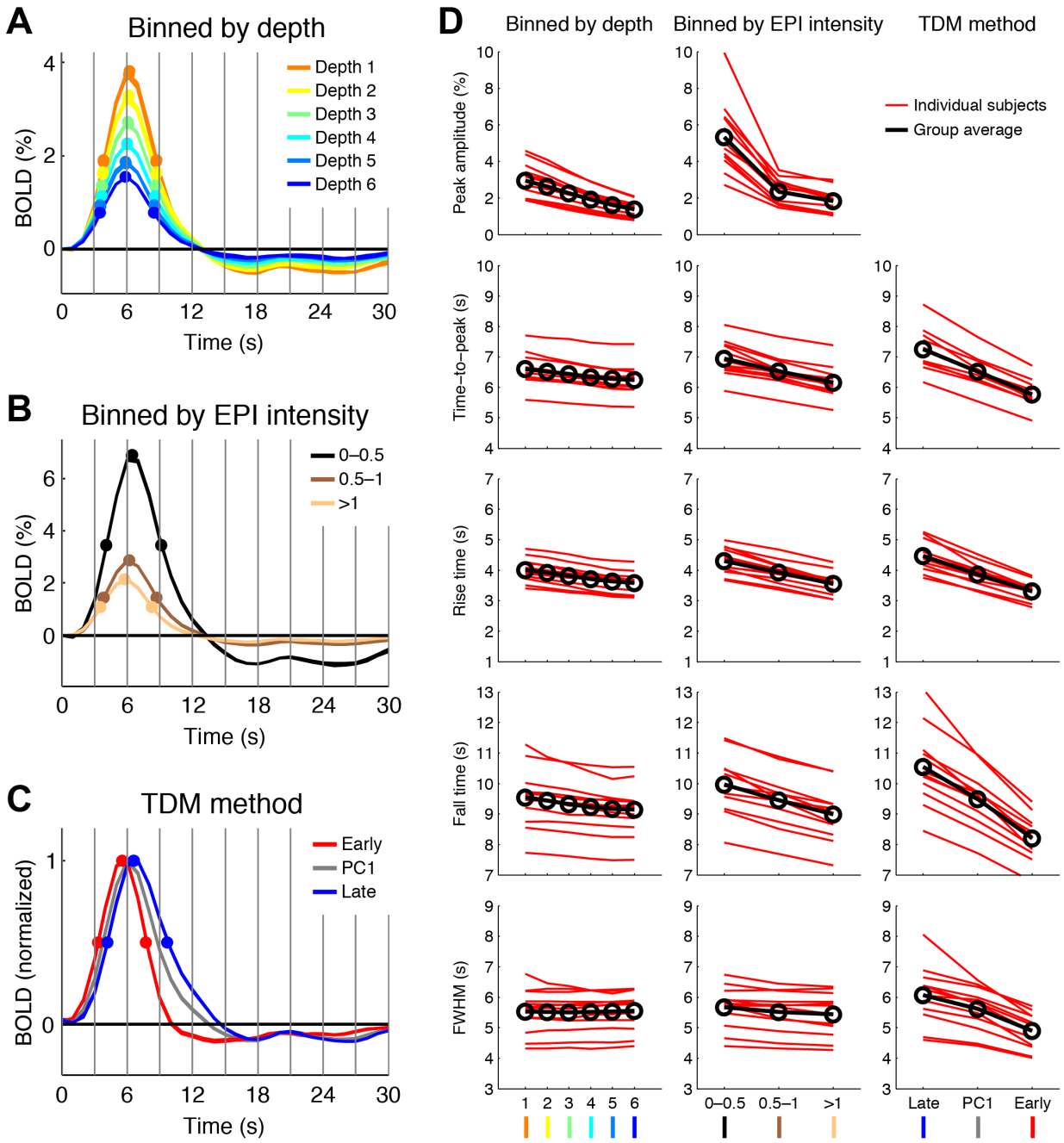
Experiment E2: Category stimuli, 7T, 0.8 mm, 2.2 s, gradient-echo

Experiment E3: Eccentricity stimuli, 7T, 1.05 mm, 2.2 s, spin-echo

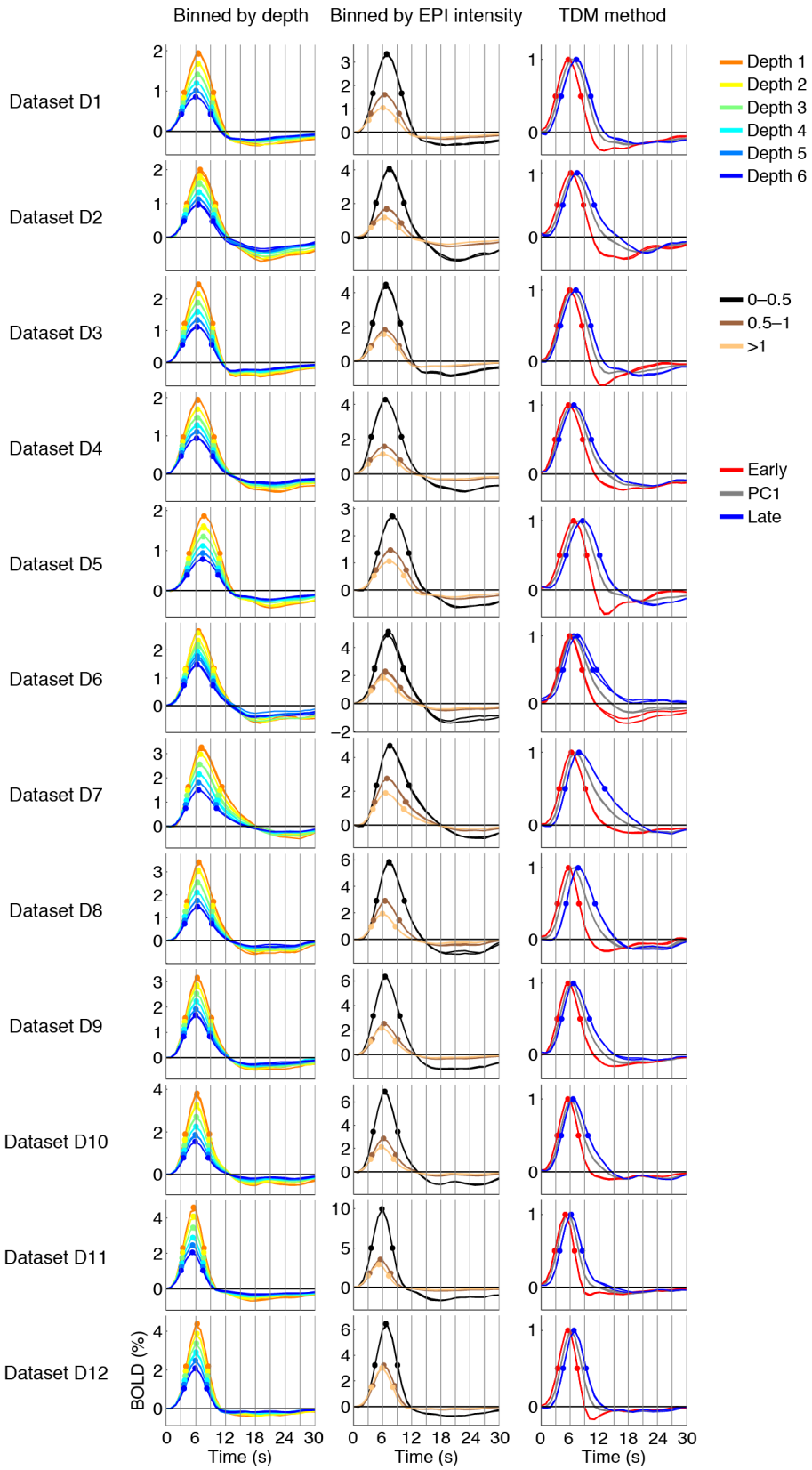
Experiment E4: Eccentricity stimuli, 3T, 2.4 mm, 1.1 s, gradient-echo

Datasets D1–D16

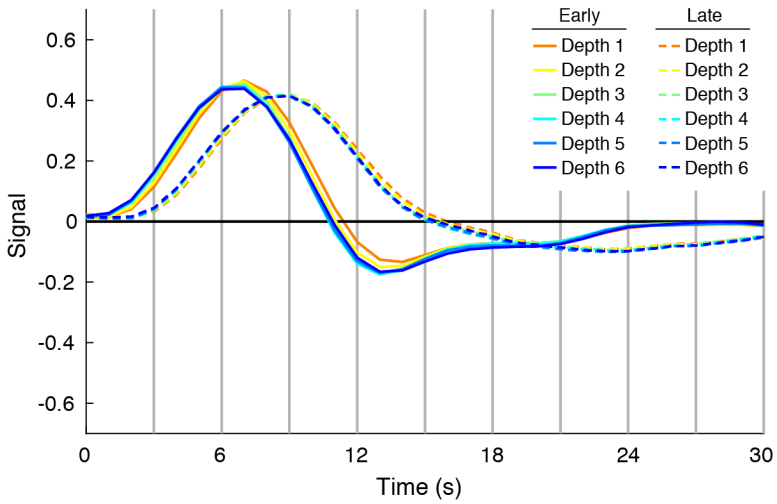
Supplementary Figure 1. Schematic of experiment. *A*, Eccentricity stimuli. Stimuli consist of six rings varying in eccentricity. A small central dot serves as a fixation point. *B*, Timing of trials and fMRI acquisition. Each stimulus lasts 3.5 s and is followed by a 0.5-s gap before the next trial. Slices are acquired at different times within each 2.2-s TR (black circles). In pre-processing, cubic interpolation is used to resample each voxel's time-series data to a rate of 1 s such that the same time points are obtained for all voxels (red crosses). Because the trial duration (4 s) is not evenly divisible by the TR (2.2 s), the experiment automatically incorporates jitter between trial onsets and slice acquisition times. *C*, Summary of datasets. Eleven subjects participated in four experiments. Each dataset corresponds to one scan session, and sixteen datasets were collected. (For details on the category stimuli used in Experiment E2, see Methods.)



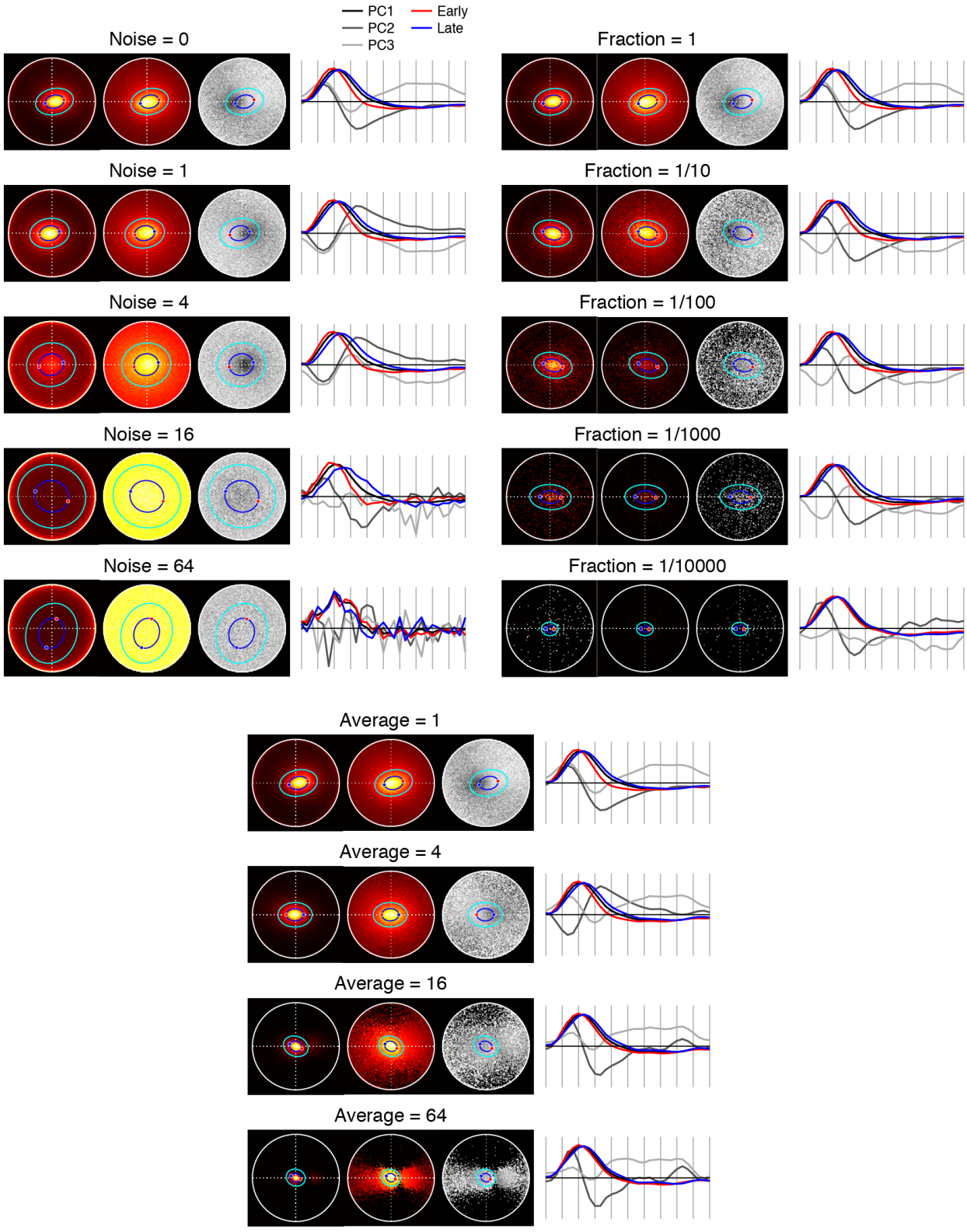
Supplementary Figure 2. BOLD timecourses exhibit systematic variation in amplitude, delay, and width. Panels A–C show detailed results for Dataset D10 (detailed results for all datasets are shown in **Supplementary Figure 3**). *A*, Timecourses binned by cortical depth (Depth 1 is superficial; Depth 6 is deep). Timecourses from split-halves of the data are shown for each bin (the two sets of traces are nearly identical, indicating high reliability). Vertical gray lines mark 3-s intervals, a convention used throughout this paper. Solid dots mark timecourse peak, rise time (time at which the signal rises to half of the peak value), and fall time (time at which the signal falls to half of the peak value). Full-width-half-max (FWHM) is calculated as fall time minus rise time. *B*, Timecourses binned by EPI intensity. Same as panel A except binning is performed with respect to bias-corrected EPI intensity. *C*, Timecourses derived by TDM. The Early and Late timecourses derived by TDM are normalized to peak at 1. For comparison, we also show the first PC of the timecourses, also normalized to peak at 1. *D*, Summary of results across datasets. Timecourse metrics obtained for individual subjects (Datasets D1–D12) and corresponding group averages are shown.



Supplementary Figure 3. Comprehensive summary of BOLD timecourses. Same format as **Supplementary Figure 2**. Qualitative patterns of results are highly consistent across datasets (e.g. time-to-peak is delayed at superficial depths). However, there is substantial quantitative variation across datasets (e.g. time-to-peak is short in Dataset D11 but long in Dataset D5). This underscores the importance of tailoring timecourse derivation to individual subjects or scan sessions.

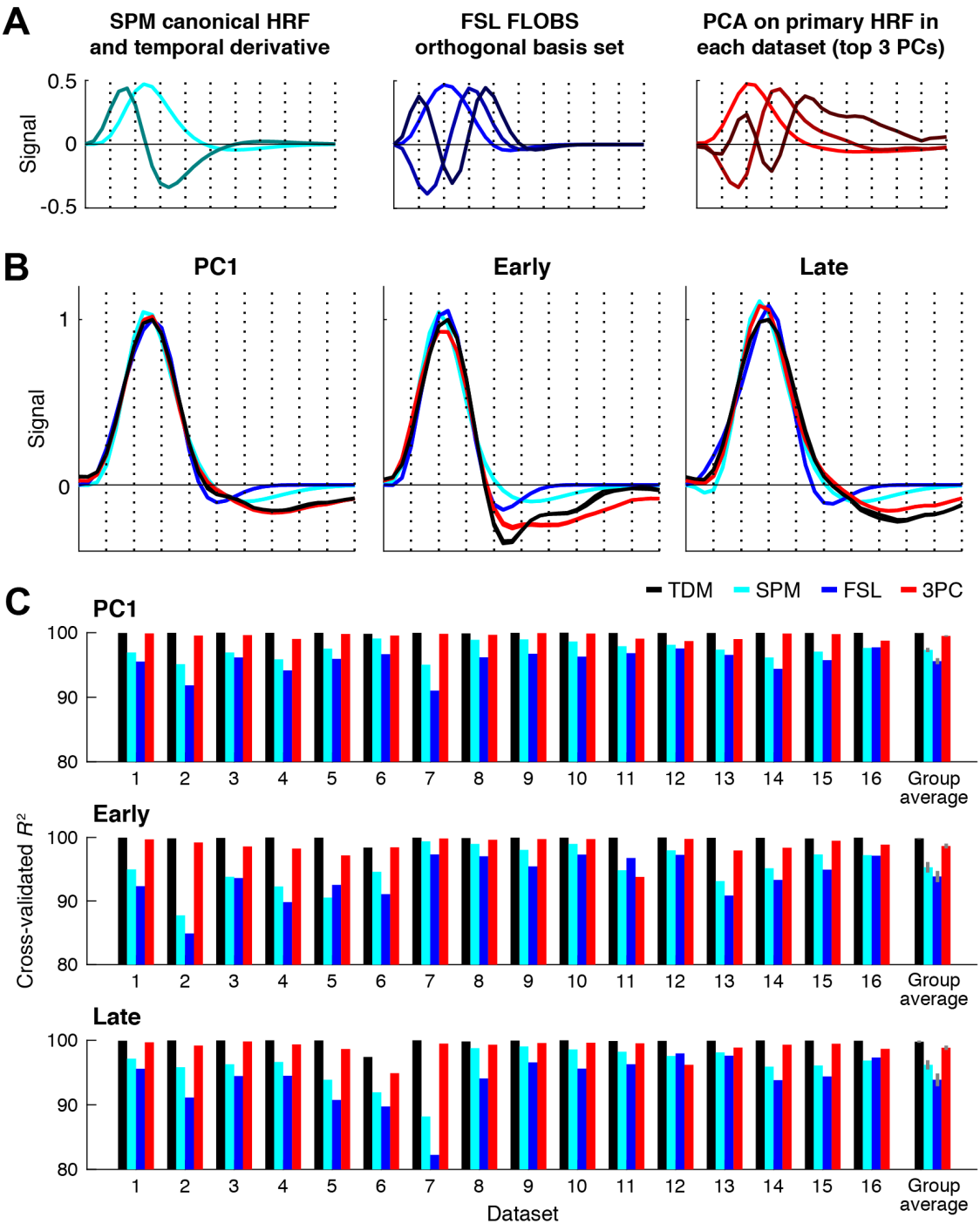


Supplementary Figure 4. Application of TDM to timecourses observed at different cortical depths. Here we plot the results of applying TDM separately to timecourses at the six cortical depths for an example dataset (D5). Moving from inner (Depth 6, blue) to outer depths (Depth 1, orange), Early and Late timecourses exhibit small increases in time-to-peak as well as in rise and fall times. This is consistent with the timing differences observed in **Supplementary Figure 2**. The overall shapes of the Early and Late timecourses are quite similar across depth. This suggests that timecourses primarily reflecting the microvasculature and timecourses primarily reflecting the macrovasculature can both be found throughout the cortical thickness.



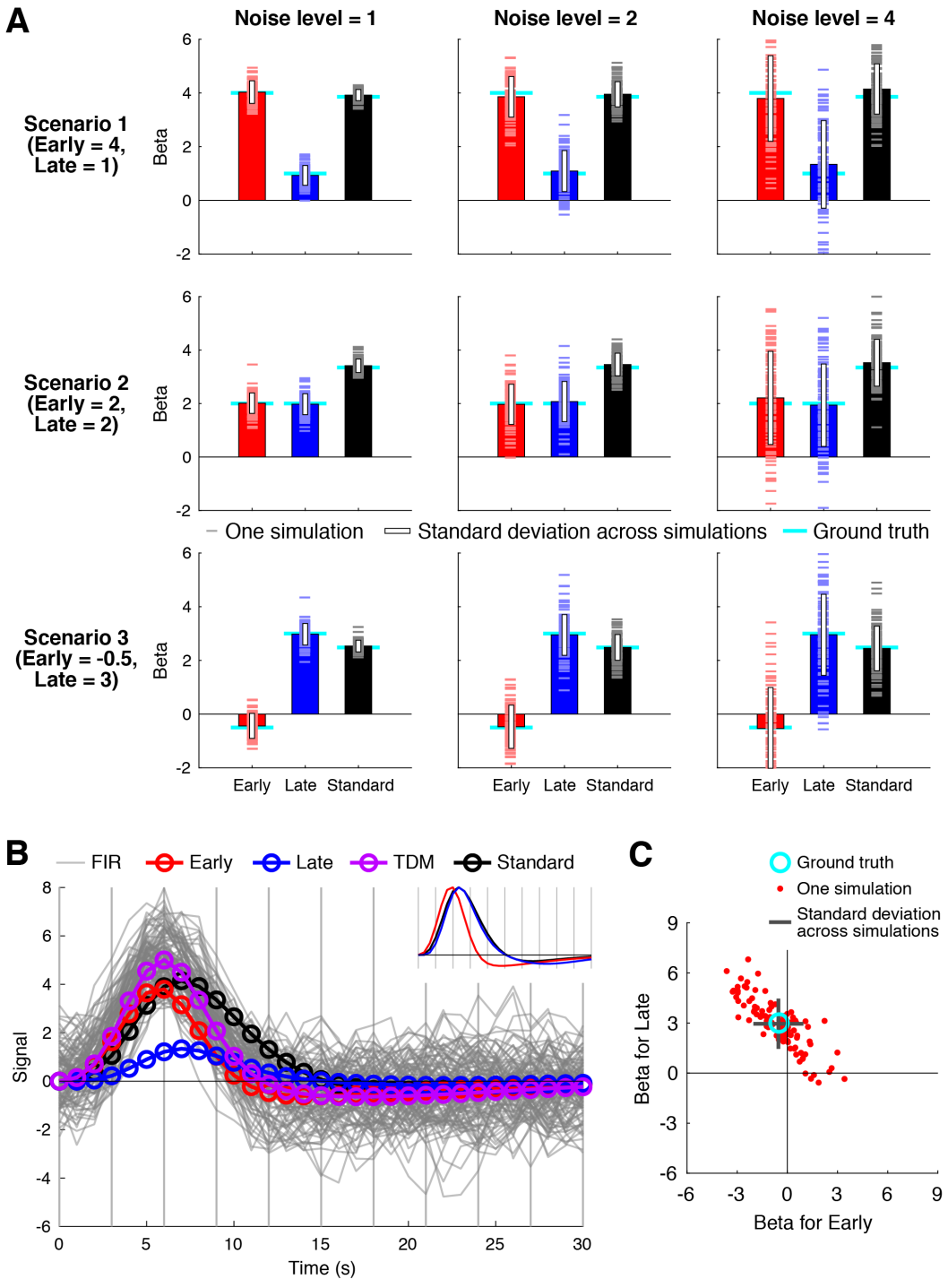
Supplementary Figure 5. Robustness of TDM to noise and other factors. Here we show the results of taking the timecourses observed in an example dataset (D4) and performing three manipulations to the timecourses before applying the TDM method (format same as **Figure 2**). In the first manipulation (upper left), the timecourses are corrupted with Gaussian-distributed noise with zero mean and standard deviation equal to 0, 1, 4, 16, or 64 (percent BOLD signal change units). This simulates increased measurement noise, or, equivalently, the availability of lesser amounts of fMRI data to estimate timecourses. In the second manipulation (upper right), the timecourses are randomly subsampled using a fraction of 1, 0.1,

0.01, 0.001, or 0.0001 of the available surface vertices. This simulates situations where only a small portion of brain is imaged. In the third manipulation (bottom), the timecourses and associated EPI intensities are averaged within groups of 1, 4, 16, or 64 vertices (groups are randomly selected and are mutually exclusive). This explores the consequences of reducing timecourse diversity through spatial averaging (which might roughly approximate low-resolution acquisition and/or spatial smoothing). Overall, we see that TDM is quite robust: the identified Early and Late timecourses are relatively unchanged from noise standard deviation 0 through 16, fraction 1 through 0.001, and group size 1 through 4. Interestingly, successful timecourse derivation can even occur in cases where the density, vector-length, and intensity images lack obvious visual features. The observation that increased noise yields distributions that are more isotropic in shape makes sense, given that random noise will correspond to random directions in the low-dimensional PCA space. Also, the observation that vertex averaging yields narrower distributions makes sense, given that averaging tends to push a sample towards its mean. Finally, in the most corrupted scenarios (bottom row), TDM provides low-quality timecourses and/or non-differentiated Early and Late timecourses. This eventual breakdown of performance is expected given that TDM is a data-driven technique.



Supplementary Figure 6. Canonical basis functions provide suboptimal timecourse estimation. We carried out a cross-validation analysis to assess how well canonical timecourses characterize the responses observed in our datasets. In this analysis, we first collected the results in which TDM is used to derive PC1 and Early and Late timecourses from split-halves of each dataset. We then assessed, using two different approaches, how well results from the first split can be used to predict the results of the second split. In the 'data-driven' approach, we simply used the timecourse from the first split as a predictor for the timecourse from the second split. In the 'basis-restricted' approach, we took the timecourse from the first split, projected the timecourse into the subspace spanned by a given set of basis functions, reconstructed the timecourse, and then used it as a predictor for the timecourse from the second split. In both approaches, the predictor timecourse is scaled to best match the observed timecourse and the quality of prediction is quantified in terms of variance explained (R^2). The use of cross-validation ensures that performance is not simply a reflection of overfitting. *A*, Basis functions. We evaluated three different sets of basis functions. For the first set (left), we used SPM's *spm_hrf.m* (default parameters) to generate the predicted response to a 4-s stimulus. The set consisted of this canonical HRF and its temporal derivative. For the second set (middle), we used FSL's

Make_flobs (default parameters) to generate a set of 3 orthogonal basis functions (these functions optimally span a collection of timecourses internally simulated by the tool). For the third set (right), we created a data-driven set of basis functions by taking PC1 from Datasets D1–D12, normalizing each of these to unit length, performing PCA, and selecting the top 3 PCs. All basis functions were prepared at a sampling rate of 1 s and normalized to unit length. Dotted lines mark 3-s increments from 0 s to 30 s. *B*, Split-half timecourses for an example dataset (D5). Each panel shows timecourses from both split-halves of the data. The data-driven timecourses are shown (black) as well as the reconstructions of these timecourses (cyan, blue, red). The distinction between the split-half results is hardly visible, indicating the high reliability of the results. Notice that the basis-restricted approaches fail to fully capture the observed timecourses. *C*, Cross-validation performance. For each timecourse (PC1, Early, Late), we plot the cross-validation performance of the data-driven approach (black) and each of the basis-restricted approaches (cyan, blue, red). Results are shown for individual datasets as well as for the group average (mean and standard error across datasets). The basis-restricted approaches consistently underperform the data-driven approach, indicating that the use of canonical basis functions fail to optimally capture the timecourse dynamics observed in our datasets.



Supplementary Figure 7. The TDM regression procedure can successfully recover early and late components. Here we perform simulations (code available at https://osf.io/j2wsc/) in order to illustrate and validate the regression-based procedure used in TDM to estimate the contributions of early and late components in hemodynamic timecourses. In these simulations, we generate synthetic data for one fMRI run (300 s, TR 1 s) in which a single experimental condition is presented (4-s trial duration) and gives rise to two distinct but overlapping timecourses. We then analyze these synthetic data using the three GLM models in this paper (FIR, Standard, TDM). Each simulation consisted of the following steps: (1) *Generate experimental regressor*. Each of 75 trials were designated to have the experimental condition presented with probability 0.25. (2) *Convolve experimental regressor with ground-truth HRFs*. For realistic timecourses, we used the fitted double-gamma functions for the group-average Early and Late timecourses from **Figure 3**. (3) *Generate low-frequency nuisance regressors*. Polynomials of degree 0 through 3 were constructed. The constant polynomial was scaled to one. The remaining polynomials were *z*-scored. (4) *Generate synthetic data*. We computed a weighted sum of the experimental and nuisance regressors.

Ground-truth beta weights for the experimental regressors were systematically varied (one beta weight for Early, one beta weight for Late), whereas the ground-truth beta weight for the constant polynomial was set to 100 and the ground-truth beta weights for the remaining polynomials were drawn from a Gaussian distribution (mean 0, standard deviation 2). (5) *Add realistic noise*. A noise time series was generated by drawing values from the normal distribution, convolving with a 3-s square wave (with amplitude $1/\sqrt{3}$ to maintain variance), and multiplying by a specific noise level. (6) *Fit GLM models.* The composite time series (experimental regressors + nuisance regressors + noise) was analyzed using the GLM models described in the Methods. *A*, Simulation results. For each of three noise levels (noise level = 1, 2, 4) and each of three experimental scenarios (early and late betas = (4,1), (2,2), (-0.5,3)), we performed 100 simulations. Each plot shows the ground-truth amplitudes (cyan lines), individual simulation results (thin lines), and the mean and standard deviation of the beta weight estimates across simulations (bars and error bars). In all cases, beta weight estimates successfully recover, on average, the ground-truth values. The reliability of beta weights is reduced for Early and Late compared to Standard, and decreases as the noise level is increased. *B*, Detailed view of timecourses. For one case (noise level = 4, betas = (4,1)), we show FIR estimates obtained in each of the 100 simulations (gray lines), estimates from the Standard GLM (black; averaged across simulations), and estimates from the TDM model (red, blue, purple; averaged across simulations). The inset shows the three timecourses used in the GLM models; these have been normalized to the same height to enhance visibility of timecourse differences. *C*, Detailed view of estimation results. For one case (noise level = 4, betas = (-0.5,3)), we show the distribution of beta weight estimates across simulations (red dots). Although the Early and Late estimates are negatively correlated, their average value is consistent with the ground-truth values. Overall, these results demonstrate that although the TDM regression procedure may suffer from decrease in reliability of beta weight estimates, the results should, on average, converge to the correct outcome.

Supplementary References

64. Uludağ, K. & Blinder, P. Linking brain vascular physiology to hemodynamic response in ultra-high field MRI. *NeuroImage* **168**, 279–295 (2018).
65. Goense, J. B. M. & Logothetis, N. K. Laminar specificity in monkey V1 using high-resolution SE-fMRI. *Magnetic resonance imaging* **24**, 381–392 (2006).
66. Budde, J., Shajan, G., Zaitsev, M., Scheffler, K. & Pohmann, R. Functional MRI in human subjects with gradient-echo and spin-echo EPI at 9.4 T. *Magn Reson Med* **71**, 209–218 (2014).
67. Duong, T. Q. *et al.* Microvascular BOLD contribution at 4 and 7 T in the human brain: gradient-echo and spin-echo fMRI with suppression of blood effects. *Magn Reson Med* **49**, 1019–1027 (2003).
68. Jochimsen, T. H., Norris, D. G., Mildner, T. & Möller, H. E. Quantifying the intra- and extravascular contributions to spin-echo fMRI at 3 T. *Magn Reson Med* **52**, 724–732 (2004).
69. Huber, L. *et al.* Sub-millimeter fMRI reveals multiple topographical digit representations that form action maps in human motor cortex. *Neuroimage* **208**, 116463 (2020).
70. Maier, A., Adams, G. K., Aura, C. & Leopold, D. A. Distinct superficial and deep laminar domains of activity in the visual cortex during rest and stimulation. *Front Syst Neurosci* **4**, (2010).
71. Self, M. W., van Kerkoerle, T., Goebel, R. & Roelfsema, P. R. Benchmarking laminar fMRI: Neuronal spiking and synaptic activity during top-down and bottom-up processing in the different layers of cortex. *Neuroimage* **197**, 806–817 (2019).