Realistic Synthetic Dataset

This document provides a general overview of the proposed synthetic dataset and a summary of their main features. These details can be also found in the supplementary material of the main paper in [ref].

This dataset resembles a demanding experimental designs, where the sources present highly relative different energy and their corresponding brain areas that correspond to different condition overlap. This renders them hardly detectable by conventional BSS methods. Moreover, the new dataset is examined under realistic noise levels and statistics and it also comprises a relatively large number of components/sources that exhibit a large range of sparsity characteristics. In order to succeed in the above and also to study the effect of HRF miss-modeling, certain interventions in the SimTB data generation procedure needed to take place.

SimTB implements a simple spatiotemporal separable model characterized by three fundamental features: a) the spatial maps are smooth (Gaussian-shaped) b) the sources are barely overlapped and c) a fixed HRF is considered. Although these features simplify the study of synthetic experiments, they are unrealistic; the overlap between the standard sources of SimTB only affects small areas with relatively low intensity, whereas, in real fMRI, some sources may exhibit high overlap [1]. Moreover, HRF naturally varies among subjects [2], [3].

For these reasons, we modified the default mode of SimTB to accomplish a more realistic synthetic fMRI-like dataset: first, we translated in space some of the sources to increase their relative overlap. Second, instead of smooth Gaussian shaped spatial maps, which show maximum activity at only one specific peak/voxel, we prefer to have more than one neighbor voxels activated with the same intensity. We achieved that, by starting from the standard SimTB's spatial maps and reducing the highest activated voxels to have the same relative energy, within a small neighborhood.

Regarding the HRF, many different parametric models have been built that capture its shape, with the most popular being two gamma distributions model [4], [5], [3]. A certain set of parameter values produces the cHRF that represents a parametric approximation of the real HRFs [4]. The cHRF is

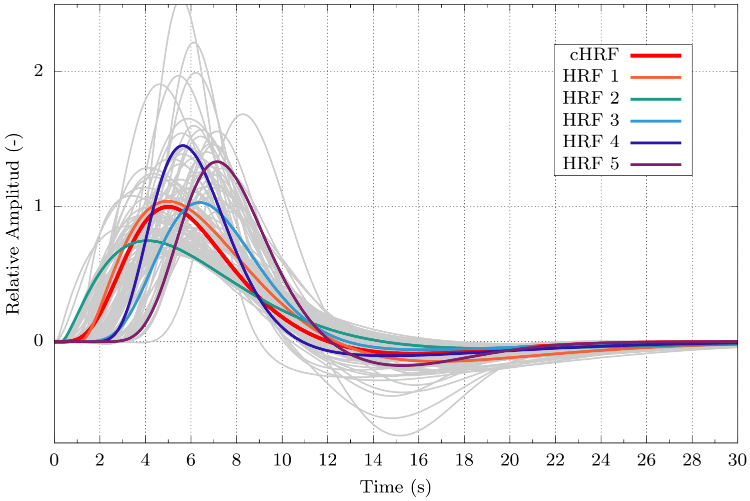


Fig. 1. Graphic representation of 100 HRFs (gray) randomly generated from the two gamma distributions model. The red curve represents the canonical HRF (cHRF) and the rest colored HFRs stands for the five selected alternatives.

TABLE I
Summary of the main features of the spatial maps of the proposed synthetic dataset, including their corresponding sparsity level.

Brain-like sources

Surce	Anatomical Correspondence*	Active voxels (ϕ_i)	Sparsity (%) (θ_i)
1	Bilateral Visual	472	95.28 %
2	Medial Frontal	467	95.33 %
3	Precuneous	447	95.53 %
4	Default Mode Network	1175	88.25 %
5	Subcortical Nuclei	670	93.30 %
6	Subcortical Nuclei - Putamen	293	97.04 %
7	White Matter Tracts (anterior)	1193	88.07 %
8	Dorsal Atention Network	818	91.82 %
9	Frontoparietal (Right dominance)	1449	85.51 %
10	Subcortical Nuclei - Thalamus	733	92.67 %
11	Right Auditory	840	91.60 %
12	Left Auditory	847	91.53 %
13	Right Hipocampus	549	94.51 %
14	Left Hipocampus	543	94.57 %
15	White Matter Tracts (posterior)	2805	71.95 %

* see SimTB toolbox [6]

Artifacts

Source	Main fatures**	Active voxels (ϕ_i)	Sparsity (%) (θ_i)
16	Sub-Gaussian	9900	1.00 %
17	Gaussian	9900	1.00 %
18	Super-Gaussian	9801	1.99 %
19	Sub-Gaussian	1386	86.14 %
20	Super-Gaussian	2816	71.84 %

** according to [7]

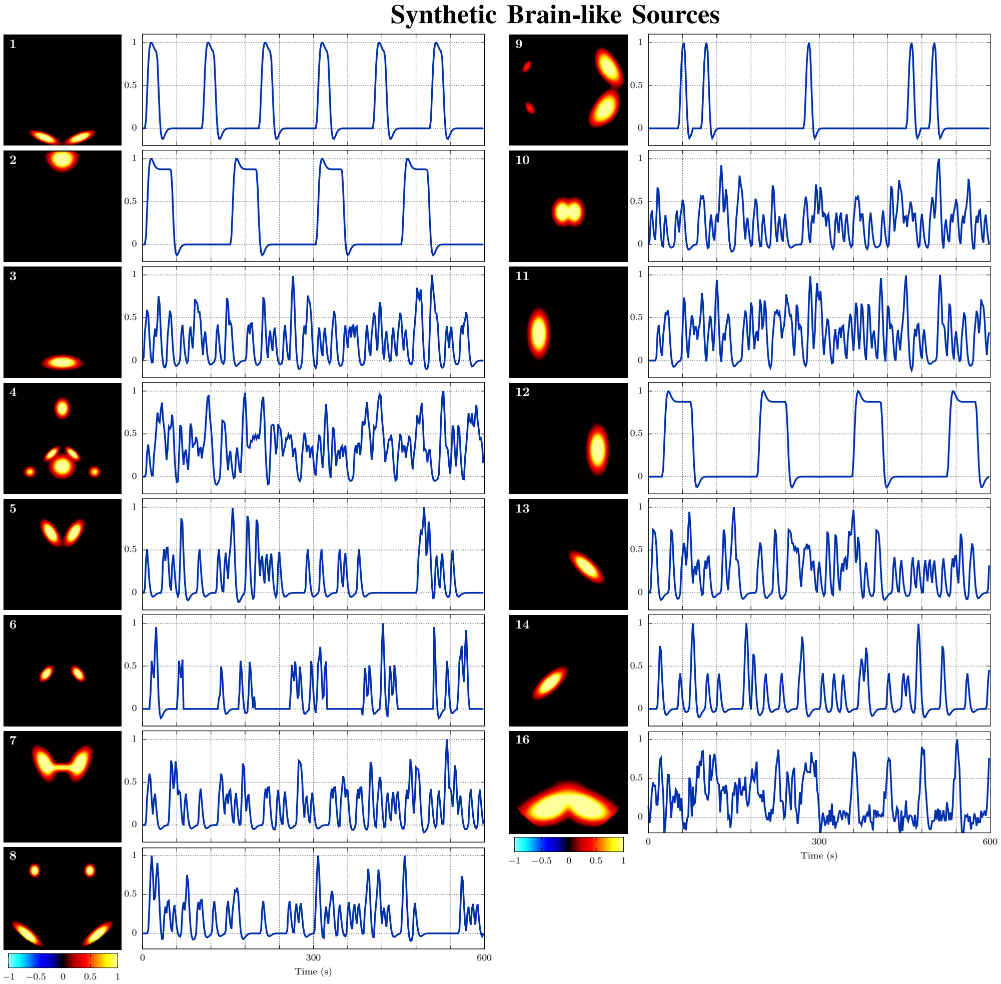
Mean values

	Group	Active voxels	Sparsity (%)
	Whole dataset	2355	76.45 %
	Brain-like sources	887	91.13 %
	Artifacts	6761	32.39 %

the default model in many different toolboxes such as SimTB [6], and and it is also used in the GLM/SPM analysis for the computation, via convolution, of the task related time-courses. Since one of the scopes of this study is to analyze the effects of HRF mismodelling, we account for the natural variability of the HRF, emulating six different *synthetic subjects*, with each one of them having its own specific HRF. Then, we generate one dataset per subject since all the brain-induced sources are HRF-dependent.

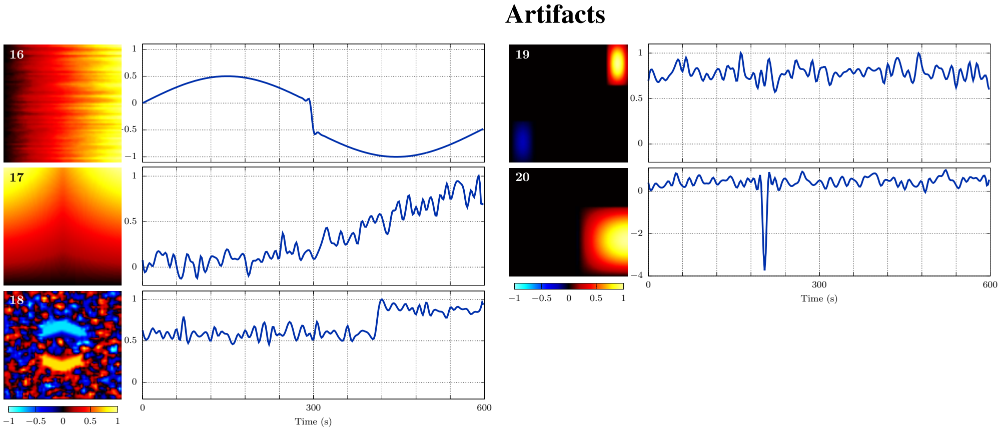
The distinct subject-specific HRFs were generated using the two-gamma distribution model and in order to ensure that they are realistic enough, we followed the procedure described next. We fed the free parameters of the two-gamma distribution model with values drawn from a uniform distribution centered on the parameter values that correspond to the cHRF. The support of the aforementioned distribution was experimentally tuned to a level that led to HRFs, which visually resemble the natural variability of HRFs. For better inspection, a total of 100 HRFs was generated (gray lines in Figure 1) and qualitatively compared with natural HRFs estimated from real fMRI studies, e.g., Figure 3 and Figure 4 of [3]. It is clear that the 100 simulated HRFs are remarkably similar to true ones. For the simulation studies that we presented in the Section IV of the main text, five randomly generated HRFs, as well as the cHRF, were used (depicted with colored curves in Figure 1). For simplicity, hereafter, and we define the "canonical" subject as the subject that will use the cHRF, we introduce the letters A, B, C, D and E to refers to the subjects that correspond to the HRFs 1, 2, 3, 4 and 5, respectively.

The subject-specific datasets comprise 15 brain-like sources. Each source is defined by its spatial map, which is represented by a single brain slice of size 100×100 voxels and a time course that comprise 300-time instances with a difference of two seconds between acquisition times (TR=2).



Synthetic Brain-like Sources

Fig. 2. Visual representation of the synthetic spatial maps and their corresponding time courses generated with the canonical HRF. The intensity of the spatial maps and time courses were normalized to have maximum intensity at 1.



Artifacts

Fig. 3. Visual representation of the synthetic artifacts. The first column depicts the spatial maps and their corresponding spatial maps for the artifacts 3, 4 and 5, whereas the second column shows the artifacts 7 and 8 from the dataset [7].

Figure 2 shows the brain-like sources for the canonical subject, which includes all the modifications that we discussed above. The rest of the subjects correspond to the same spatial maps but their time courses differ, depending on their respective HRF. Moreover, in Table I, the anatomical correspondence as well as the sparsity level of each source are also listed.

Besides the brain-induced sources, real fMRI data also contains sources that are either scanner-induced or related to other biological processes unrelated to the brain activity, such as heart-beating, breathing, movements, etc. All these phenomena are collectively referred to as *artifacts*. In the new dataset, we have also included realistic fMRI-like artifact sources, in particular the sources 3, 4, 5, 7 and 8 of the well-known dataset [7]. Special information concerning each artifact is also provided in Table I. Note that the majority of the brain-like sources are sparse, whereas most of the artifacts are dense. Finally, the subject-specific datasets are corrupted by Rician noise of SNR $= 0$. Both noise distribution and energy are realistic for the fMRI data case, [8], [9].

REFERENCES

[1] A. Protopapas, E. Orfanidou *et al.*, "Evaluating cognitive models of visual word recognition using fMRI: effects of lexical and sublexical variables," *Neuroimage*, vol. 128, pp. 328–341, 2016.

[2] G. K. Aguirre, E. Zarahn, and M. D'esposito, "The variability of human, BOLD hemodynamic responses," *Neuroimage*, vol. 8, no. 4, pp. 360–369, 1998.

[3] D. A. Handwerker, J. M. Ollinger, and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses," *NeuroImage*, vol. 21, no. 4, pp. 1639–1651, 2004.

[4] W. D. Penny, K. J. Friston *et al.*, *Statistical parametric mapping: the analysis of functional brain images*. Academic press, 2011.

[5] K. J. Friston, A. P. Holmes, and J. Ashburner, "Statistical parametric mapping (spm) (available at http://www.fil.ion.ucl.ac.uk/spm/)," 1999.

[6] E. B. Erhardt, E. A. Allen *et al.*, "SimTB, a simulation toolbox for fMRI data under a model of spatiotemporal separability," *Neuroimage*, vol. 59, pp. 4160–4167, 2012.

[7] N. Correa, T. Adalı *et al.*, "Comparison of blind source separation algorithms for FMRI using a new matlab toolbox: GIFT," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 5, 2005, pp. v–401.

[8] H. Gudbjartsson and S. Patz, "The rician distribution of noisy mri data," *Magnetic Resonance in Medicine*, vol. 34, no. 6, pp. 910–914, 1995.

[9] M. Welvaert and Y. Rosseel, "On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data," *PLOS ONE*, vol. 8, no. 11, p. e77089, 2013.