False discovery rate smoothing - Supplementary Information

A Details of fMRI data set

The fMRI data set analyzed in Section 3 was acquired and processed as follows. A spatial working memory localizer (Fedorenko et al., 2013) was performed by a single subject. On each trial, a 4x2 spatial grid is presented, and locations in that grid are presented sequentially (1000 ms per location), followed by a forced-choice probe between two grids, one of which contained all of the locations presented in the preceding series. In the easy condition, one location is presented on each presentation, whereas in the hard condition two locations are presented on each presentation. Twelve 32-second experimental blocks were interspersed with 4 16-second fixation blocks (acquisition time = 7:28). The contrast presented in Figure 1 compares the hard versus easy conditions.

fMRI acquisition was peformed using a multi-band EPI (MBEPI) sequence (Moeller et al., 2010) (TR=1.16 ms, TE = 30 ms, flip angle = 63 degrees, voxel size = 2.4 mm X 2.4 mm X 2 mm, distance factor=20%, 64 slices, oriented 30 degrees back from AC/PC, 96x96 matrix, 230 mm FOV, MB factor=4, 10:00 scan length). fMRI data were preprocessed according to a pipeline developed at Washington University, St. Louis (Power et al., 2014), including realignment for motion correction, distortion correction using a field map, and registration to a 3-mm isotropic atlas space. Preprocessed task fMRI data were analyzed at

the first level using the FSL Expert Analysis Tool (FEAT, version 5.0.6), using prewhitening and high-pass temporal filtering (100 second cutoff).

B Finding plateaus in 2D images

Algorithm 1 outlines our approach to finding plateaus, which is needed in our path-based algorithm for choosing λ. Note that each point in the grid is touched at most k times, where k is the number of neighbors of that point. Thus the algorithm runs in $\mathcal{O}(kn)$, which is effectively linear time since $k \ll n$. The algorithm is mildly sensitive to underlying numerical inaccuracies in the ADMM solution for β. It is well known that finite-precision ADMM solutions tend to slightly "round off" sharp edges in the underlying image. This produces some slight numerical noise in the degrees of freedom estimate. In our experience, this is rarely a practical concern, and can always be corrected by tightening the convergence criterion for ADMM below the plateau tolerance in Algorithm 1.

C Benchmark setup

As described in Section 5, all methods were run across a suite of scenarios, with 30 independent trials per scenario and a 10% FDR threshold. This appendix describes the method-specific settings for the two main competing methods: FDR_L and the HMRF model.

For FDR_L , we used "Method 1" from (Zhang et al., 2011) as this was suggested for fMRI-type data. We set the null-cutoff $\lambda = 0.2$. This is higher than used in (Zhang et al., 2011), which used $\lambda = 0.1$; however, they also used a 1% FDR threshold. Since λ controls the proportion above which we expect almost all p-values to be true nulls, using a λ of 0.2 is more reasonable with an FDR of 0.1. Preliminary experiments confirmed the FDR_L authors' claim that FDR_L is not very sensitive to the setting of λ.

The HMRF model has several tunable parameters and required tweaks to run the code

Algorithm 1 Our plateau-finding algorithm.

Input: grid of values β, plateau tolerance ϵ
Output: list of plateaus and their values ϕ

$tocheck \leftarrow coordinates(\beta)$
$checked \leftarrow \{\emptyset\}$
$\phi \leftarrow \{\emptyset\}$
while $tocheck$ not empty **do**
    $(x_0, y_0) \leftarrow$ pop $tocheck$ until $(x_0, y_0) \notin checked$
    $points \leftarrow \{(x_0, y_0)\}$
    $\beta_{min}, \beta_{max} \leftarrow \beta_{x_0,y_0} - \epsilon, \beta_{x_0,y_0} + \epsilon$
    $unchecked \leftarrow \{(x_0, y_0)\}$
    while $unchecked$ not empty **do**
        $(x, y) \leftarrow$ pop $unchecked$
        for each neighbor (v, w) of (x, y) **do**
            if then$(v, w) \notin checked$ and $\beta_{min} \leq \beta_{v,w} \leq \beta_{max}$
                Add (v, w) to $points$, $unchecked$, and $checked$
            end if
        end for
    end while
    Add $points$ to ϕ
end while
return ϕ

provided in the supplementary materials of (Shu et al., 2015):

- In order to compile the C++ code, we needed to change all calls to `floor(x)` with `(double((int)x))`.
- 2d grids and edge points are not supported in their implementation. To process the entire 128×128 grid, we had to embed it within the center of a $3 \times 130 \times 130$ array. This should have no effect on the result, as we specified the original lattice as the region of interest.
- The alternative density estimation procedure is parametric (as opposed to our non-parametric approach) and requires specifying the number of components in a Gaussian mixture model. We specify the correct number of components in each case, to give their model the best possible estimation (i.e. 2 for the well-separated scenarios and 1 for the poorly-separated scenarios).

- We ran with the default parameters of $sweep_b = 1000$, $sweep_r = 5000$, $sweep_lis = 1e6$, $iter_max = 5000$. These correspond to 5000 iterations of the main Gibbs sampler with a 1000-iteration burn-in. These settings are identical to those used in the HMRF paper.

We made every effort possible to be as generous as possible to both methods. This is the main reason for choosing to include the "saturated" signal regions, as these cases highlight the areas where FDR_L and HMRFs perform well, even though we expect them to be rare in practice, as evidenced by the various prior plateaus discovered by FDR smoothing in Figure 2.

D Comparisons with FDR Regression

Benchmark performance results against FDR regression (FDR-R) are presented in Table 1. We performed 100 independent trials for each of eight different scenarios, corresponding to two different plateau setups. In both setups, we used two plateaus of increased probability levels of 0.5 and 0.8; the "large regions" setup used plateaus of 40×40 and 60×60, whereas the "small regions" setup used plateaus of size 15×15 and 10×10. For each plateau setup, we tested the following four different alternative distributions:

1. $0.48N(-2, 1) + 0.04N(0, 16) + 0.48N(2, 1)$
2. $0.4N(-1.25, 2) + 0.2N(0, 4) + 0.4N(1.25, 2)$
3. $0.3N(0, 0.1) + 0.4N(0, 1) + 0.3N(0, 9)$
4. $0.2N(-3, 0.01) + 0.3N(-1.5, 0.01) + 0.3N(1.5, 0.01) + 0.2N(3, 0.01)$

FDR regression using a 100-dimensional b-spline basis comes close to the performance of FDR smoothing, but also has many conceptual and computational disadvantages. These

are essentially the same disadvantages that one would face in treating *any* spatial smoothing problem in a regression framework. For example, to handle a smoothing problem using FDR regression, one must choose the basis set and the number of basis elements. This is implicitly a choice about the smoothness of the underlying prior image, and is not straightforward in large problems or problems over an arbitrary graph structure. FDR smoothing, on the other hand, has no tunable parameters once our path-based method for choosing λ is used. Moreover, FDR regression cannot localize sharp edges in the underlying image of prior probabilities, unless these edges happen to coincide with any edges present in the basis set. FDR smoothing finds these edges automatically without requiring a clever choice of basis, and without having to tolerate undersmoothing in other parts of the image. Finally, at an algorithmic level, the important matrix operations in FDR smoothing involve very sparse matrices and benefit enormously from pre-caching. This is not true in FDR regression, which involves dense matrices and linear systems that change at every iteration.

As the table shows, FDR regression with basis functions does provide sensible answers and good FDR performance. However, the FDR smoothing approach benefits greatly by exploiting the spatial structure of the problem, resulting in better power and more interpretable summaries at lower computational cost.

E HMRF details and improvements

The HMRF model, while following the prior-dependence philosophy of FDR smoothing, makes a different distributional assumption on the dependence by placing an Ising model on the priors. This has two important side effects. First, the model is not necessarily going to discover constant regions of prior probability. This is clear when looking at the "local index of significance" (LIS) statistics produced by the HMRF, shown in Figure 1. While the LIS space is substantially smoothed, it is not constant across different plateaus like in FDR smoothing. The other core issue with the HMRF model is that its substantial

True positive rate (TPR)

	Large Regions				Small Regions			
	Alt 1	Alt 2	Alt 3	Alt 4	Alt 1	Alt 2	Alt 3	Alt 4
BH	0.364	0.215	0.128	0.366	0.212	0.123	0.090	0.194
2G	0.394	0.229	0.134	0.403	0.211	0.123	0.091	0.196
FDR-R	0.559	0.334	0.167	0.610	0.242	0.141	**0.097**	0.232
FDRS	**0.592**	**0.352**	**0.168**	**0.645**	**0.264**	**0.144**	0.093	**0.257**
Oracle	0.688	0.524	0.332	0.718	0.298	0.193	0.139	0.292

False discovery rate (FDR)

	Large Regions				Small Regions			
	Alt 1	Alt 2	Alt 3	Alt 4	Alt 1	Alt 2	Alt 3	Alt 4
BH	0.072	0.070	0.073	0.070	0.090	0.093	0.093	0.092
2G	0.089	0.083	0.083	0.089	0.092	0.096	0.098	0.096
FDR-R	0.075	0.058	0.050	0.086	0.102	0.106	0.109	0.105
FDRS	0.072	0.057	0.054	0.079	0.092	0.095	0.098	0.096
Oracle	0.101	0.100	0.100	0.101	0.097	0.101	0.101	0.098

Table 1: Results of the eight simulation studies. Each entry is an average error rate across 100 simulated data sets. FDR smoothing (FDRS) results in the highest true-positive rate for all but one of the scenarios, consistently beating both the Benjamini–Hochberg procedure (BH) and the two-groups model (2G). FDR regression (FDR-R) comes close, but slightly overshoots the desired FDR limit of 10% in the small-signal examples. (Scott et al., 2014) also report this behavior. In contrast, FDR smoothing remains (on average) under the nominal FDR across all experiments.

complexity results in a very difficult model to fit. The implementation provided by the authors performs an EM algorithm with Gibbs sampling and required more than three days to run the examples with the suggested number of iterations, compared to minutes with FDR smoothing on the same examples and the same compute cluster. More to the point, the final fit shows clear bias to local optima that over-estimate the strength of the signal region. The result is a model which performs well only when the regions are clearly segmented and the signal region is saturated, and which otherwise fails to adhere to the specified FDR threshold. See Appendix E for more details on the HMRF model, its configuration, and suggestions from the HRMF authors on ways to improve the runtime and fit of the model; we did not incorporate these suggestions in our benchmarks as they were either purely computational speedups or were intuitive suggestions that would require developing entirely new methods.

In an effort to provide fair evaluation, we contacted the first and second authors of the HMRF paper (Shu et al., 2015). They provided several suggestions for improving the speed of the algorithm and its performance. The following speedup suggestions were offered:

- Reduce the number of burn-in iterations.
- Monitor the stability of the parameter estimations in order to stop earlier than the maximum number of iterations.
- Stop the backtracking line search at a fixed number of steps in the Newton's method step.
- Use an updated pseudo-random number generator as the code relies on an outdated generator which may be slower than the most up-to-date version.

Note that all of the above suggestions would reduce the running time of the algorithm, but would not likely result in an improved fit or better performance on the benchmarks. The main performance improvement suggested was to preprocess the z-scores so as to detect the

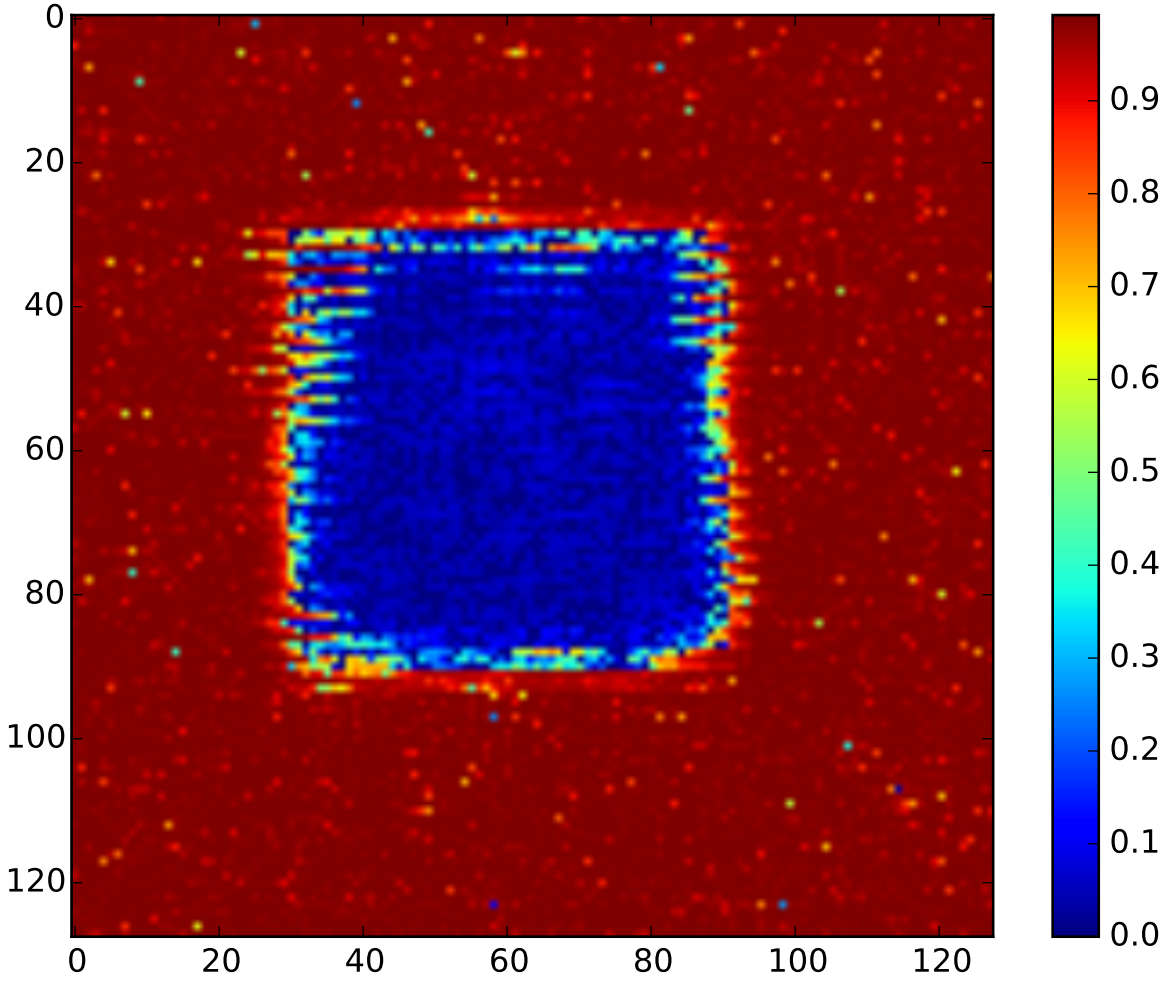

Figure 1: The inferred "local index of significance" statistics inferred by the HMRF model on the example in Figure 7. The Ising model assumption, combined with the difficult-to-fit distribution it induces, results in a model that overestimates the strength of the signal region.

different regions first, then run separate HMRFs on each region. One way to do this would be to run FDR smoothing, then treat each plateau as a different region and fit an HMRF to them. It is unclear whether this approach would truly address the underlying issues we observed in the benchmarks. Thus, while this is an interesting idea and may be effective, it would constitute an entirely new method and therefore we leave it to future work.

F Correlated noise example with large bandwidth

Figure 2 shows an example of a dataset from the experiment in Section 5.3. The highly correlated noise creates clear regions of false positives that are difficult to distinguish from the true positive regions. Specifically, the bottom left panel ("True Discoveries") shows all true positives, whereas the bottom right panel ("Estimated Discoveries") shows the discoveries reported by the FDR smoothing algorithm. The algorithm reports many false positives in the lower-right area of the image due to the highly correlated noise that produces clumped outliers in the observed data (second row, right column). Without prior knowledge of the correlation of the noise, it becomes virtually impossible to separate true signals from grouped outliers.

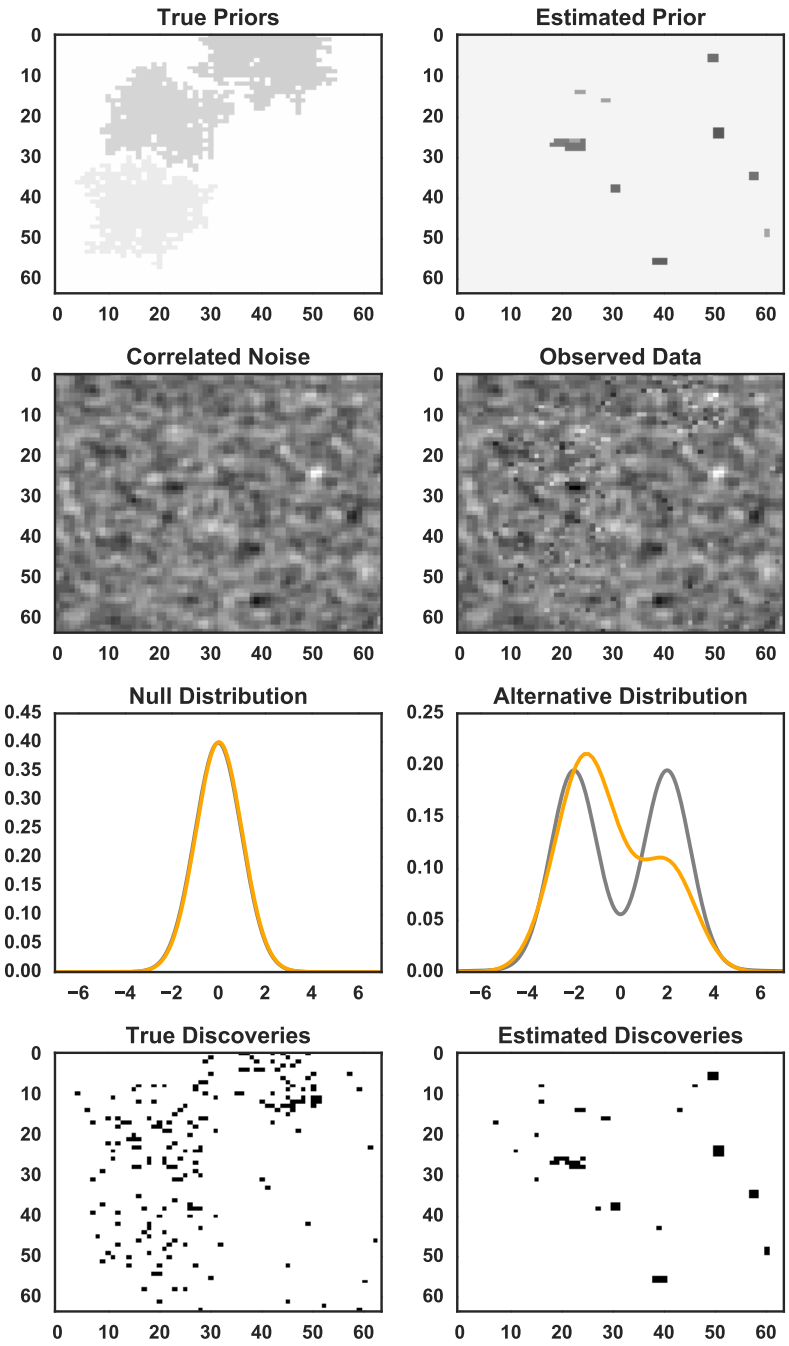

Figure 2: An example of a dataset generated with a bandwidth just greater than 1. The left figure in the second row shows the highly-correlated noise added to the model. The corresponding right figure shows the resulting data that the model is given, with clear examples of phantom plateaus.

References

E. Fedorenko, J. Duncan, and N. Kanwisher. Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci U S A*, 110(41):16616–21, Oct 2013. doi: 10.1073/pnas.1315235110.

S. Moeller, E. Yacoub, C. A. Olman, E. Auerbach, J. Strupp, N. Harel, and K. Uğurbil. Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magn Reson Med*, 63(5):1144–53, May 2010. doi: 10.1002/mrm.22361.

J. D. Power, A. Mitra, T. O. Laumann, A. Z. Snyder, B. L. Schlaggar, and S. E. Petersen. Methods to detect, characterize, and remove motion artifact in resting state fmri. *Neuroimage*, 84:320–41, Jan 2014. doi: 10.1016/j.neuroimage.2013.08.048.

J. G. Scott, R. C. Kelly, M. A. Smith, P. Zhou, and R. E. Kass. False discovery-rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 2014. to appear.

H. Shu, B. Nan, and R. Koeppe. Multiple testing for neuroimaging via hidden Markov random field. *Biometrics*, 2015.

C. Zhang, J. Fan, and T. Yu. Multiple testing via FDRL for large scale imaging data. *Annals of statistics*, 39(1):613, 2011.