Application of Deep Knockoffs for fMRI to Generate Surrogate Data

Alec Flowers 321786, Alexander Glavackij 322968, Janet van der Graaf 327759
Machine Learning Course CS-433, EPFL, Switzerland

***Abstract*—The study of neural structure and activity is a promising field for a better understanding of our brain, as well as for diagnostics of brain disorders. Functional Magnetic Resonance Imaging (fMRI) is one of the techniques used to detect active brain regions while performing tasks, i.e. selecting significant regions. Recently, a new procedure to select significant features, called Knockoff filtering, gained attention. In this paper, we apply the knockoff framework to generate surrogate data and perform non-parametric tests using the surrogate data to build brain activation maps. We implement three different knockoff frameworks, evaluate their quality, and compare the resulting activation maps to the ones produced using parametric testing from a general linear model (GLM), the standard method to generate activation maps. Our results show that the activation maps built with the knockoff framework are comparable to the ones created by the GLM. These results open the door for further investigations of the knockoff framework on fMRI data.**

I. Introduction

Neuroscience, the study of brain structure and its functional relations is key to the understanding of brain disorders. A common technique used in this field is functional magnetic resonance imaging (fMRI), which enables mapping neural activity of the brain [1] to real-word tasks. The activation of certain brain regions for different tasks - for example the activation of the motor cortex when moving an arm - is assessed by measuring changes of blood-oxygenation-level-dependant (BOLD) response over time, while a subject is performing a certain task [2]. An increase in blood oxygenation in a brain region indicates an increase in metabolic requirements, and thus, neuron activation. The result of an experiment is a time-series for every voxel in the brain. Figure 1 shows an example BOLD response. The measurements are subject to multiple errors, e.g. measurement noise and we turn to statistics to help us interpret the BOLD signal. The task of finding the brain regions which exhibit significant activation is a feature selection problem, which results in a statistical activation map (SAM). Traditionally, statistical significance is examined by testing whether the null-hypothesis H_0 is true. H_0 is rejected if the probability of it being true is lower than the threshold α. Adjusting α controls the False Discovery Rate (FDR), the amount of features which have been falsely identified as significant. The parametric statistical tests are predicated on assumptions about the probability distribution of the error. While sometimes appropriate, there are times these assumptions fail or cannot be validated. Recently, a procedure called *Knockoff Filtering* [3] has been proposed to identify significant features while controlling the FDR, without making assumptions about the distribution of the measurement error i.e. in a non-parametric way. This framework gives strong statistical guarantees while preserving predictive power. It has been successfully applied to several feature selection problems - identifying genetic polymorphisms associated with diseases [4], selection of bio-markers for cancer [5], and to identify bacterial strains [6].

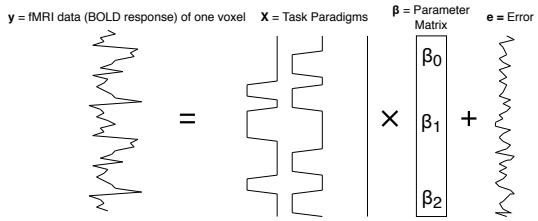


Fig. 1: Basic principle of the GLM in fMRI. The GLM finds the parameters β_i, given the regressors $\mathbf{X}$, which best explain the fMRI data $\mathbf{Y}$ by minimizing the error $\mathbf{e}$ [2].

In this work we apply the Knockoff framework to fMRI data to generate activation maps, i.e. find brain regions which are significantly active in specific tasks. To do this, in Section II, we lay the technical background of this work. It includes the general linear model (GLM) method, the parametric method to generate activation maps, a sketch of the Knockoff framework, and a primer on non-parametric thresholding. Afterwards, in Section III, we present our core contribution: we implement the Knockoff framework on fMRI data and use non-parametric significance statistics to generate activation maps. In Section IV, we discuss our attempts to improve knockoff generation and compare the resulting activation maps generated by the Knockoff framework to the activation maps generated by the GLM. Lastly, in Section V, we draw a conclusion and give an outlook on future work.

II. Background

A. General Linear Model

The GLM encompasses a class of statistical methods which assume that in task-based fMRI experiments the fMRI data is composed of a linear combination of model factors and uncorrelated noise [2]. The model factors include *Task Paradigms*, cf. Figure 1, which indicate the time-points a subject performed a certain task. To properly model the BOLD response, the task paradigms are first convolved with a hemodynamic

response function. The convolved task paradigms are then used to reconstruct the BOLD response with a regression analysis:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}. \tag{1}$$

The $\beta_i \in \boldsymbol{\beta}$ indicate the influence of each task paradigm (X) on the overall fMRI data. Each factor is then evaluated for statistical significance, for example via the student's t-test. The result of this analysis is an activation matrix $A \in \{0,1\}^{v \times t}$, where t denotes the number of task paradigms and v denotes the number of voxels. Thus, a voxel in the fMRI data is considered to be active during a task if the β_i of the task paradigm was significant in the GLM. To control for the multiplicity of testing over multiple voxels, we apply the Bonferroni correction which sets the significance cutoff more strictly at $\alpha_{Bonferroni} = \alpha \,/\, v$.

B. Knockoffs

Feature selection refers to the process of selecting a subset of relevant features X_s which best explain a response variable Y. More formally, this is $Y \perp\!\!\!\perp X_s \subseteq \mathbf{X} \in \mathbb{R}^{n \times p}$. A feature $X_j,\ j \in \{1, \dots, p\}$ is assumed to be unimportant if $Y \perp\!\!\!\perp X_j | (X \setminus X_j)$, i.e. if X_j is conditionally independent of Y once $(X \setminus X_j)$ is known. A random vector $\widetilde{X} \in \mathbb{R}^p$ is a knockoff copy for any feature $X \in \mathbb{R}^p$ sampled from distribution P_x, if the joint law of $(X, \widetilde{X})$ obeys

$$(X, \widetilde{X}) \overset{d}{=} (\widetilde{X}, X)_{swap(j)} \text{ for each } j \in \{1, \dots, p\}. \tag{2}$$

The symbol $\overset{d}{=}$ indicates equality in distribution, $(\cdot)_{swap(j)}$ is defined as the operator swapping $\widetilde{X}_j$ with X_j. Swapping is only allowed, if the underlying probability distribution is not changed by the swap. Since the knockoff features are not related to the outcome Y, they can be used as negative controls for the "real" features. There exist multiple methods to use the knockoffs as negative controls. In [4], the authors use a lasso regression on the original data set expanded by the knockoffs for each original feature. The significant features are then determined by comparing the regression coefficient of the original features and the knockoff features. This approach requires the data X and the corresponding knockoffs $\widetilde{X}$ to be the predictors in the regression model. However, in the GLM, the fMRI data is the predicted value Y, which is regressed on by the task paradigms. Thus, we cannot employ the methodology used in [4]. As a result, we employ another non-parametric test method to test for significance using the knockoff features.

C. Non-Parametric Permutation Tests

In situations where parametric assumptions do not hold, a non-parametric approach is the only valid method of analysis. [7]. We take inspiration from the permutation tests laid out in [7]. Instead of assuming a known null distribution, we generate one by creating N surrogates of fMRI time courses and computing functional activation values for these. We compare the empirical values of the activation against the surrogate distribution to calculate p-values. By conducting hypothesis testing at each voxel we produce a p-value, p_v, for $v \in \{1, \dots, V\}$ voxels. If we do not take into account the multiplicity of testing, the Type-1 error is not controlled over the entire image as 5% of voxels are expected to reject the null by design. This is the multiple comparison problem and we call these *uncorrected p-values*. To correct for this we turn to the single threshold test that provides strong control over image-wise type 1 error through consideration of a maximal voxel statistic. Let T_v be the statistic computed on the empirical data for each voxel. Let t_i^{max} be the maximum voxel statistic for $i \in \{1, \dots, N+1\}$ over the empirical data and N surrogates. We calculate the "permutation" distribution of the maximal statistic denoted T_i^{max} over the entire image and compute *corrected p-values*

$$p_v = \frac{\#\{i : T_i^{max} \geq T_v\}}{N+1} \tag{3}$$

[7] If p_v is less than critical value α we reject H_0.

III. Methodology

A. fMRI Data Set

The data set we use in this work was made available by the Medical Image Processing Laboratory (MIPLAB) jointly run by the EPFL and University of Geneva [8]. It comprises fMRI data (BOLD response) collected from 100 subjects over seven different types of tasks (motor-, emotional-, gambling-, relational-, language-, social-, and working memory tasks), each task with multiple conditions. For the sake of presentation in this paper, we limit ourselves to the motor task, however, the pipeline we develop in the two following subsections can be applied to all provided tasks. The motor task data is a matrix $Y_{s \times p \times n}$, where s denotes the number of subjects, p the number of regions and n the number of timepoints. In this data set, voxels v have already been mapped onto related regions p according to [9]. The dimensions of the motor task data is as follows: $s = 100$, $p = 379$, and $n = 284$. Mapping these values back to the notation used in Section II-B, we interpret a brain region as a feature p_i in our data set and a value of a time-series of feature p_i as a sample n_{i,p_i}. The motor task data was collected with five different task paradigms: movement of the left and right hand, the left and right foot, and tongue. Thus, the GLM would regress five β_i for every region p.

B. Generating Knockoffs

In the remaining parts of this work, we refer to the data set as Y instead of X and the knockoff copies as $\widetilde{Y}$ instead of $\widetilde{X}$, since the fMRI data is the Y being reconstructed by the GLM. A copy of feature Y_i is a knockoff copy, if the exchange property (2) is fulfilled, i.e. the distribution d is not changed by swapping the real feature Y_i with the knockoff feature $\widetilde{Y}_i$. Thus, a knockoff generator is required to generate features which exhibit the same distributional properties as the original features. This can be achieved through different methods, in this work we compare three: two types of Gaussian knockoffs (GKO) and Deep Knockoffs (DKO).

The GKO generator tries to generate knockoffs which have the same mean μ_i and co-variance matrix Σ as the original features, thus, it matches the first two moments. Knockoffs are then generated by sampling values from a gaussian distribution with μ_i and Σ. A GKO generator does not match the third and fourth moments, however, if the real features are distributed gaussian the exchange property 2 still holds. Another problems arises with GKOs: if the number of features p is large, the eigenvalues of Σ tend to get very small. This leads to numerical instabilities and the knockoffs having no statistical power, i.e. being highly correlated to the original features [10]. Indeed, calculating the eigenvalues of Σ for our data set with $p = 379$ leads to eigenvalues in the range of $\lambda \sim 1 \times 10^{-15}$.

In [10], the authors introduce a workaround: instead of generating knockoffs which match the full Σ, they choose a Σ_{approx} and generate knockoffs which match Σ_{approx}. To calculate Σ_{approx}, they cluster the features using estimated correlations as a similarity measure with single-linkage clustering. Σ_{approx} is then calculated on the representatives of the resulting clusters. As a result, the maximum correlation between representatives is reduced, and thus, eigenvalues are larger and statistical power is increased. This implies that a knockoff for the representative of a cluster is automatically the knockoff for every feature in that cluster. In [11], the authors present another method to calculate Σ_{approx}: they computing a low rank factor model of the covariance matrix. We proceed with both methods: we refer to the GKOs constructed as in [10] as Clustered Gaussian knockoffs (CGKOs) and the GKOs constructed as in [11] as Low Rank Gaussian knockoffs (LGKOs).

In contrast to a GKO generator, a DKO generator can match higher moments. Consequently, if the underlying distribution of the real features is not gaussian, the DKO should produce better knockoffs, i.e. knockoffs whose distribution is closer to the original distribution. A DKO generator uses a neural network to generate knockoffs. The network takes a vector of the original data set $y_{1\times p} \in Y_{n\times p}(\mathbb{R})$ and a random noise vector $v_{1\times p}$ as inputs. The output is a vector containing a knockoff for every feature $\widetilde{y}_{1\times p}$ The loss function the network uses is the difference between the distribution of the knockoff features and the original features. Further details on DKOs can be found in [4]. In the following, we proceed with the following three knockoff generators: CGKO, LGKO, and DKO.

C. Generating Activation Maps

Let $\mathcal{G}(Y) = \widetilde{Y}$ denote a knockoff generator which outputs a knockoff $\widetilde{Y}_{n\times p}$ given an input $Y_{n\times p}$. We use $\mathcal{G}$ to generate l knockoffs $K = \{\widetilde{Y}_1, \ldots, \widetilde{Y}_l\}$. We use these l knockoffs for our permutation distribution. As noted in Section II-C they have no relationship to the task paradigms and meet the weak non-parametric test criteria of having the same distribution as the empirical data. Afterward, we apply the GLM on every $\widetilde{Y} \in K$ and use the resulting beta values as our statistic which results in $t\times p\times l$ beta-values. With the surrogate and empirical beta-values we calculate T_i^{max}, T_i^{min}, and create *corrected p-values* according to Formula 3 for the empirical betas. Note that we are performing a two-sided t-test thresholding with an $\alpha = 0.025$ since we are interested in both positive and negative beta values. We visualize our results in brain plots and compare our thresholds to the standard.

	Pre-process	**Pre-Process Parameters**	**Model Hyperparameters**
LRKO	Low Rank Approx.	$rank = 120$ $shrink$ = True	None
CGKO	Dendogram Clustering	$max\_corr = 0.4$	None
DKO	Dendogram Clustering	$max\_corr = 0.4$	$\gamma = 1.0$, $\lambda = 0.1$, $\delta = 0.1$, $lr = 0.01, epochs = 1000$ $batch\_size = 0.5 \cdot n$

Table I: *rank* refers to the rank of the low rank factor model, *shrink* is a boolean whether to use Ledoit-Wolf shrinkage, *max_corr* specifies the maximum correlation to cut the dendograms, γ is a penalty that encourages matching of higher moments. λ is a second order penalty to encourage second order knockoffs. δ is a de-correlation penalty to make the knockoffs as different as possible to preserve statistical power. All generators were trained on one subject.

IV. Results

Our main goal was to find the generator which produces the best knockoffs. We tuned each knockoff generator and display the diagnostics with the optimal hyperparameters in Table I. We focused on trying to improve the DKO generator. We first tuned the hyperparameters. Next we tried appending multiple subjects together and use that as our training data. Furthermore, we changed the batching procedure to respect the time-series order. Appending the data of multiple subjects would return great diagnostics for the multi-subject knockoff but poor diagnostics for individual subjects. The DKO for 1000 epochs would take 1 hour to train on a Tesla P4 GPU provided by Google Colab. In the end, the results were inconclusive and we leave this open for future work.

Figure 2 shows the diagnostics presented in [4] for the CGKO, LGKO and DKO generators. These diagnostics quan-

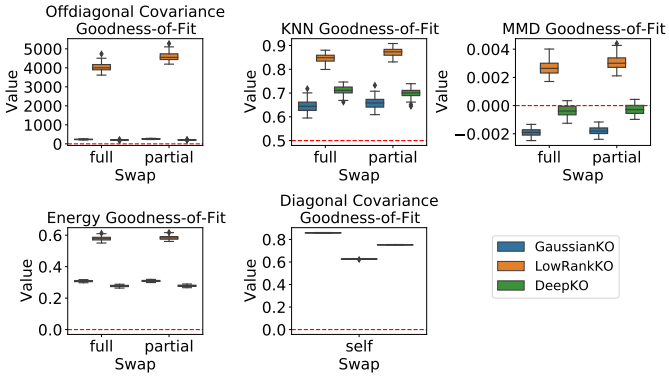


Fig. 2: Diagnostics of the three methods for knockoff generation (CGKO, LGKO, DKO) [4]. The red dashed lines indicate the ideal values for every diagnostic.

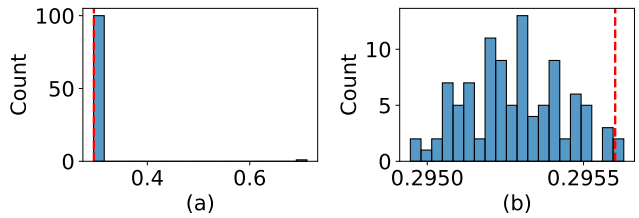


Fig. 3: Task 'motor', subject 1, condition 1. T_i^{max} distribution for DKO. (note this is just one side of the hypothesis test) (a) Shows the entire distribution including the t_i^{max} from the empirical betas. (b) We drop the t_i^{max} that is calculated on the empirical betas to zoom in on the distribution of the knockoff t_i^{max}. The red dashed lines indicate the critical threshold value.

tify the goodness-of-fit (GOF) of the conditional model producing knockoffs $\tilde{Y}$ by measuring the compatibility of the joint distribution of $(Y, \tilde{Y})$ with the exchange property 2. These diagnostics verify whether the two-sample problem $P_Y = P_{\tilde{Y}}$ holds true, and thus can be used as a quality measure of the knockoffs. In [4], the authors present the four diagnostics.

The DKOs show the best off-diagonal covariance, MMD, and energy GOF. The CGKOs display the best 1-NN GOF, the other diagnostics very close to the DKOs. The LGKO are the worst performing knockoffs, they exhibit the worst diagnostics across the board, except for the diagonal covariance GOF. We conclude that the DKOs exhibit the best diagnostics.

After evaluating the three presented knockoff generators, we use them to generate activation maps. We use the process detailed in III-C and look into distribution of T_i^{max} (cf. Figure 3) for the DKO. The 100 knockoffs produce t_i^{max} (these results are reflected for T_i^{min}) that are very tightly clustered together, for this one example the values are within .005 of each other. When we inspect the 100 knockoffs we find that they look very similar to one another with values deviating by only a small amount.

As there is no-ground truth in fMRI activation data, we compare our thresholding technique with the current standard of GLM. The goal was not to match the GLM, but rather compare the two methods. The confusion matrix is calculated over 1 subject and all 5 task paradigms, i.e. $379 \times 5 = 1895$ total values. For both the GKO and DKO there is a clustering pre-processing step where regions with very high correlations are put into a group and a candidate from this group is chosen. If a candidate from the group is thresholded, then the entire group is considered significant. We were using correlation as a proxy for similarity, however, even highly correlated signals may have very different beta values and we postulate this could be one source of deviation between the DKO, CGKO and GLM. The LRKO did not have this pre-processing step and more closely matches the GLM but is much stricter and only thresholds 18 values compared to the 86 of the GLM.

V. Conclusion

In this paper we laid the groundwork for applying the non-parametric technique of knockoff filtering to build statistical activation maps for fMRI data. This domain provided the unique challenge of time-series data that is both temporally and spatially correlated. We have tuned and analyzed three different knockoff generating methods and compared the results to the current best practice. According to the diagnostics we presented, the deep knockoffs are the best knockoff copies. The statistical parametric maps generated with the knockoff approach are comparable to the ones generated by the GLM. Future work could combine fMRI data of multiple subjects and generate knockoffs for all at once. Furthermore, future research could also use the fact that fMRI data is correlated in time and explore different neural network to generate time-series knockoffs.

CGKO

GLM \ CGKO	**0**	**1**
0	1788	21
1	78	8

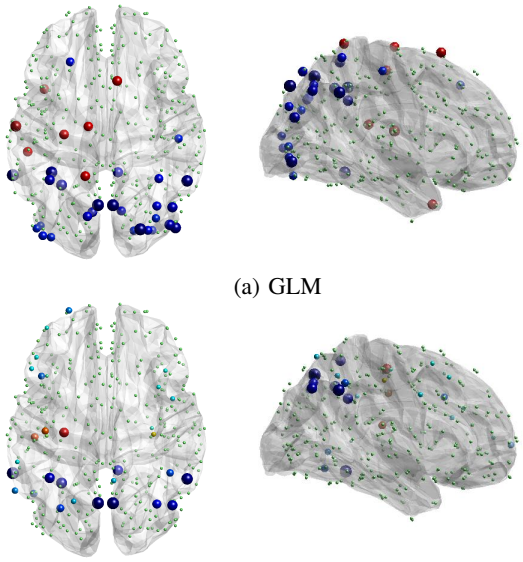
LRKO

GLM \ LRKO	**0**	**1**
0	1807	2
1	70	16

DKO

GLM \ DKO	**0**	**1**
0	1719	90
1	69	17

Table II: Task 'motor', subject 1. Confusion matrices for the three types of knockoff generators: CGKO, LRKO and DKO; where the true value is taken as the results of the GLM. 1 corresponds to an active region and 0 to a non-active region of the brain.



(a) GLM

(b) DKO

Fig. 4: Task 'motor', subject 1, condition 1. Examples of brain plots of a single subject obtained by (a) GLM and (b) thresholded activation map after non-parametric tests using surrogate data obtained by the DKO generator.

References

[1] B. R. Buchbinder, "Chapter 4 - functional magnetic resonance imaging," in *Neuroimaging Part I*, ser. Handbook of Clinical Neurology, J. C. Masdeu and R. G. Gonzàlez, Eds., vol. 135, Elsevier, 2016, pp. 61–92. DOI: https://doi.org/10.1016/B978-0-444-53485-9.00004-0.

[2] S. Huettel, A. Song, and G. McCarthy, *Functional Magnetic Resonance Imaging*, ser. Functional Magnetic Resonance Imaging v. 1. Sinauer Associates, 2004, ISBN: 9780878932887.

[3] R. F. Barber and E. J. Candès, "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, vol. 43, no. 5, pp. 2055–2085, 2015, ISSN: 00905364. [Online]. Available: http://www.jstor.org/stable/43818570.

[4] Y. Romano, M. Sesia, and E. Candès, "Deep knockoffs," *Journal of the American Statistical Association*, vol. 0, no. 0, pp. 1–12, 2019. DOI: 10.1080/01621459.2019.1660174.

[5] A. Shen, H. Fu, K. He, and H. Jiang, "False discovery rate control in cancer biomarker selection using knockoffs," *Cancers 11*, vol. 744, no. 6, 2019, ISSN: 2072-6694. DOI: 10.3390/cancers11060744.

[6] C. Chia, M. Sesia, C.-S. Ho, S. Jeffrey, J. Dionne, E. Candès, and R. Howe, "Interpretable signal analysis with knockoffs enhances classification of bacterial raman spectra," Jun. 2020.

[7] T. E. Nichols and A. P. Holmes, "Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples," Tech. Rep., 2001.

[8] EPFL. (2020). Medical image processing laboratory, [Online]. Available: https://miplab.epfl.ch/ (visited on 12/13/2012).

[9] M. F. Glasser, T. S. Coalson, E. C. Robinson, C. D. Hacker, J. Harwell, E. Yacoub, K. Ugurbil, J. Andersson, C. F. Beckmann, M. Jenkinson, S. M. Smith, and D. C. Van Essen, "A multi-modal parcellation of human cerebral cortex," *Nature*, 2016. DOI: 10.1038/nature18933. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/27437579.

[10] E. Candès, Y. Fan, L. Janson, and J. Lv, "Panning for gold: Model-x knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 3, pp. 551–577, 2018. DOI: https://doi.org/10.1111/rssb.12265.

[11] A. Askari, Q. Rebjock, A. d'Aspremont, and L. E. Ghaoui, "FANOK: knockoffs in linear time," *CoRR*, vol. abs/2006.08790, 2020. arXiv: 2006.08790. [Online]. Available: https://arxiv.org/abs/2006.08790.