NeuroImage 245 (2021) 118686

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

The unreliable influence of multivariate noise normalization on the reliability of neural dissimilarity

J. Brendan Ritchie[1,*], Haemy Lee Masson[2], Stefania Bracci[3], Hans P. Op de Beeck[1]

[1] Department of Brain and Cognition, Leuven Brain Institute, KU Leuven, 3000 Leuven, Flemish Brabant, Belgium
[2] Department of Cognitive Science, Johns Hopkins University, Baltimore, USA
[3] Centre for Mind/Brain Sciences, University of Trento, Rovereto, Italy

ABSTRACT

Representational similarity analysis (RSA) is a key element in the multivariate pattern analysis toolkit. The central construct of the method is the representational dissimilarity matrix (RDM), which can be generated for datasets from different modalities (neuroimaging, behavior, and computational models) and directly correlated in order to evaluate their second-order similarity. Given the inherent noisiness of neuroimaging signals it is important to evaluate the reliability of neuroimaging RDMs in order to determine whether these comparisons are meaningful. Recently, multivariate noise normalization (NN_M) has been proposed as a widely applicable method for boosting signal estimates for RSA, regardless of choice of dissimilarity metrics, based on evidence that the analysis improves the within-subject reliability of RDMs (Guggenmos et al. 2018; Walther et al. 2016). We revisited this issue with three fMRI datasets and evaluated the impact of NN_M on within- and between-subject reliability and RSA effect sizes using multiple dissimilarity metrics. We also assessed its impact across regions of interest from the same dataset, its interaction with spatial smoothing, and compared it to GLMdenoise, which has also been proposed as a method that improves signal estimates for RSA (Charest et al. 2018). We found that across these tests the impact of NN_M was highly variable, as also seems to be the case for other analysis choices. Overall, we suggest being conservative before adding steps and complexities to the (pre)processing pipeline for RSA.

1. Introduction

> Primum non nocere
> – Latin translation of saying (“first, do no harm”) attributed to Hippocrates (c. 460–370 bce).

Representational similarity analysis (RSA) has become a staple of the multivariate pattern analysis (MVPA) toolkit in cognitive neuroscience. Methodologically, the core construct of the approach, representational dissimilarity matrices (RDMs), provide a common and straightforward format for summarizing and directly comparing datasets from different types of modalities to evaluate their second-order similarities. By converting multivariate signals in condition-rich experiments into RDMs, neural data acquired with fMRI, EEG/MEG, or cellular recordings can be directly compared both to each other and also to RDMs derived from behavioral judgments and computational models (Kriegeskorte, Mur, Bandettini, 2008a). Theoretically, by focusing attention on the “representational geometry” of multivariate datasets (Kriegeskorte and Kievet, 2013), RSA has its roots in the long tradition of psychological theories and methods that characterize the relationship between mental representations in terms of similarity structure (Attneave, 1950; Shepard, 1964). Thus, not only does RSA provide a unified analytic framework for formating and comparing datasets; it also promises a means for bridging the gap between psychological constructs and their neural implementation.

As with any neuroimaging method, the viability of RSA is constrained by data quality. Although all neuroimaging data is noisy, for RSA the issue is especially pressing in light of the massive number of comparisons that are sometimes necessary to construct an RDM. For example, the classic study of Kriegeskorte et al. (2008b) included 92 image conditions, which requires calculating 4186 pairwise neural dissimilarity values to construct a single RDM. In principle, one approach to determining the reliability of an RDM would be to evaluate the stability of each of these comparisons individually (Bobadilla-Suarez et al. 2019; Ritchie and Op de Beeck, 2019). However, in practice, researchers have focused on more global properties of RDMs when evaluating their within- and between-subject reliability. For example, since different samples for the same condition should be more similar than samples from different conditions, if activity patterns are reliable, a common procedure is to calculate all pairwise similarity correlations between independent splits of the data. If the on-diagonal correlations, reflecting self-similarity of conditions, are greater than the off-diagonal values, this suggests that on average the conditions can be differentiated (Haxby et al. 2001; Nili et al. 2020; Ritchie, Bracci, and Op de Beeck, 2017). However, since it is the off-diagonal values that are compared when carrying out RSA, another common method is to estimate between-subject reliability of these values by using a leave-one-subject-out procedure: RDMs for all but one subject are averaged and correlated with that of the remaining subject. The average across folds then gives a point estimate of the “noise ceiling”; that is, an upper bound of

[*] Corresponding Author: J. Brendan Ritchie, Psychologisch Instituut, Tiensestraat 102 – box 3714, 3000 Leuven, Belgium.
E-mail address: j.brendan.w.ritchie@gmail.com (J.B. Ritchie).

https://doi.org/10.1016/j.neuroimage.2021.118686.
Received 28 June 2021; Received in revised form 21 October 2021; Accepted 26 October 2021
Available online 31 October 2021.

how much another RDM can on average correlate with individual neural RDMs (Nili et al. 2014).

The importance of reporting the reliability of neural dissimilarity has naturally led to proposals for how it can be improved. One major focus has been on the choice of dissimilarity metric, with different distance metrics having been proposed as an alternative to the standard 1 – *r* correlation distances (Allefeld and Haynes, 2014; Guggenmos, Sterzer and Cichy, 2018; Nili et al. 2014; Nili et al. 2020). Walther et al. (2016) go a step further, proposing not only cross-validated Mahalanobis distance as a superior dissimilarity metric, but also recommending that within-subject reliability of neural RDMs can be improved by multivariate noise normalization (NN_M). In standard fMRI pipelines, a GLM is fit to the BOLD signal of individual voxels and the activity patterns that are analyzed with RSA are the beta estimates of this model. However, the BOLD signal can be influenced by many sources of noise, which are not captured by the GLM. Some of these sources of noise may have a spatial component (Friston, Jezzard, and Turner, 1994), in which case, it is possible that estimating the structure intrinsic to the noise can be used to improve the estimates of the signals of interest, and thereby improve the reliability of the data used to construct neural RDMs. NN_M offers a method for improving the reliability of fMRI data by improving the estimate of beta values based on voxel noise using information gleaned from the residuals of the GLM that is standardly used in first level analysis (Walther et al. 2016). More specifically, NN_M normalizes the beta weights by the covariance of the run-specific noise, in contrast to univariate noise normalization, which uses only the variance (Misaki et al. 2010).

Through both simulation and reanalysis of four datasets, Walther et al. found that NN_M improves the split-half within-subject reliability of the off-diagonal values of neural RDMs regardless of the choice of dissimilarity metric. Applying the NN_M approach to MEG data, Guggenmos, Sterzer and Cichy (2018) also found a marked improvement. However, not all results have been positive, running counter to the results of these studies. Charest, Kriegeskorte and Kay (2018) compared the effects of NN_M to those obtained with GLMdenoise, which estimates the number of noise predictors using a data-driven cross-validation procedure. Contrary to the previous studies, they found that NN_M in fact made reliability worse, and only revealed a benefit when combined with the noise estimates derived from GLMdenoise. Even more concerning, Liu et al. (2021) found that, after carrying out NN, an observed interaction in representational dissimilarity for adjacent fingers and age was no longer significant. Taken as a whole, these studies suggest there is at present equivocal support for the effectiveness of NN_M for improving the reliability of neural RDMs.

In the present study we revisited the issue of the effectiveness of NN_M at improving the reliability of neural dissimilarity estimates regardless of the choice of dissimilarity metric. First, we attempted to replicate the findings of Walther et al. with three fMRI datasets from previous studies (Bracci and Op de Beeck, 2016; Lee Masson et al. 2018; Ritchie and Op de Beeck, 2019). Second, unlike the previous work on NN_M, we did not restrict ourselves to sensory-motor ROIs or a single ROI per dataset. Third, we also evaluate the impact of NN_M on both between-subject reliability, or the noise ceiling, and RSA effect sizes. Fourth, we compared the effect of NN_M with and without spatial smoothing, which has a minor positive effect on RSA, at least when using the 1 - *r* correlation as the dissimilarity metric (Hendriks et al. 2017; Op de Beeck, 2010). Finally, we attempted to reproduce some of the findings of Charest et al. by comparing results with NN_M to those obtained with GLMdenoise or NN_M when using the noise estimates of GLMdenoise.

2. Materials and methods

2.1. Datasets

We reanalyzed fMRI datasets from three previously published studies, described below. In each case, the experiments were approved by the ethics committee of UZ/KU Leuven and all methods were performed in accordance with the relevant guidelines and regulations. All studies were carried out a 3T Phillips scanner with a 32-channel coil at the Department of Radiology of UZ Leuven. MRI volumes were collected echo planar (EPI) T2*-weighted scans with virtually identical parameters (Table 1). Preprocessing, including slice time correction and motion correction, was carried out with SPM8 or 12. First level analysis was also carried out with SPM and the GLM included all of the stimulus conditions as well as six motion correction parameters (translation and rotation in the x,y, and z axes). We note that in a standardly constructed GLM, although a single design matrix is used to model all the data, beta estimates are specific to each run and independent of each other. Further differences in the preprocessing of the images are noted below, and full details of the analysis pipelines can be found in the original studies. Stimuli of the datasets, and the ROIs used in the present study, are depicted in Fig. 1.

Dataset 1 (D1). The first dataset came from a study (N = 10) investigating the role of abstraction in category learning behavior using activity patterns from early visual cortex (Ritchie and Op de Beeck, 2019). Stimuli consisted of 16 square-wave annular gratings varying in four levels of spatial frequency and orientation (Fig. 1). Subjects completed 12 runs of a rapid-event related design in which each image appeared and flashed for 2 s (phase reversing at 4 Hz) followed by 2 s of fixation. The region of interest (Fig. 1) was anatomically defined V1 (Benson et al. 2012). For the present study the data was also smoothed at two levels: 6 mm and 9 mm FWHM. The original data was not normalized, and analysis was carried out within the native brain space of individual participants. In the present work, the data was again analyzed after transforming the data to a normalized brain space. All analyses on the normalized space data was identical to that carried out on the native brain space data, except that the normalized ROI image were thresholded to have a minimal increase of voxels within an ROI compared to the ROI image in the native brain space (mean increase = 44; SD = 29). For RSA the model RDM used was based upon the pairwise similarity judgments for the grating stimuli from both the in-scanner judgments of participants and a separate group of participants (N = 10) who performed the task off-line.

Dataset 2 (D2). The second dataset came from a study (N = 14) contrasting activity patterns for object category vs shape in multiple regions of the ventral and dorsal visual pathways (Bracci and Op de Beeck, 2016). Stimuli consisted of 54 greyscale natural images of 6 object types and 9 orthogonal shape types (Fig. 1). Across two sessions, subjects completed 16 (in one case 14) runs of a rapid-event related design in which two repeats of each image appeared for 1.5 s followed by 1.5 of fixation, in a pseudorandom order. The main ROI considered was bilateral object-selective lateral occipitotemporal cortex (LOTC), defined by a functional contrast of chairs > scrambled images based on separate localizer runs (Fig. 1). We also consider two other ROIs from the study: superior parietal lobe (SPL) and early visual cortex (BA17). Both regions were defined by a contrast of all localizer conditions > baseline. For the present study the data was smoothed at two further levels at 6 mm and 9 mm FWHM. Data was also normalized. For RSA the model RDM included the similarity judgments ratings for shape based on a multiple arrangement task.

Dataset 3 (D3). Data came from a study (N = 21) contrasting response patterns for social vs non-social actions across a large number of brain regions (Lee Masson et al. 2018). Stimuli consisted of 75 videos (3 s) with 39 depicting human-to-human touch interaction and 36 showing human-to-object interaction (Fig. 1; Lee Masson and Op de Beeck, 2018). The stimulus set included 3 videos depicting the same interaction with different actor pairs. Here we keep these conditions separate, as in the original study. Subjects completed 6 long runs of a rapid-event related design in which each video was shown followed by 3 s fixation in pseudorandom order. The main ROI considered was the bilateral temporoparietal junction (TPJ), which was defined by a contrast of observed touch vs baseline within an anatomical mask (Fig. 1).

Table 1
Summary of fMRI datasets. * indicates the acquisition voxel size for D3.

Dataset	Number of Subjects	Number of Conditions	Number of Repetitions per run	Condition duration (s)	Number of runs	Voxel size (mm)	TR (s)	TE (ms)	Flip angle (deg)	Field of view (mm)	Volume Dimensions	Number of Volumes per run
1	10	16	2	2	12	3	2	30	90	216	72 x 72 x 37	94
2	14	54	2	1.5	16 (or 14)	3	2	30	90	216	72 x 72 x 37	230
3	21	75	1	3	6	2.7*	2	30	90	216	72 x 72 x 37	239

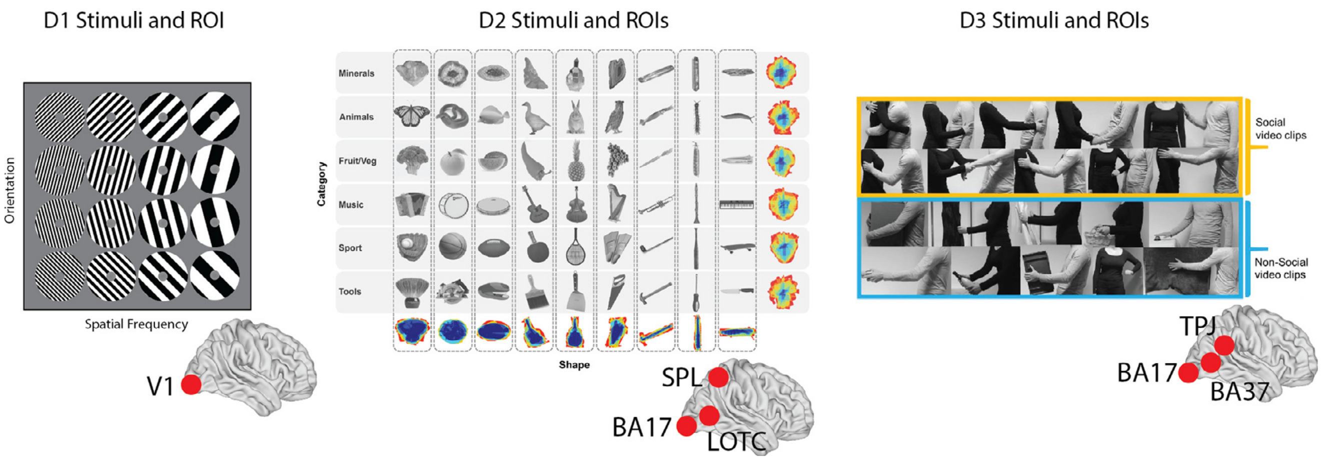

Fig. 1. Stimuli and approximate location of ROIs of the fMRI datasets. Stimulus images for D2 and D3 reproduced. With permission from Bracci and Op de Beeck (2016) and Lee Masson et al. (2018), respectively. For D3, only a subset of the video stimuli are depicted.

We also considered early visual cortex (BA17) and a lateral occipital region (BA37) as ROIs. Both were defined by the same contrast within anatomically defined masks. Data was normalized, and up-sampled to 2 mm resolution from 2.7 mm (Table 1). We also considered two levels of smoothing: 4 mm and 6 mm FWHM. For RSA the model RDM was the binary model matrix for the social/non-social division of the videos.

2.2. Multivariate noise normalization

Procedurally the process of $\mathrm{NN_M}$ was carried out as follows. Let **Y** be a matrix of size *S* (number of total scans across runs) x V (number of voxels in an ROI) consisting of the raw BOLD signal amplitudes after preprocessing, and let **X** be an *S* x *P* (number of predictors) sized design matrix from a GLM. We used the design matrix from SPM, which consists of separate columns for each experimental condition (for each run) as well as run-wise nuisance predictors for the six head motion parameters and run-specific constants. The ordinary least square estimate for the beta values for each predictor of X across all acquisitions in Y is then:

$$B = inv\left(\mathbf{X}^T * \mathbf{X}\right) * \ \mathbf{X}^T * Y \tag{1}$$

Where **B** is a *P* x *V* sized matrix of the beta weights for each predictor (rows) for each voxel (columns) of the ROI. In turn, the estimate of the residuals of the model is then:

$$R = Y - X * B \tag{2}$$

Where **R** is a *S* x *V* matrix reflecting the residual information of each scan (rows) for each voxel (columns). Following Walther et al. (2016, Eq. 4), the *V* x *V* variance-covariance matrix for the run *k* is then:

$$\Sigma_k = \frac{1}{S}\mathbf{R}_k^T * \mathbf{R}_k \tag{3}$$

Where $\mathbf{R}_\mathrm{k}$ is the portion of the residual matrix **R** corresponding to the scans in run *k*. Because the number of features (voxels) may be greater than the number of samples (scans) Σ_k can be rank deficient and non-invertible. To address this, previous studies have used the optimal shrinkage factor of Ledoit and Wolf (2004) to regularize $\mathbf{R}_\mathrm{k}$ towards the diagonal matrix. We used the oracle approximating shrinkage (OAS) factor, which outperforms the method of Ledoit and Wolf when the number of samples is much less than the number of features (Chen et al. 2010). After shrinkage, the multivariate noise normalized versions of the beta weights are then calculated as:

$$\mathbf{B}_k^* = \mathbf{B}_k * \Sigma_k^{-\frac{1}{2}} \tag{4}$$

Where $\mathbf{B}_\mathrm{k}$ is a *C* (number of conditions) x V matrix corresponding to the portion of **B** for run *k* that only includes the beta weights for the experimental predictors (i.e. excluding the nuisance predictors). More generally, **B*** is then a *P*' x *V* matrix, where *P*' is the number of conditions multiplied by the number of runs. As the estimates of the residuals and $\mathrm{NN_M}$ are carried out independently for each run, the values in **B*** can then be used to estimate neural dissimilarity values based on the cross-validated metrics described below.

This procedure for $\mathrm{NN_M}$ was carried out on all three datasets for the (smoothed and unsmoothed) imaging data, which was masked to select only the voxels within the specified ROIs. For D1, the same analysis was also carried out on the first 6, 8, 10, and 12 runs of the unsmoothed data.

Another option for normalizing the responses of each voxel is to down-weight the beta estimates of noisier voxels based on the standard deviation of their noise:

$$b_{k,p}^{+} = \frac{b_{k,p}}{\sigma_{k,p}} \tag{5}$$

Where $b_{k,p}$ is the beta estimate of voxel *p* for some predictor, for run *k*, and $\sigma_{k,p}$ is the standard deviation of its residual, or the square root of the diagonal value in Σ_k for the voxel *p* (Eq. 3). This *univariate* noise normalization ($\mathrm{NN_U}$) is similar to using the t-values instead of beta estimates in order to improve classifier performance and RDM reliability (Charest et al. 2018; Misaki et al. 2010). This alternative form of normalization was also carried out on all three datasets, for the unsmoothed data, masked by the ROI images.

Implementation of both $\mathrm{NN_M}$ and $\mathrm{NN_U}$ was carried out in Matlab using inbuilt SPM functions and custom code.

2.4. Dissimilarity metrics

An RDM is typically constructed as a *C* x *C* matrix of all pairwise estimates of dissimilarity between *C* conditions across features (voxels) based on some metric. We considered four different dissimilarity metrics in our analysis, the first three of which were evaluated by Walther et al. (2016).

The first metric is the commonly used 1 – *r*, or correlation, distance (Cor), which has been a mainstay of RSA since its initial development (Haxby et al., 2001; Kriegeskorte, Mur, and Bandettini, 2008; Op de Beeck et al., 2008; Kietzmann et al., 2019). To compute this metric, the portions of **B** for the conditions *i* and *j* are averaged across runs, and the resulting beta row vectors $\bar{b}_i$ and $\bar{b}_{j,}$ are linearly correlated and the Pearson's correlation coefficient, r_{ij}, is subtracted from 1. Note that it is common practice to subtract the mean from beta vectors before correlating them, as was the case for the study from which D3 was selected (Lee Masson et al. 2018). However, this has the potential to distort the relationships between conditions (Walther et al. 2016). We elected to not do mean subtraction, which entails that the correlation distance can be influenced by global signal differences across voxels. However, we note that preliminary analysis found that run-wise mean subtraction had no appreciable influence on the results when NN_M was employed. It is also possible to utilize a cross-validated version (Guggenmos et al. 2018), however we selected to use the simpler, and clearly most common, method.

The second metric is pairwise classifier accuracy (Cla), which is the first of the cross-validated metrics we consider. For this we used linear discriminant analysis (LDA), which maps samples onto a discriminant axis that maximizes the between-class variance, while minimizing the within-class variance. A decision values is then positioned orthogonal to the discriminant and decisions about the label for training data are based on the position of a sample on the discriminant relative to the decision value. For LDA (and the other metrics described below) we used a leave-one-run-out cross-validation procedure: the data from the training runs were averaged and used to estimate the discriminant and decision value, which was then used to label the test data. This was carried out for all cross-validation folds, and the dissimilarity metric is then the pairwise classifier accuracy. LDA makes an assumption of homoscedasticity: that the beta weight vectors for *i* and *j* have identical multivariate Gaussian distributions that differ only in their mean; that is, they have the same within-class variance-covariance matrix sigma Σ. These details are elaborated on below. Linear support vector machines (SVM) are also a popular algorithm that can be used to estimate neural dissimilarity (Walther et al. 2016), which do not make the same distributional assumptions and in some cases may be superior in performance to LDA (Misaki et al. 2010). However, we preferred LDA because it is analytically simpler (and therefore computationally faster), and because it is mathematically closely related to the other two distance metrics we utilized, which have been emphasized in previous work on NN_M (Diedrichsen et al. 2016; Guggenmos et al. 2018; Walther et al. 2016).

The third metric was the pairwise cross-validated squared Euclidean distance (Euc), which can be expressed as:

$$d^2_{euc} = \left(\bar{b}_i - \bar{b}_j\right)_{TR} * \left(b_i - b_j\right)^T_{TS}$$

Where $\bar{b}_i$ is the average beta weight vector from the training runs, denoted *TR*, and b_i is the beta weight vector from the test run, denoted *TS*. Procedurally the data for the conditions *i* and *j* was demeaned, split into the training and test partitions (leave-one-run-out), and the run-wise averages were computed for the training partition, which was then multiplied by the transposition of the difference between the beta weight vectors in the test partition. The final dissimilarity value was the results of this procedure when averaged across cross-validation folds.

Finally, when the variance-covariance matrix of the training data, or within-class "scatter", is added to Eq. 5, we get what Walther et al. (2016) call the cross-validated squared Mahalanobis distance (Mal) as a metric of dissimilarity:

$$d^2_{mal} = \left(\bar{b}_i - \bar{b}_j\right)_{TR} \Sigma^{-1}_{TR} * \left(b_i - b_j\right)^T_{TS} \tag{6}$$

The within-class scatter, Σ_{TR}, is just the equally weighted variance-covariance matrices for the training samples of beta weights for conditions *i* and *j* (Misaki et al. 2010):

$$\Sigma_{TR} = 0.5 * cov(\mathbf{B}_i)_{TR} + 0.5 * cov(\mathbf{B}_j)_{TR} \tag{7}$$

Where $\mathbf{B}_i$ is a matrix of all the samples for condition *i* in the training data. Procedurally the variance-covariance matrices can be rank deficient and so they were also regularized using OAS (Chen et al. 2010). When defined in this way, Mal is closely related to LDA since the pairwise difference along the discriminant, as a portion of Eq. 6, describes the feature weights for an LDA classifier:

$$w = \left(\bar{b}_i - \bar{b}_j\right)_{TR} \Sigma^{-1}_{TR} \tag{8}$$

The decision value along the discriminant can then be calculated by summing the mean beta weight patterns for the two conditions from the training data:

$$c = w * \frac{1}{2}\left(\bar{b}_i + \bar{b}_j\right)_{TR} \tag{9}$$

Then for any beta weight vector from the training set *b*, we can compute the value v:

$$v = w * b^T \tag{10}$$

If $v > c$ (or is positive), then the Fisher discriminant rule says to guess that *b* is from class *i,* otherwise if $v < c$ (or is negative), then the rule says to guess that *b* is from class *j*.

Two clarifications are worth making about Mal as a dissimilarity metric, before continuing. First, Walther et al. (2016) describe their preferred metric, the linear discriminant contrast (LDC), or "crossnobis" distance, as identical to Mal (Walther et al. 2016, Eq.9; see also Nili et al. 2020, p.6). This assertion is misleading, since LDC is defined as the value one obtains when one first carries out NN_M and then uses Euc as a dissimilarity metric (Dierdrichsen et al. 2016, Eq.4; Walther et al. 2016, Eq.7;). However, the variance-covariance matrices used for NN_M (Eq.3) and Mal (Eq.7) are clearly not the same and Mal can be used without NN_M (van Meel and Op de Beeck, 2020; Ritchie and Op de Beeck, 2019; Ritchie et al. 2020). To avoid confusion, we treat Euc and Mal as wholly distinct metrics. Since the assumption that NN_M somehow converts Euc to Mal is obscure, we did not follow Guggenmos et al. (2018) in replacing the within-class scatter from Eq. 7 with the identity matrix. Second, LDA is essentially a discretized version of Mal, and so will necessarily contain less information than a continuous metric (Walther et al. 2016). To address this drawback, one possibility is to use the distance between *v* and *c* as a dissimilarity metric, or the equivalent decision value for other classifiers, to weight the classifier accuracies (Guggenmos et al. 2018, Eq. 3). However, it is unclear to us what the advantage is of this metric compared to simply calculating Mal directly and so we do not consider it in the present work.

The calculation of the dissimilarity values using all these metrics was carried out with custom code in Matlab along with the CoSMoMVPA toolbox (Oosterhof et al. 2016).

2.4. Estimating reliability

The reliability of individual RDMs were estimated in a number of ways. First, following the analysis of Walther et al., our primary form of evaluation for reliability was to assess the within-subject reliability. For this we split the data of individual subjects into the odd and even runs and then constructed RDMs for these partitions based on the different metrices described above. The two RDMs were then Pearson's *r* correlated with each other to determine their reliability (Charest et al. 2018;

Guggenmos et al. 2018; Walther et al 2016). Other measures of reliability are possible. For example, Walther et al. (2016) propose the sum-of-squares differences for testing the reliability between two sets of RDMs because it will be influenced by scaling factors and provides a better estimate for distance metrics. We opted for the Pearson's r correlation as our estimate of reliability for a number of reasons. First, it is the most familiar method. Second, it is the same method used for estimating between-subject reliability. Third, unlike the sum-of-squares difference it can be used with Cor. And fourth it preserves continuity with the analyses that are typically used to evaluate RSA effect sizes (e.g. multiple regression).

Second, for the between subject reliability we calculated the noise ceiling (Cronbach, 1949; Nili et al. 2014; Op de Beeck et al. 2008). Individual RDMs were constructed using the above dissimilarity metrics for the full datasets, and then a leave-one-subject-out procedure was employed: the RDMs for all but one subject were averaged and then (Pearson's r) correlated with the left-out subject's RDM, and then the coefficients were averaged across all iterations of the procedure. This has sometimes been described as the “lower” bound of reliability with the upper bound defined by the same procedure except that the data of the one subject is also included in the group average (Nili et al. 2014). However, this latter procedure is explicitly described as generating overestimates of the reliability, while the former procedure is not. Since it is not clear how it provides an accurate description of the relevant feature, which is the explainable variance, we do not consider this overfitting estimate. Crucially, while in standard applications the noise ceiling is treated as a point estimate, there is of course a distribution based on the leave-one-subject-out folds, which is reflected in the results to follow.

Third, for Cor we also evaluated the relationship between the on- and off diagonal values, or “exemplar discriminability index” (EDI) as it has also been called (Haxby et al. 2001; Nili et al. 2020). The data of individual subjects was again split in half to odd and even runs, and then the pairwise correlations were performed between the two splits. In principle, the patterns for a condition should be self-similar across the two splits, in which case the diagonal values will be higher than the off diagonal values (Ritchie, Bracci, and Op de Beeck, 2017). The on- and off-diagonal values are then averaged, and their difference score is reported (on-minus-off). Positive values indicate that the patterns for each condition are on average reliably self-similar. Testing for the significance of this index of reliability assumes that the distribution of the difference scores are 0-mean normal given the null hypothesis. Although this assumption is not in fact true of the distribution, the simulations of Nili et al. (2020) suggest that the false positive rate is in fact low and so the test remains valid.

2.5. Comparison to model RDMs

To evaluate the impact of $\mathrm{NN_M}$ on the effect sizes for RSA, we compared individual subject RDMs to distinct model RDMs for each dataset (described above). The bottom half of individual RDMs (based on their full data, as constructed for the between-subject reliability analysis) were converted to column vectors and rank-order correlated (Spearman's ρ) with the lower half of the model RDM, which was also converted to a column vector. For discussion of why rank-order correlations such as Spearman's ρ are typically used to test RSA effects, rather than Pearson's *r,* see Kriegeskorte, Mur, and Bandettini (2008, Appendix). Notably, when binary coding is used, as with the social/non-social touch model for D3, dissimilarity values in effect have a Bernoulli distribution (taking values 0 or 1 at different frequencies) and so assuming linearity is inappropriate.

2.6. GLMdenoise

GLMdenoise is a method that estimates beta values and automatically derives nuisance predictors, and the optimal number of these predictors, directly from an fMRI dataset via cross-validation procedures (Kay et al. 2013). The guiding idea behind the pipeline is that this procedure can improve the beta estimates by iteratively determining how much of the variance between conditions and noise predictors can be captured. Unlike SPM, GLMdenoise analyzes the data from all runs together, and so outputs only a single beta weight for each condition. Although each run has its own set of noise regressors the number of these regressors are fixed across datasets and are determined by the steps described below. GLMdenoise also includes a number of polynomial regressors to characterize the baseline signal level, which shifts over time in each run. Polynomials of degrees 0 to round(L/2) are included, where L is the length of the runs in minutes. Thus, given the relatively short run length, there were three such regressors for D1 and five for D2 and D3, which had comparatively longer run lengths.

GLMdenoise has a number of steps, which we briefly summarize (for full details see Kay et al. 2013). First, a seed HRF for each condition is generated based on the stimulus durations (2, 1.5, and 3 s). Second, the signal estimate is determined by keeping either the HRF or nuisance regressors fixed and the ordinary least squares estimate is computed until there is a convergence of parameter estimates, based on when R^2 is 99% between the current and previous HRF iteration. Third, having set the form of the HRF, a leave-one-run-out cross-validation procedure is used to determine best fit of the GLM for each voxel. The predictions from the folds are combined and the total GLM model is compared to the data using R^2. Fourth, voxels are selected for the noise pool based on $R^2 < 0$ as determined at the previous step. Although this step can be used to exclude voxels outside the brain, we truncated the raw data using the whole brain mask generated by the GLM from SPM (Charest, Kay, and Kriegeskorte, 2018). Fifth, PCA is run on the noise pool, while projecting out the polynomial regressors. It is the resulting principle components that constitute the noise regressors for each run. Sixth, again, cross-validation is used to evaluate the model fit while varying the number of noise regressors by iteratively increasing the numbers of the principle components from the previous step that are used as predictors. The only way this step impacts the beta estimates is if there are correlations between the noise and condition regressors. This procedure also allows for greater variance to be captured. Seventh, the number of PC is selected based on the median of R^2 across the runs. A final step, performing a procedure to bootstrap to estimate error bars, was not performed to conserve computational resources.

We used the resulting denoised data in four ways. First, we applied the analysis separately to the odd and even splits in order to determine the impact on within-subject reliability. Note that this allowed that the number of PCs would be different between the odd and even splits. Also, since the predictors are concatenated, resulting in a single beta estimate for a condition from all runs, we could only estimate the within-subject reliability for the Cor measure since cross-validation was not possible. Second, we applied GLMdenoise to the full datasets to evaluate the between-subject reliability and RSA effect sizes, again only using RDMs constructed with the Cor metric. Third, we attempted to replicate the analysis of Charest et al., in which the design matrix generated by GLMdenoise is used to carry out NN, by plugging the matrix into Eqs. 1 and 2. However, this procedure required restructuring the design matrices in order to ensure run-specific estimates of beta values for each condition both for $\mathrm{NN_M}$ and to make utilization of the cross-validated dissimilarity metrics possible. To this end the estimates for each condition were assigned a distinct column in the design matrix (i.e. were no longer concatenated). No other modification of the design matrices was carried out. We note that as GLMdenoise already cross-validates across runs the assumption of independence between runs made by $\mathrm{NN_M}$ and cross-validated metrics is violated. However, the reconfigured design matrices were only used to evaluate the within-subject reliability, where the cross-validation at least occurs independently for the odd and even splits. Finally, we compared the results to those obtained when the SPM design matrices were split for the odd and even runs and the predictors for the experimental conditions concatenated across runs. This allowed us to assess whether any improvement in within-subject reliability obtained with GLMdenoise could also be achieved with SPM by changing the design matrix to only estimate a single beta value per condition.

2.7. Statistical analyses

To statistically assess the impact of $\mathrm{NN_M}$ on reliabilities and RSA effect sizes we modeled the individual correlation coefficients using linear mixed effects (LME) models with subject as a grouping factor and NN, dissimilarity metric, and their interaction, as fixed effects. There were random effects for the intercept, and $\mathrm{NN_M}$ and metric for each level of the grouping variable (that is, subject). To assess the overall fit of the models we report the R^2. Because the metric predictor included four classes, the LME estimates the effect of each metric and its interaction with $\mathrm{NN_M}$. Thus, to also assess the main effects of choice of dissimilarity metric and the interactions we applied an ANOVA to the LME. Although this set of statistical procedures carries the common (and commonly violated) assumption that the data is normally distributed, we utilized LME in this case because, by grouping the data by subject, it provides a more appropriate way of testing the impact of $\mathrm{NN_M}$ and metric as the data for a subject across metrics and $\mathrm{NN_M}$ is not independent. Rather, it reflects the same data being analyzed in a number of different ways. A main effect of $\mathrm{NN_M}$ tells us that the mean correlation across subjects is impacted by $\mathrm{NN_M}$ across metrics, where at the individual level $\mathrm{NN_M}$ generates an increase in the correlation value, not just the population and of metric regardless of $\mathrm{NN_M}$. An interaction tells us that these fixed effects influence each other. We also tested the paired differences (or difference scores; see Fig. 3C) in mean correlation coefficients for each metric, when calculated with and without NN, using paired t-tests.

When evaluating the impact of $\mathrm{NN_M}$ when run-number was manipulated for D1, we included it as a fixed effect. When the data was smoothed, we further added level of smoothing (none, Level 1, or Level 2) as a fixed effect in the LME. When evaluating the impact of $\mathrm{NN_M}$ on the EDI, the LME model only included $\mathrm{NN_M}$ and smoothing as fixed effects, since only Cor was used as a dissimilarity metric. When evaluating the impact on within-subject reliability smoothing was included along with $\mathrm{NN_M}$ and metric as fixed effects, and therefore the model contained multiple two-way interactions and a three-way interaction term. As before we evaluated main effects using an ANOVA applied to the LME and carried out paired t-tests of the mean individual correlations for each metric, with and without $\mathrm{NN_M}$. Since for the smoothed data analysis this resulted in far more statistical tests for each dataset, we controlled for multiple comparisons by reporting the FDR adjusted p-values. For GLMdenoise, we carried out paired t-tests of the mean correlations obtained with GLMdenoise relative to those for baseline and $\mathrm{NN_M}$. The same tests were performed when comparing $\mathrm{NN_M}$ with the GLMdenoise design matrices against $\mathrm{NN_M}$ with the SPM design matrices and against baseline.

2.8. Exploratory analysis of residual variance-covariance matrices

In an exploratory analysis we investigated the covariance structure of the run-wise variance-covariance matrices derived from the GLM residuals based on the spatial distance between voxels in an ROI as well as the goodness-of-fit (GoF) of the GLMs for each voxel.

For each subject we constructed matrices for the initial three ROIs based on the pairwise Euclidean distance between voxels in voxel space. Since plausibly covariance between voxels decreases with their distance, we used the exponentially decaying distance between voxels as an estimate of similarity in spatial position:

$$s_{ij} = \exp\left(-d_{ij}\right) \tag{11}$$

Where d_{ij} is the pairwise Euclidean distance between voxels i and j. Van Bergen and Jehee (2018) also consider Eq. 11 as a predictor of noise correlations between voxels and vary parameters for the rate and starting value of the decay in order to better fit their data. For the present preliminary analysis both values were set to 1. The off-diagonal values of the decaying distance matrix for each subject was correlated (Pearson's r) with covariance values of each of the run-specific variance-covariance matrices and then averaged. The group averages of these correlations were then tested for significance with a two-sided t-test. The between-run reliability of the run-specific covariance values was also estimated using a similar procedure for calculating the between-subject reliability of neural RDMs: for each run, the off-diagonal covariance values were correlated with the average values of the remaining runs, and then the across-run average of these correlations was calculated and tested for significance using a two-sided t-test.

We considered whether the above correlations (between the decaying distance matrices and variance-covariance matrices) might predict whether $\mathrm{NN_M}$ improved within-subject reliability. For this analysis we focused solely on Euc as a measure since, as reported below, it produced the largest improvements in within-subject reliability for D1 and D2. For all three datasets we subtracted individual within-subject reliability coefficients without $\mathrm{NN_M}$ from those obtained with $\mathrm{NN_M}$. We also considered whether these difference scores might be predicted by the ratio between voxels and experimental conditions which describe the size of the variance-covariance matrices and relate to the possibility of rank deficiency described above. Pooling across all three datasets, we rank-order correlated (Spearman's ρ) all three variables with each other: (i) the average correlations between the decaying distances between voxels and run-specific covariance values from the residuals; (ii) the change in within-subject reliability after $\mathrm{NN_M}$ when using Euc as the dissimilarity metric; and (iii) the ratio between voxels and experimental conditions.

We also evaluated whether the GoF of the GLMs for each voxel might bare a relationship to the covariance between voxels and predict whether $\mathrm{NN_M}$ improved within-subject reliability. The measure of goodness-of-fit we used was the model-based SNR:

$$\mathrm{SNR}_{mb} = \frac{\sigma_S^2}{\sigma_N^2} \tag{12}$$

Which divides the variance of the explained BOLD signal by the variance of the unexplained BOLD signal (Welvaert and Rosseel, 2013). Calculation of the SNR_{mb} was carried out using code adapted from the MACS SPM toolbox (Soch and Allefeld, 2018). We evaluated this GoF measure in two ways. First, we determined the proportion of voxels in an individual subject ROI where the explained signal was greater than the unexplained signal (i.e. $\mathrm{SNR}_{mb} > 1$). We also made a dissimilarity matrix based on the pairwise absolute difference in GoF, which was correlated with the run-averaged residual variance-covariance matrix. Both of these sets of values derived from the GoF were then rank-order correlated (Spearman's ρ) with the change in within-subject reliability after NN, when Euc was used as a metric.

3. Results

We investigated whether $\mathrm{NN_M}$ improves the estimate of condition specific changes in the BOLD signal of fMRI, and thereby improves the reliability of RDMs, and whether this effect of $\mathrm{NN_M}$ interacts with choice of dissimilarity metric. We reanalyzed three datasets (D1-D3) from previous studies in order to evaluate the impact of $\mathrm{NN_M}$ on the within- and between-subject reliability of neural RDMs constructed using four different metrics, as well as the RSA effect sizes based on correlations with dataset-specific model RDMs. We further evaluated the results of $\mathrm{NN_M}$ for other ROIs from two of these datasets, compared its impact across different levels of spatial smoothing, and finally compared it to results obtained with GLMdenoise, which has also been suggested as a method for improving signal estimate for the purpose of RSA. Because many statistical tests were performed for each portion of the results, they are reported in accompanying tables.

Table 2
Summary of linear mixed effects (LME) modeling of the results depicted in Fig. 2.

Dataset	Effect	R^2	F	df_1	df_2	p
Table 2A: Within-subject reliability						
1	LME	0.84	29.91	7	72	9.66E-18
	NNM		19.36	1	72	3.67E-05
	Metric		11.63	3	72	2.66E-06
	Interaction		10.72	3	72	6.56E-06
2	LME	0.92	16.56	7	104	1.63E-14
	NN		5.9	1	104	0.02
	Metric		17.7	3	104	2.37E-09
	Interaction		18.76	3	104	8.43E-10
3	LME	0.51	3.45	7	160	0.002
	NNM		13.38	1	160	9.25E-04
	Metric		2.13	3	160	0.1
	Interaction		1.29	3	160	0.28
Table 2B: Between-subject reliability						
1	LME	0.88	46.78	7	72	2.64E-24
	NNM		21.5	1	72	8.37E-17
	Metric		52.53	3	72	4.22E-18
	Interaction		44.1	3	72	2.75E-16
2	LME	0.97	93.51	7	104	5.94E-42
	NNM		81.2	1	104	1.08E-14
	Metric		138.49	3	104	3.49E-36
	Interaction		123.42	3	104	3.92E-34
3	LME	0.65	26.81	7	160	4.36E-24
	NNM		58.58	1	160	1.73E-12
	Metric		4.31	3	160	5.90E-03
	Interaction		1.43	3	160	0.24
Table 2C: Effect sizes						
1	LME	0.93	28.59	7	72	2.04E-18
	NNM		23.29	1	72	7.57E-06
	Metric		41.74	3	72	9.67E-16
	Interaction		34.72	3	72	5.45E-14
2	LME	0.9	16.22	7	104	2.84E-14
	NNM		29.94	1	104	3.10E-07
	Metric		21.91	3	104	4.48E-11
	Interaction		20	3	104	2.61E-10
3	LME	0.93	6.79	7	160	4.61E-07
	NNM		34.7	1	160	2.20E-08
	Metric		3.86	3	160	1.06E-02
	Interaction		4.92	3	160	2.70E-03

3.1. Multivariate noise normalization does not consistently improve the within- or between-subject reliability or RSA effect sizes

We began with within-subject analyses following the approach of Walther et al. For each dataset we calculated the correlation between the odd and even run RDMs of each subject both with and without NN_M and across dissimilarity metrics. For the four datasets they analyzed, Walther et al. considered regions from primary motor and sensory cortex (M1/S1) as well as high-level visual cortex. Thus, the choice to include D1, with V1 as a ROI and gratings as stimuli, was to have an equivalent early sensory area include in our analysis. Similarly, D2 was included, with LOTC as an ROI and natural object images as stimuli, to have a similar dataset to two of those considered by Walther et al. However, Walther et al. only evaluated sensorimotor or visual areas and did not consider any ROI that is known to be responsive to more abstract or cognitive relationships between experimental conditions. This was our motivation for including D3, with its large number of social/non-social touch videos, and to initially focus on an area like TPJ, which is known to be recruited by social cognition and theory of mind (Saxe and Kanwisher, 2003). Inclusion of D3 was especially important for evaluating how well NN_M applies more broadly.

For D1, the results were well described by the LME model with significant main effects of NN_M, metric, and interaction (Table 2A). All paired t-tests for the metrics were also significant (Fig. 2A). So, for D1, across metrics, NN_M improved within-subject reliability, though the improvement tended to vary with choice of metric. For D2, the results were well described by the LME model with significant main effects for NN, metric, and interaction (Table 2A). Only the paired t-tests for Euc and Mal were significant (Fig. 2A). So, for D2, NN_M only improved within-subject reliability when using the two distance metrics. For D3, the results were well described by the LME model and there was a significant main effect of NN, but not metric or interaction (Table 2A). However, the effect of NN_M for D3 was in the wrong direction, though only the paired *t*-test for Cor was significant (Fig. 2A). So, for D3, NN_M made within-subject reliability worse.

Next, we evaluated the impact of NN_M on the between-subject reliability by calculating the noise ceiling with and without NN, for each of the four metrics. For D1, the results were well described by the LME model and there were significant main effects of NN_M, metric, and interaction (Table 2B). Only the paired *t*-test for Cla was not significant (Fig. 2B). So, for D1, NN_M improved between-subject reliability but the size of this improvement varied with metric (Fig. 2B). For D2, the results were well described by the LME model and there were significant effects of NN_M, metric, and interaction (Table 2B). Only the paired *t*-test for Clas was not significant (Fig. 2B). So, as with D1, for D2 NN_M tended to improve between-subject reliability, but the size of this improvement varied with metric (Fig. 2B). For D3, the results were also well described by the LME model, with significant effects of NN_M and metric, but no interaction (Table 2B). All paired tests were also significant, but unlike with D1 and D2, the differences were in the wrong direction, with NN_M consistently lowering the between-subject reliability (Fig. 2B). So, for D3, NN_M also made the between-subject reliability worse.

Finally, we considered how NN_M might influence actual RSA effect sizes since it was conceivable that it might impact the explainable variance without changing the mean correlations with model RDMs. For D1, the results were well described by the LME model with significant main effects of NN, metric, and interaction (Table 2C). The paired differences were only significant for Cor and Euc (Fig. 2C). Although the mean correlations were much lower for D2 than D1, the same pattern was observed. The LME model explained most of the variance and there were main effects of NN, metric, and interaction (Table 2C). The paired differences were also only significant for Cor and Euc. So, for both D1 and D2, NN_M only seemed to increase the RSA effect sizes when using two of the metrics. For D3, the LME model was again significant, and the same pattern was observed as was seen for the within- and between-subject reliability: there were main effects of NN, metric, and interaction (Table 2C). But this was once again in the wrong direction for NN, with mean correlations significantly lower across all metrics (Fig. 2C).

Two observations are worth making about the findings summarized so far. First, regarding the positive results for D1 and D2, among metrics NN_M had very little influence on any of the mean effect sizes when Cla was used while it had the greatest impact when Euc was used. This substantial improvement seemed to occur because the correlations for Euc were low to begin with. In contrast, the mean reliabilities were already much higher when Cor and Mal were used as metrics and tended to be at least slightly improved by NN_M. However, if one chose the metric based solely on the magnitude of the baseline RSA effect sizes when NN_M was not utilized, then Mal should be selected as NN_M did not significantly increase the mean correlations. Second, it is notable that the baseline within-subject reliability is considerably lower for D3 than the other two datasets. This is likely the case for a number of reasons: that it is a cognitive region; that the study included a much larger number of conditions; and finally that there were far fewer runs (only half as many as for D1). In part for these reasons we next considered the impact of NN_M on other ROIs.

3.2. Univariate noise normalization has less impact on within-subject reliability

We next assessed whether it was specifically the normalization by the covariance between voxels that produced the discrepant results across datasets. To do this, we performed univariate noise normaliza-

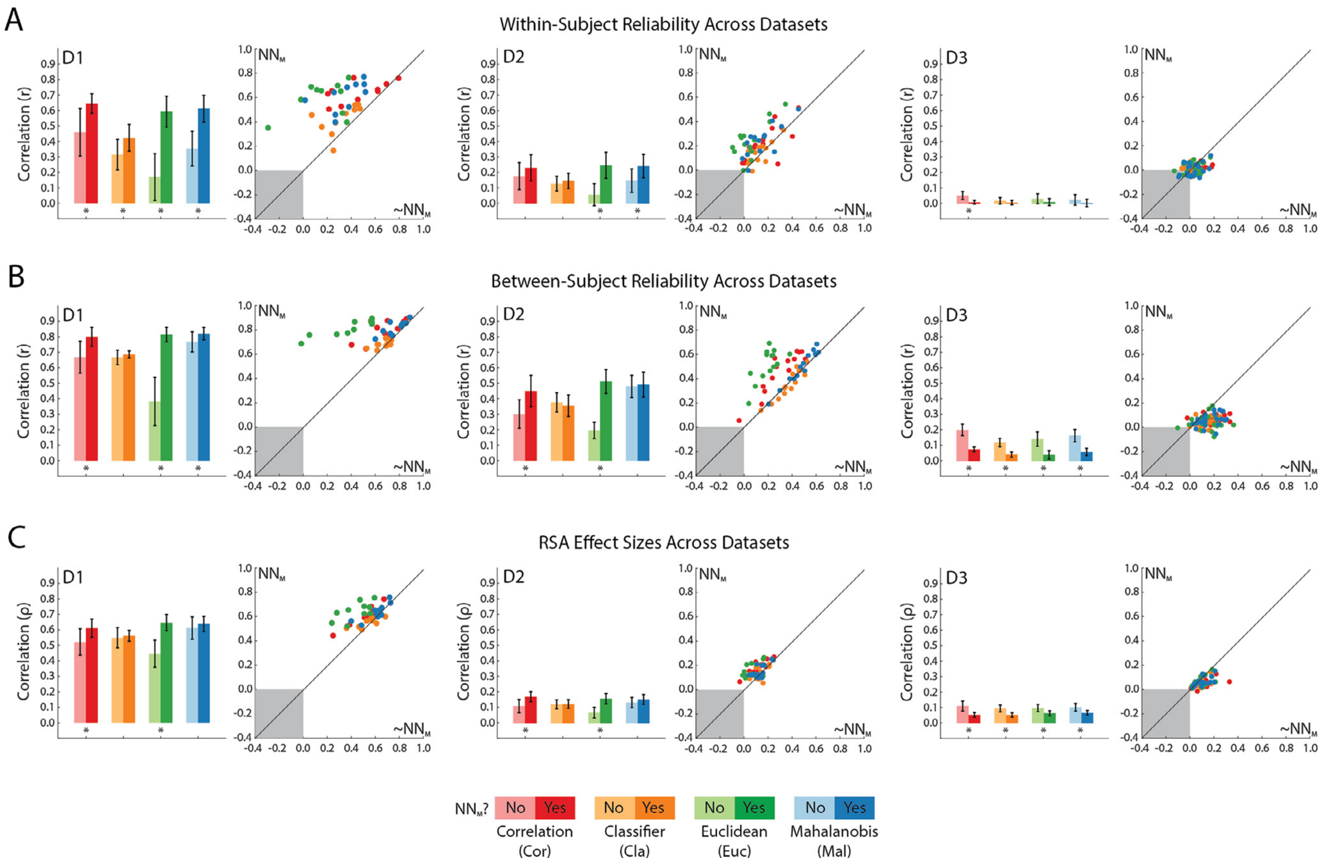

Fig. 2. Multivariate noise normalization (NN_M) does not consistently improve the reliability of neural RDMs, or effect sizes, across datasets. (A) Bars indicate the group mean of within-subject split-half reliabilities and error bars are the standard error of the mean. Colors indicate the different dissimilarity metrics. Lightened bars on the left of each pair indicate the mean value when NN_M was not carried out prior to RSA, while saturated colored bars to the right indicate mean correlations when NN_M was carried out. * = $p < 0.05$ based on two-sided paired t-tests. Error bars are the standard error of the mean (SEM). Also depicted are scatter plots of all individual subject data points across all four dissimilarity metrics. (B) Bars indicate the group mean of the between-subject reliability or “noise ceiling”. All conventions are the same as in (A). (C) Bars indicate the mean rank-order correlations between individual neural RDMs and target model RDMs for each dataset. All conventions are the same as in (A).

tion (NN_U), and compared its impact on the within-subject reliability across all three datasets.

For D1, we found that NN_U also significantly improved reliability compared to baseline across all four of the dissimilarity metrics, but to a significantly lesser degree than NN_M when Euc and Mal were used to construct the RDMs (Fig. 3A). For D2, NN_U only improved reliability when Cor was used, with no significant difference for any of the other metrics for which reliability was effectively unchanged from baseline. Thus, NN_M also significantly improved reliability compared to the results with NN_U, when either Euc or Mal were used as metrics. Finally, for D3, NN_U had no significant impact on the within-subject reliability, regardless of the metric used. Crucially, unlike NN, there was no general trend of decreasing the reliability either.

These results suggest that the differential impact of v across datasets and metrics is specifically a result of applying multivariate NN_M and normalizing by the run-wise covariance between residuals of the voxels in an ROI, as NN_U had comparatively far less influence (positive or negative) on within-subject reliability.

3.3. The improvement in within-subject reliability from multivariate noise normalization is consistent across number of runs

Another question is whether the difference in findings we observed between D1 and D2 compared to D3 is a result of the difference in the amount of data per subject. D1 and D2 contained 12 and 16 (or 14) runs per subject, while D3 only contained 6 runs. To assess the amount of runs on within-subject reliability, we further analyzed the data of D1. For each subject we carried out the same analysis as before, splitting the odd and even runs, based on the first 6, 8, 10, or full 12 runs of a subject (Fig. 3B).

For D1, the results were well described by the LME model with significant main effects of NN, metric, and run number, with a significant interaction between NN_M and metric (Table 3). With fewer runs, there was no significant increase in reliability when Cor was used as a metric, while the reverse pattern, of decreasing differences, was suggested when Cla was used. Regardless of the number of runs, NN_M resulted in a substantial improvement when Euc and Mal were used as dissimilarity metrics. Notably, the baseline reliability achieved with Cor with only 6 runs, even without NN_M, was much higher than for any of the cross-validated metrics. Furthermore, even with only 6 runs, the reliability for D1, even without NN_M, tended to be higher than for D3, and similar or higher than what was observed for D2.

Table 3
Within-subject reliability (# of runs).

Dataset	Effect	R^2	F	df_1	df_2	p
1	LME	0.79	38.73	8	311	1.72E-42
	NNM		15.8	1	311	8.75E-05
	Metric		24.01	3	311	5.30E-14
	Runs		75.84	1	311	1.84E-16
	N x M		21.8	3	311	7.71E-13

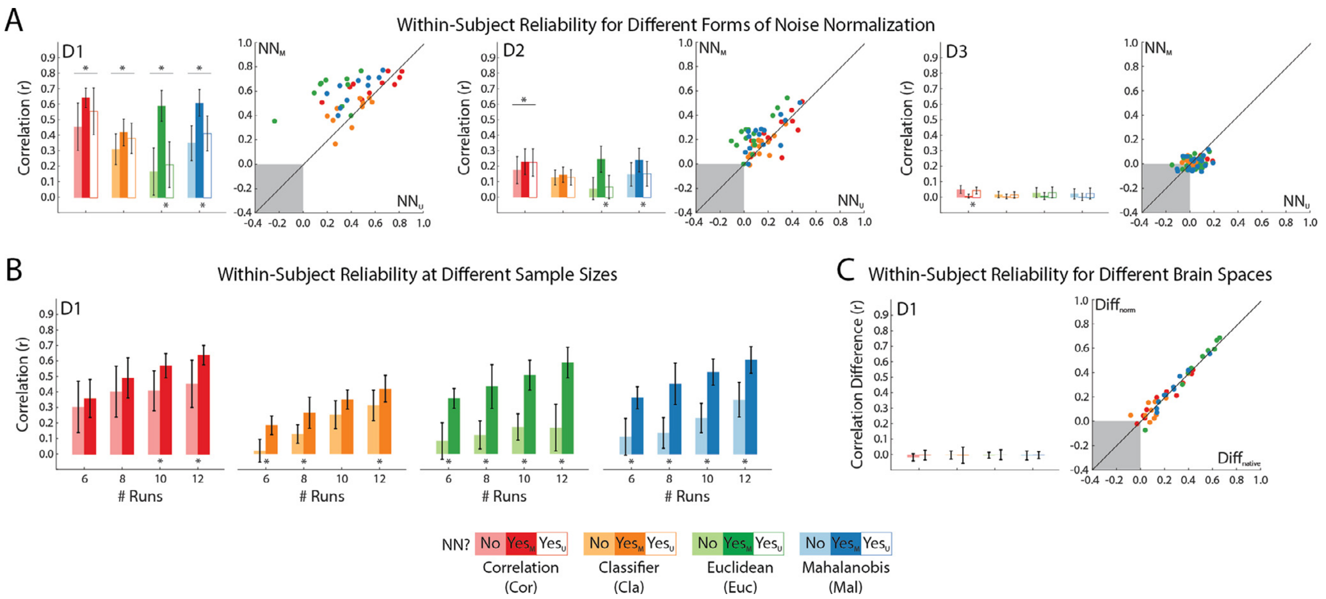

Fig. 3. Comparing the effects of multivariate vs univariate noise normalization, number of runs, and normalizing the brain images on within-subject reliability of neural RDMs. (A) – (B) Bars indicate the group mean of within-subject split-half reliabilities and error bars are the standard error of the mean. (A) White bars with colored outlines indicate results when carrying out univariate noise normalization (NN_U). (B) Pairs of bars indicate the results for different numbers of runs per subject. All other conventions are the same as in Fig. 2A. (C) Bars indicate the difference of the within-subject reliabilities subtracting the effects observed based on results when data was in subjects' native brain space vs when it was in a normalized brain space. Errors are the standard error of the mean. Scatter plot indicates the difference in within-subject reliability as a result of NN_M either in the native subject space ($Diff_{native}$) or the normalized space ($Diff_{norm}$).

These results show that it is possible to obtain substantial improvement in within-subject reliability of individua RDMs with as few as 6 runs, which was the same number of runs for D3. Thus, the discrepant results we observed for that dataset, are unlikely to be merely a result of run sample size.

3.4. *The improvement in within-subject reliability from multivariate noise normalization is not impacted by normalizing brain space*

Another possibility is that the relatively consistent results observed for D1 relative to D2 and D3 was a result of the fact that it was carried out on data in the native brain space of each individual subject and not a normalized brain space as was the case for the other two datasets. For it is conceivable that, to the extent that the noise information present in the residuals has a spatial component, this might be distorted through normalization.

To test this, we first normalized the data for D1 before carrying out the exact same analyses for constructing a GLM for the data, applying NN, and evaluating the impact of the method on within-subject reliability across dissimilarity metrics. When the within-subject reliabilities obtained for the normalized space data were subtracted from those for the native space data, these difference scores were never significantly different from chance (Fig. 2C). The lack of influence of normalization is also apparent when looking at individual data points which largely fall along the diagonal in Fig. 2C, indicating that there was no appreciable difference in the impact of NN_M, whether data from a normalized or native brain space is used.

These results suggest that the difference in results for D1 from D2 and D3 are unlikely to simply be a result of whether data is analyzed in a subject's native brain space.

3.5. *The impact of multivariate noise normalization is consistently inconsistent across regions of interest*

For the three datasets we considered, and the three evaluations we performed, the impact of NN_M was most consistently beneficial for D1, more mixed for D2, and generally quite negative for D3. Unlike D1, D2 and D3 both came from studies that considered a large number of ROIs. Given our discrepant findings, we next asked whether similar results might obtain for a selection of the other ROIs from these studies. Since the ROI for D1 was a portion of V1, we also considered the equivalent ROIs for D2 and D3 to determine whether the impact of NN_M would be consistently more positive in early visual cortex, across datasets. In both cases in the original studies these regions were labeled BA17. For D3 we initially focused on a non-visual region, and this raised the question of whether this might play a role in the negative influence of NN_M we observed. Thus, for D2, we also considered a ROI outside of the ventral visual pathway, SPL. For D2 we had initially focused on a high-level visual region and so we also considered a similar ROI for D3: a portion of lateral occipitotemporal cortex, labeled as BA37 in the original study. For the pairs of new ROIs for D2 and D3, we then again carried out the same analyses as before. Results are depicted in Fig. 4 and statistical tests are reported in Table 4.

For both datasets, the correlations for within-subject reliability for neural RDMs for BA17 were well described by the LME model. There were significant main effects for Metric and Interaction for D2, and only Metric for D3 (Table 4A). However, for D2 there was only a significant increase for Euc, and no significant differences in mean reliability for D3 (Fig. 4A). Thus, for D2, NN_M seemed to have very little positive impact on the mean within-subject reliability of neural RDMs for BA17, and for D3, NN_M seemed to have no impact, though there was possibly a slightly trend in a positive direction. For D2 (SPL) the LME model was significant, and there were main effects of NN_M and metric, but no interaction (Table 4A). Like with TPJ for D3, for all metrics NN_M decreased within-subject reliability, though only the effect for Cor was significant. The results for D3 (BA37) were qualitatively similar to those for BA17, though the LME model was not significant even though there were significant effects of NN, metric, and interaction, and only a significant pairwise difference for Cor and no effect or a possible positive trend otherwise (Fig. 4A).

When it came to the results for the between-subject reliability, for both datasets the correlations for all ROIs were well-described by the LME model and all main effects were significant, with the exception of NN_M for D2 (SPL) and D3 (BA17) (Table 4B). For D2, for BA17 there was

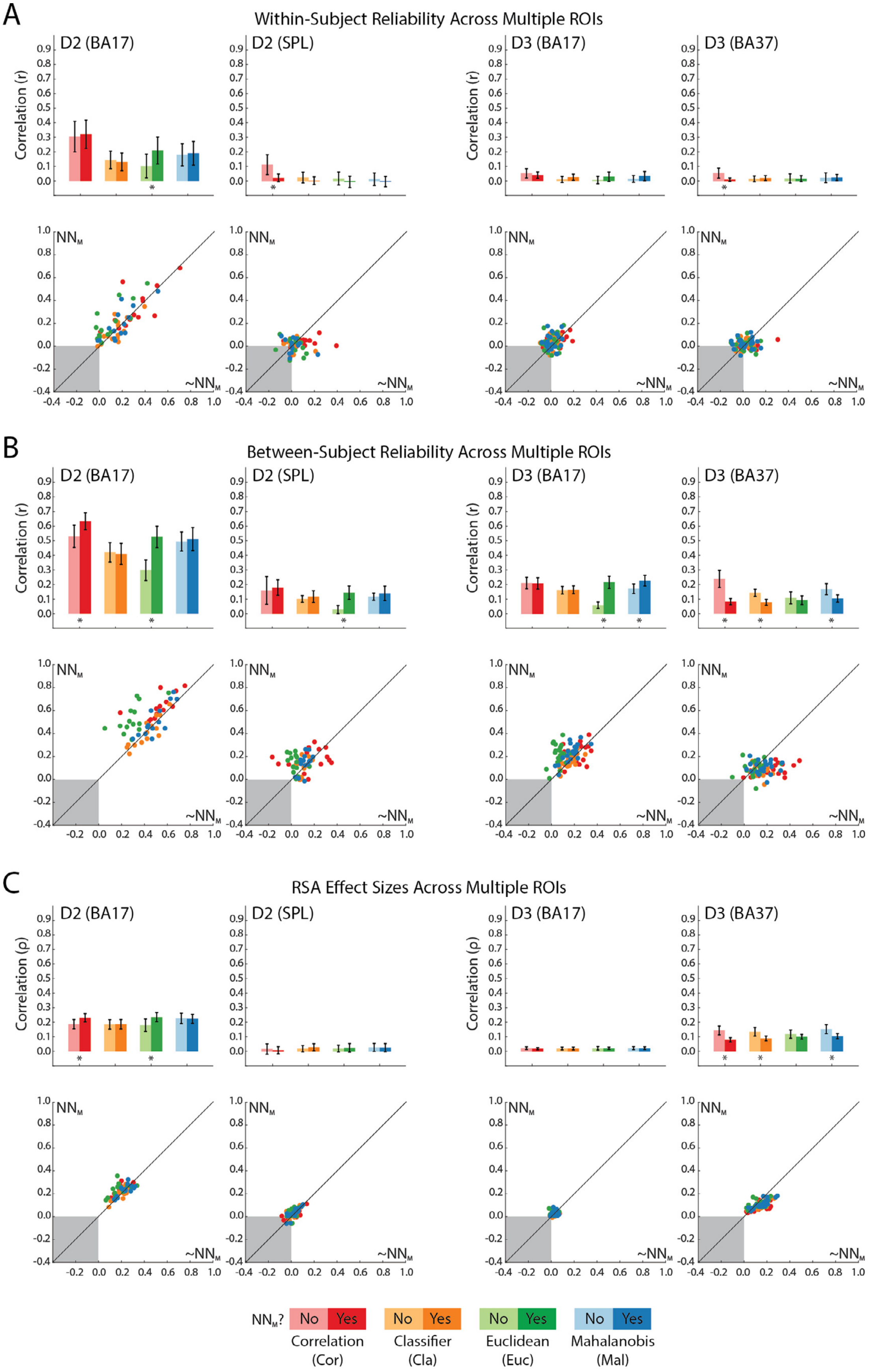

Fig. 4. Multivariate noise normalization (NN$_M$) does not consistently improve the reliability of neural RDMs, or effect sizes, in different ROIs. (A) Bars indicate the group mean of within-subject split-half reliabilities and error bars are the standard error of the mean. (B) Bars indicate the group mean of the between-subject reliability or "noise ceiling". (C) Bars indicate the mean rank-order correlations between individual neural RDMs and target model RDMs for each dataset. All conventions are the same as in Fig. 2A.

Table 4
Summary of linear mixed effects (LME) modeling of the results depicted in Fig. 4.

Table 4A: Within-subject reliability

Dataset	Effect	R^2	F	df_1	df_2	p
2 (BA17)	LME	0.97	19.18	7	104	2.99E-16
	NNM		0.41	1	104	0.52
	Metric		24.51	3	104	4.49E-12
	Interaction		14.14	3	104	8.53E-08
2 (SPL)	LME	0.27	4.65	7	104	1.45E-04
	NNM		13.25	1	104	4.27E-04
	Metric		4.38	3	104	0.006
	Interaction		1.99	3	104	0.12
3 (BA17)	LME	0.7	2.93	7	160	6.50E-03
	NNM		0.92	1	160	0.340
	Metric		4.64	3	160	2.16E-05
	Interaction		2.33	3	160	0.08
3 (BA37)	LME	0.33	1.92	7	160	0.07
	NNM		10.65	1	160	0.001
	Metric		3.6	3	160	0.02
	Interaction		2.95	3	160	0.03

Table 4B: Between-subject reliability

Dataset	Effect	R^2	F	df_1	df_2	p
2 (BA17)	LME	0.94	48.87	7	104	4.35E-30
	NNM		23.4	1	104	4.58E-06
	Metric		53.5	3	104	5.34E-21
	Interaction		44.92	3	104	1.06E-18
2 (SPL)	LME	0.38	5.96	7	104	3.91E-05
	NNM		0.48	1	104	0.49
	Metric		5.23	3	104	0.002
	Interaction		3.05	3	104	0.03
3 (BA17)	LME	0.86	33.42	7	160	2.50E-28
	NNM		0.02	1	160	0.88
	Metric		56.65	3	160	5.21E-25
	Interaction		43.62	3	160	1.18E-20
3 (BA37)	LME	0.53	26.69	7	160	8.64E-16
	NNM		66.08	1	160	1.12E-13
	Metric		13.78	3	160	4.82E-08
	Interaction		9.3	3	160	1.02E-05

Table 4C: Effect sizes

Dataset	Effect	R^2	F	df_1	df_2	p
2 (BA17)	LME	0.93	14.57	7	104	4.24E-13
	NNM		17.29	1	104	6.60E-05
	Metric		14.26	3	104	7.48E-08
	Interaction		16.47	3	104	7.88E-09
2 (SPL)	LME	0.86	0.73	7	104	0.64
	NNM		0.86	1	104	0.36
	Metric		0.55	3	104	0.65
	Interaction		1.09	3	104	0.36
3 (BA17)	LME	0.83	0.89	7	160	0.42
	NNM		0.5	1	160	0.48
	Metric		0.37	3	160	0.77
	Interaction		0.46	3	160	0.71
3 (BA37)	LME	0.95	22.68	7	160	1.00E-15
	NNM		43.64	1	160	5.56E-10
	Metric		19.2	3	160	1.10E-10
	Interaction		17.27	3	160	9.12E-10

Table 5
Summary of linear mixed effects (LME) modeling of the results depicted in Fig. 5.

Table 5A: On-off diagonal diffrence

Dataset	Effect	R^2	F	df_1	df_2	p
1	LME	0.62	9.99	3	56	2.26E-05
	NNM		8.65	1	56	0.005
	Smoothing		1.44	1	56	0.24
	Interaction		0.01	1	56	0.91
2	LME	0.99	28.84	3	80	9.62E-13
	NNM		11.21	1	80	0.001
	Smoothing		3.52	1	80	0.06
	Interaction		4.59	1	80	0.04
3	LME	0.97	19.41	3	122	2.36E-10
	NNM		18.08	1	122	4.180E-05
	Smoothing		29.07	1	122	3.46E-07
	Interaction		15.23	1	122	1.56E-04

Table 5B: Within-subject reliability (smoothing)

Dataset	Effect	R^2	F	df_1	df_2	p
1	LME	0.87	27.72	15	224	4.57E-38
	NNM		32.97	1	224	2.99E-08
	Metric		20.24	3	224	1.16E-11
	Smoothing		19.28	1	224	1.73E-05
	N x M		37.03	3	224	1.62E-19
	N x S		14.74	1	224	1.60E-04
	M x S		0.26	3	224	0.85
2	LME	0.94	33.18	15	320	2.43E-49
	NNM		6.39	1	320	0.01
	Metric		48.9	3	320	4.44E-26
	Smoothing		13.79	1	320	2.41E-04
	N x M		83.48	3	320	5.24E-40
	N x S		43.46	1	320	1.76E-10
	M x S		0.84	3	320	0.47
3	LME	0.71	3.15	15	488	2.43E-04
	NNM		13.79	1	488	2.28E-04
	Metric		4.12	3	488	0.01
	Smoothing		5.31	1	488	0.02
	N x M		4.02	3	488	0.008
	N x S		7.86	1	488	0.005
	M x S		2.11	3	488	0.10

a significant increase in mean reliability due to NN_M when Cor and Euc were used as metrics. For D3, there were significant increases in mean reliability due to NN_M when Euc and Mal were used as metrics for BA17 (Fig. 4B). Thus, across metrics, the impact of NN_M on between-subject reliability was also quite variable when considering the early visual ROIs for D2 and D3. For D2 (SPL) there were significant main effects of metric and interaction, but not NN_M (Table 4B). Only when Euc was used as a metric was there a significantly increased in the mean correlations due to NN_M. For D3 (BA37) the pattern of results was similar to those for TPJ with significant main effects, but significant decreases in reliability for all metrics except Euc (Fig. 4B).

Finally, when it came to the RSA effect sizes, the correlations for D2 (BA17) and D3 (BA37) were well described by the LME model and all main effects were significant (Table 4C). For D2 (BA17) there were only significant increases in the mean correlations for Cor and Euc, while for all metrics NN_M resulted in significant decreases in mean effect sizes. These results further suggest that the mixed findings we previously observed were not merely a result of the ROIs that were initially selected. Although the R^2 values appeared similar for the other two ROIs, the LME did not significantly capture any variance for D2 (SPL) or D3 (BA17) for which there was very little variation (Fig. 4C). These null findings were somewhat expected, given the weak (D2) or non-existent (D3) findings from the original studies (Bracci and Op de Beeck, 2016; Lee Masson et al. 2018).

3.6. Multivariate noise normalization interacts with spatial smoothing in different ways, for different datasets

Previous work suggests that multivariate BOLD signals can be modestly enhanced by levels of spatial smoothing, which can improve the EDI values and reliability of RDMs for RSA (Hendriks et al. 2017; Op de Beeck, 2010). We also sought to evaluate how spatial smoothing and NN_M might interact, given that the noise targeted by the method is also presumed to be spatially distributed and so may be influenced by the possible enhancement of the spatially distributed signal. To this end, we investigated the on-off diagonal difference (EDI) and the within-subject reliability at different levels of smoothing (with and without NN_M). For the analysis we focused on the initial trio of ROIs. Results are depicted in Fig. 5 and LME modeling statistics are summarized in Table 5.

For D1, the correlations were well captured by the LME model with a significant main effect of NN, but not smoothing or interaction (Table 5A). There were significant pairwise differences in the mean EDI with no smoothing, and at Level 1 (Fig. 5A). For D2, the correlations were well captured by the LME model with a significant main effect of NN_M and interaction, but not smoothing (Table 5A). There were significant pairwise differences in the mean EDI with no smoothing and at

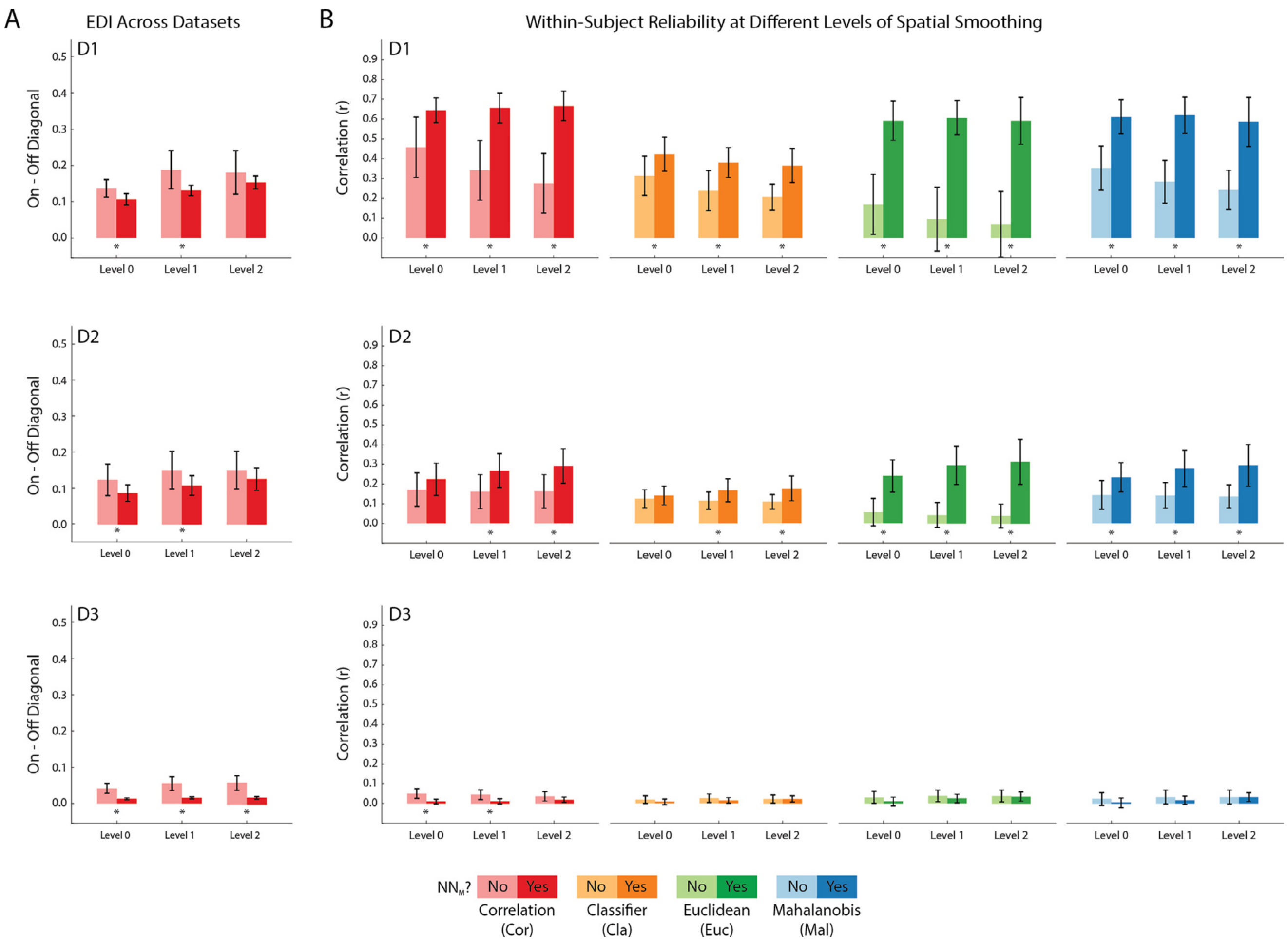

Fig. 5. Multivariate noise normalization (NN_M) interacts with levels of spatial smoothing. (A) Bars indicate the mean EDI, with and without NN_M, for three levels of spatial smoothing. Error bars are standard error of the mean. * = $p < 0.05$ based on two-sided paired t-tests (B) Bars indicate the group mean of the split-half within subject reliability, with and without NN_M, across all three levels of spatial smoothing. * = FDR-adjusted $p < 0.05$ based on two-sided paired t-tests. Conventions are otherwise the same as in Fig. 2A.

Level 1 (Fig. 5A). For D3, the correlations were well captured by the LME model and all main effects were significant (Table 5A). There were significant pairwise differences in the mean EDI at all three levels of smoothing (Fig. 5A). Notably, in all cases where there was a pairwise difference in mean EDI, this was because NN_M tended to decrease the values of this index of on-diagonal reliability of individual RDMs, though this decrease tended to diminish with increased levels of smoothing.

Next, we evaluated the impact of NN_M and smoothing on within-subject reliability. For D1, the correlations were well captured by the LME model, and there were significant main effects for all three fixed effects (NN_M, metric, and smoothing), and the interactions between NN_M and metric and NN_M and smoothing level (Table 5B). All pairwise tests were significant, with within-subject reliability appearing to decrease with greater levels of smoothing and NN_M undoing this effect (Fig. 5B). For D2, the correlations were well captured by the LME model, and there were significant main effects for all three fixed effects (Table 5B). The interactions between NN_M and metric and NN_M and smoothing level were also significant (Table 5B). Once smoothing was introduced all pairwise comparisons were significant and the pattern observed for D1 was reversed; for D2, the baseline within-subject reliability was not influenced by smoothing, but consistently increased with level of smoothing when NN_M was performed (Fig. 5B). Finally, for D3, the correlations were well captured by the LME model with a significant main effect for all three fixed effects (Table 5B). There was also a significant interaction between NN_M and metric and NN_M and smoothing. A significant decrease in mean within-subject reliability was only observed for the first two levels of smoothing for Cor (Fig. 5B). However, smoothing seemed to undo the negative impact of NN_M.

In summary, although NN_M and spatial smoothing are in principle motivated by similar considerations about the spatial distribution of noise and signal, respectively, these two factors tended to interact in different ways depending on the dataset in question. We return to the topic of the spatial properties of the residual variance-covariance matrices below.

3.7. GLMdenoise has a more consistently beneficial impacts on the reliability of neural dissimilarity and RSA effect sizes than multivariate noise normalization

Charest et al. (2018) propose GLMdenoise as a method for improving signal estimates, and thus the reliability of neural RDMs. Their results suggest that, when preprocessing is carried out with GLMdenoise, superior results are obtained compared to NN_M. Furthermore, they found that NN_M could be improved when using the design matrix of GLMdenoise to estimate the residuals. We revisited both findings of Charest et al. (2018) again using the initial trio of ROIs. First, we compared the results obtained with Cor for within- and between-subject reliability of neural RDMs and RSA effect sizes with those obtained when employing GLMdenoise. Second, we compared the results obtained with NN_M for within-subject reliability when using GLMdenoise design ma-

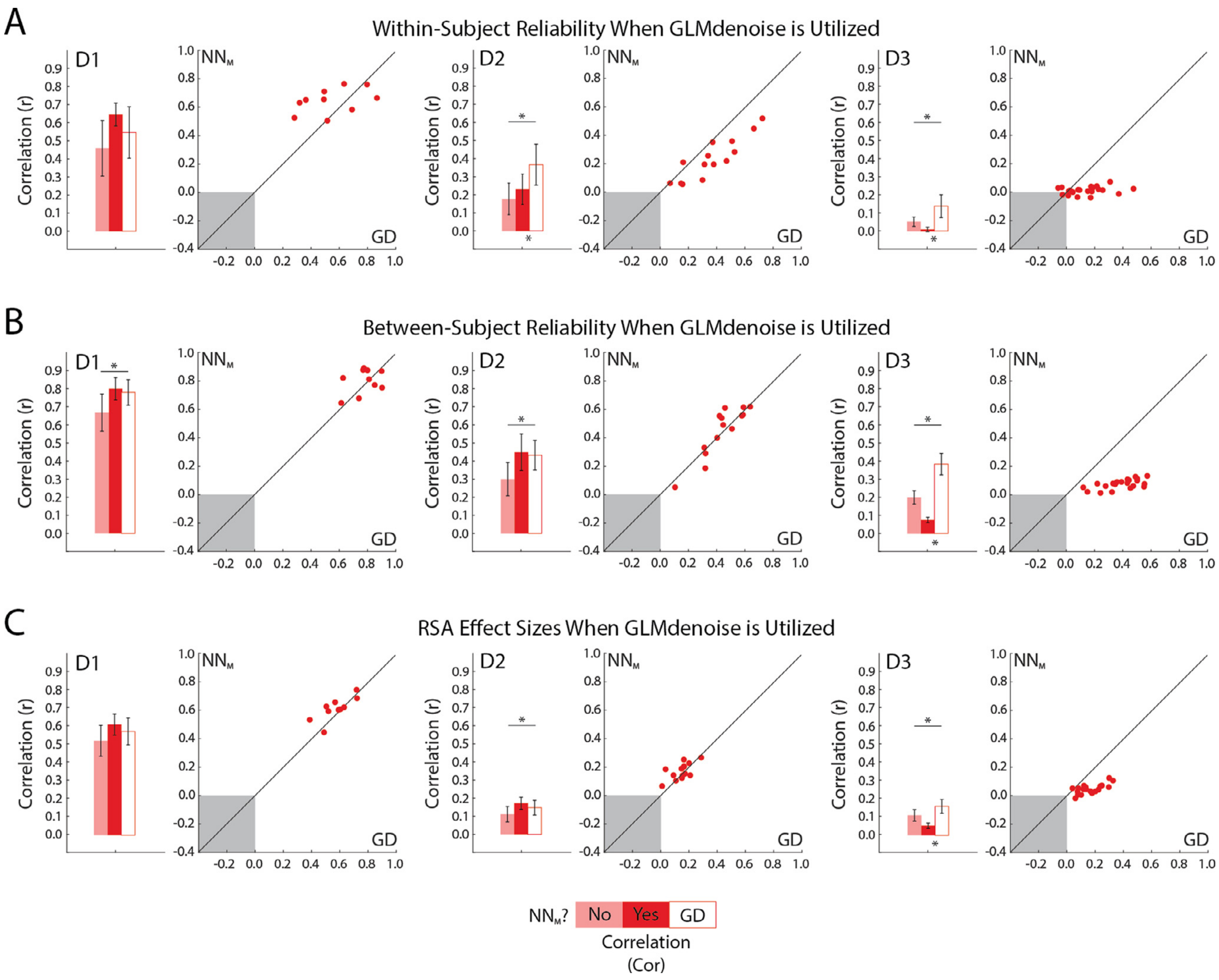

Fig. 6. GLMdenoise has a consistently more beneficial impact on neural RDM reliability and RSA effect sizes than multivariate noise normalization. (A) Bars indicate the group mean of within-subject split-half reliabilities when Cor is the dissimilarity metric either with or without NN_M, or with GLMdenoise (GD). Error bars are the standard error of the mean. (B) Bars indicate the group mean of the between-subject reliability or "noise ceiling". (C) Bars indicate the mean rank-order correlations between individual neural RDMs and target model RDMs for each dataset. All conventions are the same as in Fig. 1A.

trices relative to those from SPM. Since of primary interest was the difference afforded by GLMdenoise, we did not model the results with LME models and instead simply compared the pairwise relationships using t-tests. The baseline and NN_M results in the accompanying figures are the same as plotted in Fig. 2 and are replotted below for visual comparison.

When it came to within-subject reliability, for D1 GLMdenoise did not significantly increase the mean within-subject reliability relative to the mean correlations obtained either with or without NN_M (Fig. 6A). Since NN_M did improve the within-subject reliability before (Fig. 2A), we cannot entirely rule out a positive impact of GLMdenoise and a false negative test result (Nieuwenhuis, Forstmann, and Wagenmakers, 2011). For D2, and most notably for D3, the results were by comparison unambiguously positive: GLMdenoise significantly improved mean reliability compared to baseline and in contrast to NN, which had no effect or made reliability worse (Fig. 6A). For both D1 and D2, GLMdenoise significantly improved the between-subject reliability of neural RDMs compared to baseline, and to a similar level as NN_M. For D3, GLMdenoise again provided a substantive improvement in reliability relative to baseline unlike NN_M (Fig. 6C). Finally, for RSA effect sizes, the impact of GLMdenoise was mixed (Fig. 6C). For D1 the mean effect was not significantly higher than baseline or lower than what was obtained with NN_M, though again, the difference between baseline and NN_M was significant (Fig. 2C). For D2, while the improvement afforded by GLMdenoise was significantly higher than baseline, and not significantly different than the mean effect obtained after NN_M. Once more and most notably, GLMdenoise significantly improved the effect sizes for D3 relative to the mean effect obtained with NN_M and relative to baseline.

When NN_M was carried out with the GLMdenoise design matrix (Fig. 7A), there was still a significant increase in within-subject reliability for D1 when Euc and Mal were used as metrics, but the improvement was not significantly different than the mean reliabilities when NN_M was performed with the SPM design matrix. For D2, there were only improvements for Euc, but these were again not significantly different than those obtained using the SPM design matrices. Finally, for D3 reliability was not significantly higher than for the SPM design matrix, while there was still a significant drop relative to baseline when Cor was used as a metric. When compared to the results depicted in Fig. 6, it is notable that for D2 and D3 there was no longer a substantial improvement when using Cor as a metric.

Given that the only alteration of note to the analysis was the expansion of the design matrix, these results suggests that the improvement provided by GLMdenoise reported above (Fig. 6) may have less to do with the data-driven method for deriving noise regressors, and more to do with the fact that the analysis derives single estimates per condition

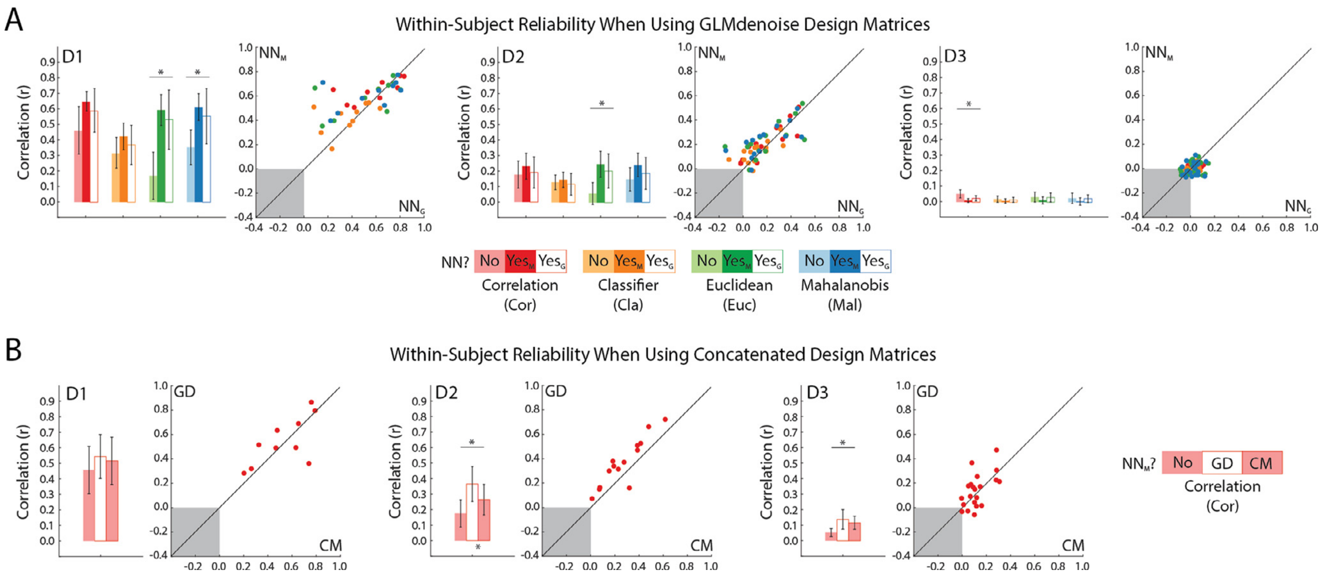

Fig. 7. Using different design matrices to assess within-subject reliability. (A) Bars indicate the group mean of within-subject split-half reliabilities either without NN_M, or with NN_M using either the SPM design matrix (Yes/ NN_M) or the GLMdenoise design matrix (Yes_G/NN_G). (B) Bars indicate the group mean of within-subject split-half reliabilities when Cor is the dissimilarity metric either without NN_M (No) with GLMdenoise (GD), or with concatenated SPM design matrices (CM). All conventions are otherwise the same as in Fig. 1A.

due to concatenation across runs of the experimental condition predictors. For if the apparent benefit of GLMdenoise was due to the superior noise estimates, then one would predict a similar (though perhaps diminished) improvement in reliability for D2 and D3 even when averaging across multiple runs when using Cor as the metric. The fact that the results for this measure differ between Fig. 5 and Fig. 6A suggest that this may not the case. To test this, we altered the SPM design matrices for the odd and even partitions so that they also concatenated across runs, resulting in single beta estimates per condition for each partition (Fig. 6B. While for D1 there was again no significant increase in within-subject reliability relative to baseline the mean after GLMdenoise, there was a significant increase for D2 and D3 (though in the former case, less than what was obtained with GLMdenoise). Thus, the relative improvement observed with GLMdenoise compared to the results obtained with NN_M may partially depend on differences in the structure of the design matrix. Though we also note that Charest et al. (2018) also concatenated their design matrices across runs, and still found a relative increase in within-subject reliability when using GLMdenoise.

3.8. *The covariance structure derived from GLM residuals is related to the distance between voxels, but not goodness-of-fit*

The results so far invite the question: what sort of structure do the residual variance-covariance matrices have and how might it predict the impact of NN_M? We carried two sets of exploratory analyses as a step towards answering this question.

First, a fundamental assumption of NN_M is that the noise information contained in the residuals of a GLM is spatially distributed, which suggests that the noise covariance may reflect the relative position of voxels inside a volume. Such a possibility is consistent with a high-degree of between-run reliability, across datasets, in the residual covariance values for the initial trio of ROIs (Fig. 6A). For each subject we constructed matrices of the exponentially decaying distances between voxels and correlated them with the off-diagonal values of the run-specific variance-covariance matrices. The resulting coefficients were then averaged across runs. We found that for all three datasets the distances matrices on averaged significantly correlated with the covariance values (Fig. 8A). This was especially true for D3, where over a third of the variance was on averaged accounted for by the distance relations between voxels. Though these correlations may be inflated due to the spatial upsampling that was performed for that dataset. While in all cases the ROIs were bilateral, the distance between the hemispheric components are much farther for LOTC and TPJ and as these ROIs were defined in terms of functional contrasts, they also tended to be asymmetric. Both properties can be seen in Fig. 8B, which depicts the mean variance-covariance matrix for a representative subject from D1-D3 and the corresponding decaying distance matrix. Especially for the subjects from D2 and D3 the spatial structure component of the covariance values can be seen from visual inspection alone.

We next considered how the correspondence between the distance between voxels and the covariance of their residuals might relate to any improvement in within-subject reliability resulting from the application of NN_M. We focused on the results obtained with Euc since NN_M produced the most substantial improvement when Euc was used as a metric (Fig. 1A). For each subject we calculated the change in within-subject reliability after NN_M by subtracting the baseline correlation values. As anticipated by the results depicted in Fig. 2A and 8A, when data was pooled across datasets the change in reliability due to NN_M negatively correlated with the correspondence between voxel distance and covariance (Fig. 8C). We also correlated the changes in within-subject reliability with the ratio between voxels and conditions, which again showed a negative correlation (Fig. 8C). Finally, when the changes in reliability were correlated with the voxel/condition ratios for each subject, there was a positive correlation (Fig. 8C).

Second, the structure of the residual variance-covariance matrices is dependent on how well the constructed GLMs account for the signal fluctuation of each voxel. Thus, it is possible that the impact of NN_M may be related to the GoF of the voxels in an ROI. For each of the initial ROIs, we determined the proportion of voxels in an ROI where the explained variance of the BOLD signal was greater than the unexplained variance ($SNR_{mb} > 1$). We also constructed dissimilarity values based on the absolute difference in GoF between voxels, which was correlated with the run-averaged residual variance-covariance matrix of each subject.

For all three datasets, the proportion of voxels with greater explained BOLD signal variance was < 0.5, suggesting that for the majority of voxels there was greater unexplained variance in the signal (Fig. 8D). However, only for D1 was there any significant positive correlation between the dissimilarities of GoF and the pairwise residual covariance between

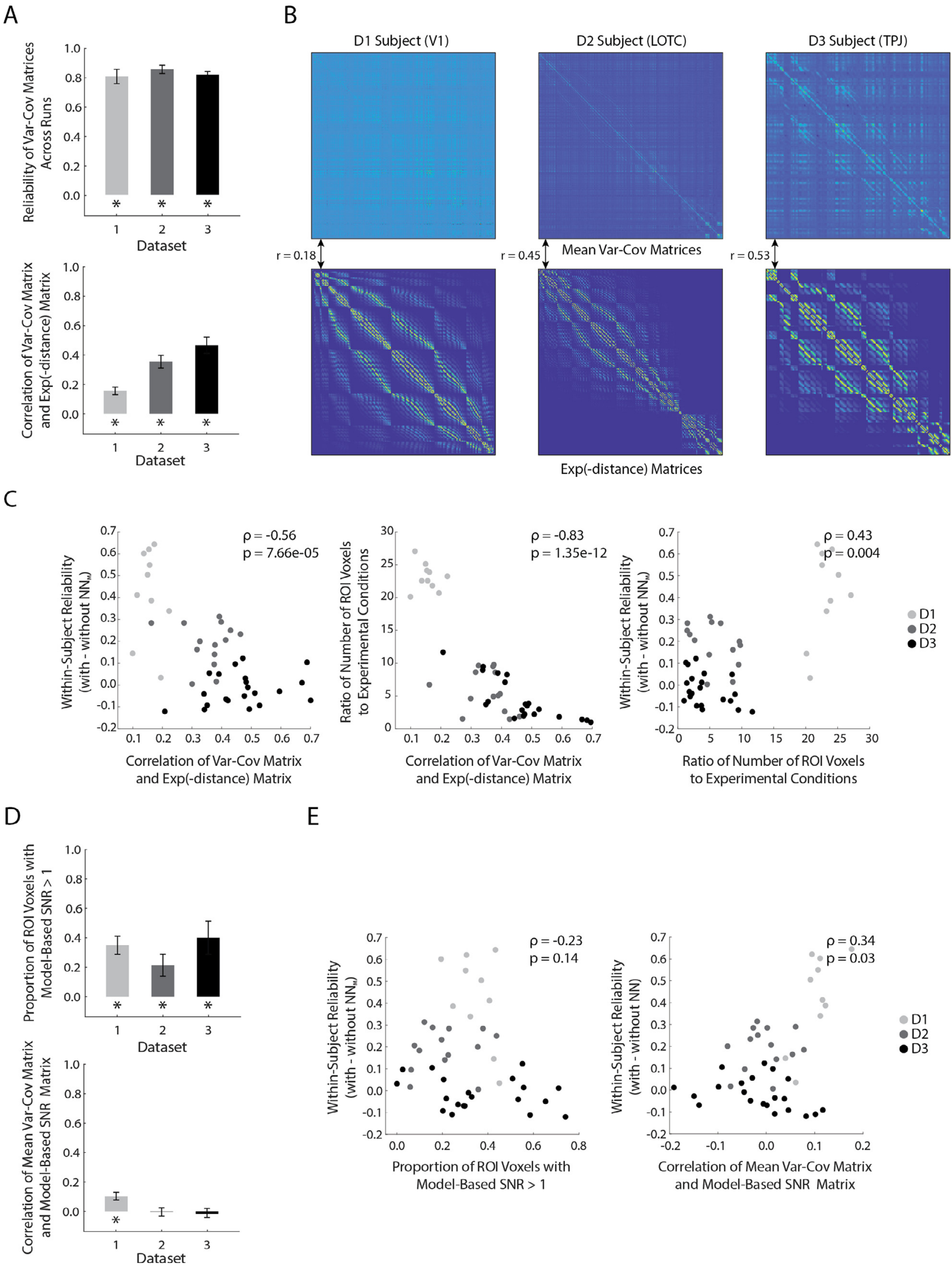

Fig. 8. Exploratory analysis of the structure of residual variance-covariance matrices. (A) Mean between-run reliability of variance-covariance (var-cov) matrices and their mean correlation with distance metrices. * = $p < 0.05$ based on two-sided paired t-tests. Error bars are the standard error of the mean (SEM). (B) Matrices for representative subjects for each dataset with correlations closest to the group average (correlations shown are for the mean var-cov matrices that are depicted). Matrices have been rescaled to arbitrary units (0 – 100) and scales are labeled to conceptually distinguish between them. (C) Scatter plots depicting the correlations between variables related to structure of var-cov matrices (data pooled across datasets D1-D3). (D) Mean values related to the model-based SNR. Error bars are the standard error of the mean (SEM). (E) Scatter plots depicting correlations between values related to the model-based SNR and impact of NN_M on within-subject reliability.

voxels (Fig. 8D). Nor did the proportion of voxels with $\mathrm{SNR}_{mb} > 1$ predict individual variation in the change in within-subject reliability afforded by $\mathrm{NN_M}$ when pooling across datasets (Fig. 8E). However, there was a positive correlation between changes in within-subject reliability due to $\mathrm{NN_M}$ and the correlations between the variance-covariance matrices and SNR_{mb} matrices.

Although we again emphasize that these analyses were exploratory, these preliminary results are nonetheless still instructive. First, they confirm that part of the covariance structure is well captured by the similarity in position of voxels within a volume. Second, they are suggestive that these spatial relations may be predictive of the influence of $\mathrm{NN_M}$ on within-subject reliability. Third, they are also suggestive that the relationship between the number of voxels and conditions, or features and samples, which influences the invertibility of the variance-covariance matrices, may also be predictive of whether $\mathrm{NN_M}$ has a beneficial effect. In particular, having more features per sample may partially account for whether $\mathrm{NN_M}$ improves reliability. Finally, GoF did not provide a good predictor of either the covariance structure between voxels or the impact of $\mathrm{NN_M}$ on within-subject reliability. Based on these results, future research might more systematically investigate how the covariance structure derived from residuals may vary across datasets and ROIs within datasets.

4. Discussion

$\mathrm{NN_M}$ has been billed as a useful method for improving the within-subject reliability of neural RDMs irrespective of the choice of dissimilarity metric (Guggenmos et al. 2018; Kriegeskorte and Diedrichsen, 2019; Nili et al. 2020; Walther et al. 2016). At the same time, other results suggest that it may be less beneficial than originally proposed and even inferior to GLMdenoise when it comes to boosting RDM reliability (Charest et al. 2018). We revisited this issue with three of our own datasets. We also evaluated the impact of $\mathrm{NN_M}$ on between-subject reliability, RSA effect sizes, the influence of $\mathrm{NN_M}$ across ROIs, the influence of spatial smoothing when carrying out $\mathrm{NN_M}$, and finally we compared the results obtained to those generated when using GLMdenoise. The results of our investigation were mixed. Our findings have implications for: (i) whether $\mathrm{NN_M}$ should be prescribed to boost reliability when performing RSA; (ii) whether developing pipeline tools like $\mathrm{NN_M}$ is well-motivated in the first place; and (iii) how the choice of whether to use $\mathrm{NN_M}$ may depend on differing methodological and theoretical motivations for using RSA.

4.1. Multivariate noise normalization and principles of beneficence and non-maleficence in data processing

$\mathrm{NN_M}$ comes strongly recommended. Walther et al. (2016, p.197) state that: “Activation patterns (usually formed by regression coefficients) should be subjected to multivariate noise normalization to improve RDM reliability, regardless of dissimilarity measure.” Similarly, Guggenmos et al. (2018, p.444) offer that for time-series RSA: “multivariate noise normalization is a highly recommended preprocessing step irrespective of other analytic choices.” Do our results support these prescriptions?

Before answering this question, it is worth highlighting that any prescribed intervention for data analysis should meet two normative principles, which mirror those used to evaluate the ethical implications of biomedical research (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979). The first is one of *beneficence*: does the intervention tend to improve the signal estimate for the purpose of detecting the effects of interest? Clearly promotion of the effectiveness of $\mathrm{NN_M}$ has been largely centered on whether the method is beneficial in some way. However, equally important is the principle of *non-maleficence*: does the intervention tend to not worsen the signal estimate for the purpose of detecting the effects of interest? In other words, with data as with health: first, do no harm.

We believe our results call into question whether $\mathrm{NN_M}$ satisfies either principle. Therefore, they do not support the above prescriptions. Consider beneficence first. The original finding of Walther et al. (2016) was that $\mathrm{NN_M}$ tended to improve within-subject reliability, however, it is unclear that such a finding provides sufficient evidence of a benefit for RSA. First, they did not show that it had any discernable influence on the between-subject reliability or RSA effect sizes. Second, they did not show that it had a consistent benefit across multiple ROIs within the dataset, even though any use of $\mathrm{NN_M}$ would presumably be applied uniformly when a study considers multiple regions throughout the brain. When we carried out these further analyses, across just a few additional ROIs, we found that $\mathrm{NN_M}$ was only consistently beneficial when using Euc, and Cor, as a metric. It has been suggested that Cla is a less desirable metric because it is discrete (Guggenmos et al. 2018; Walther et al. 2016). In line with previous findings we found use of Cla resulted in the lower mean reliabilities and effect sizes even when $\mathrm{NN_M}$ was applied. Reliabilities and effect sizes were also not consistently improved when using its continuous cousin Mal as a metric. This latter result is significant when compared to those obtained when using Euc paired with $\mathrm{NN_M}$, which is closest to the “crossnobis” distance that has been promoted as the preferred metric for improving reliability (Diedrichsen et al. 2016; Guggenmos et al. 2018; Nili et al. 2020; Walther et al. 2016). For we found little difference in between-subject reliability and effect sizes when using this approach vs Mal without $\mathrm{NN_M}$. Thus, our results do not support the claim that $\mathrm{NN_M}$ is of clear benefit, regardless of metric.

However, these varied positive results were only observed for D1 and D2. For when it comes to non-maleficence, the negative impact of $\mathrm{NN_M}$ across analyses for D3 was perhaps our most consistent finding. Only when using Euc and Mal as metrics to construct RDMs for BA17 did we see any positive impact of $\mathrm{NN_M}$ for this dataset. We note that this outcome may not be entirely surprising for some. The fundamental assumption of $\mathrm{NN_M}$ is that the noise information in the residuals of the GLM have a spatial component, which itself may be connected to topographic organization in a region. Our exploratory analysis showed that the structure of the covariance values for bilateral TPJ were especially well-captured by the distance between voxels within the ROI (Fig. 8A). And yet, $\mathrm{NN_M}$ consistently reduced the reliabilities and effect sizes observed for this region. This is unlikely solely to be a result of the fact that the hemispheric components for the bilateral ROI were far apart since the same was true of the LOTC ROI for D2. Instead, for D3 the spatial structure did not appear to include information that was meaningful for improving BOLD signal estimation. When stimuli are videos any such topographic response to an image frame will be combined with that for all of the other video frames, and so the noise may have a more complex spatiotemporal structure. So, it may be that $\mathrm{NN_M}$ is less suitable when dynamic stimuli are used in part because standard GLMs also do not capture the dynamic nature of video stimuli either. However, our results for D2 (SPL) also raise the question of whether $\mathrm{NN_M}$ is suitable at all when one leaves the cortical realm of sensorimotor regions. Thus, it may also be that $\mathrm{NN_M}$ is not suitable for all regions one may investigate. In which case, if $\mathrm{NN_M}$ is intended to be applied uniformly across ROIs, then it does not satisfy the principle of non-maleficence since there is the risk that in some regions one may simply be multiplying beta estimates with spatially unstructured noise.

More cautiously, our results suggest that further analytic trials would be required to evaluate the effectiveness of $\mathrm{NN_M}$ and when and where it should be utilized. In contrast, the results obtained with GLMdenoise were more reliably beneficial. Indeed, the clearest positive impact of the approach was seen for D3. Thus, our findings may be more consistent with the recommendations of Charest, Kriegeskorte, and Kay (2018) to use this boutique approach to first-level analysis prior to carrying out RSA. However here a number of considerations suggest further analysis is still necessary. First, we found that the positive influence of GLMdenoise may have more to do with the single beta estimates generated by the analysis rather than the more complex procedure for choosing noise regressors, as similar results are achievable simply by restructuring standard design matrices so that only a single beta estimate is produced. Second, GLMdenoise is not compatible with using cross-validated metrics even if one reorganizes the design matrix since the cross-validation procedure steps it depends on violate the independence between runs. As an alternative to Cor, GLMdenoise would also be compatible with using non-cross-validated forms of either Mal and Euc as metrics (Guggenmos et al. 2018; Ritchie et al. 2020), However, the fact would remain that, if one believes cross-validation provides more accurate estimates of dissimilarity relations (Bobadilla-Suarez et al. 2019; Walther et al, 2016), then GLMdenoise is not a viable analytic option. We note that leveraging other methods, such as Bayesian RSA (Cia et al. 2019) may also help to further clarify the relationship between GLMdenoise and NN_M, and more general concerns about how noise in BOLD signals might be estimated and exploited in the service of carrying out RSA.

4.2. Questioning the motivation for multivariate noise normalization

Our results provide reason to doubt whether NN_M is as widely applicable as has been proposed. However, we believe that they also point to more fundamental issues with NN_M, which call into question the motivation for introducing, and promoting, such new analytical tonics as a curative for the ills of noisy data.

The first issue is that it remains not entirely clear when and why NN_M works. The underlying assumption is that the noise structure contained in the residuals will have a spatial component that can be leveraged to improve the estimate of condition-specific variation in the BOLD signal. There are parts of the brain where such an assumption seems eminently plausible, such as early visual or motor cortex, and indeed in such regions NN_M seems to produce the best results. Our preliminary results also suggest that some of the structure of the variance-covariance matrices central to NN_M simply reflects the distance between voxels. However, the reality is that without further analysis the form of the noise remains unknown, and where there is no spatial component to be found one is again simply multiplying signal estimates with unstructured noise. In such cases, NN_M may have all the benefit of bloodletting. The risk posed by the unknown is not unique to NN_M. In recommending GLMdenoise as a first level analysis for RSA, Charest et al. also acknowledge that, since GLMdenoise is data driven, the noise that is removed by the approach is left unspecified unless further analyses are taken. They further acknowledge that the noise may vary across experiments and participants. The risk with NN_M is that, since it is not data-driven, it may even vary across areas and stimulus types. Furthermore, whereas GLMdenoise is part of an open source Matlab toolbox, the application of NN_M is not as yet standardized – though one implementation can be found in the Decoding Toolbox (Hebart et al. 2015).

The fact that NN_M may inject even more uncertainty into how we interpret neuroimaging results is problematic when we consider the trend towards more transparent, replicable processing pipelines (Esteban et al. 2019). This point is well-illustrated by the results of Botvinik-Nezer et al. (2020) who investigated the effect of pipeline flexibility on neuroimaging findings. In their study, 70 teams analyzed the same neuroimaging dataset to test the same collection of hypotheses. The variation in analysis approaches, and subsequent results, were striking: no two teams used identical pipelines, and even in cases where the statistical maps of the brain were correlated in some stages of the processing pipeline the reported significant results still differed. Of the nine hypotheses tested, there was near consensus of a negative results for three of them (< 10 % of teams found an effect) while there was majority agreement of a significant result for only one of them > 80 % of teams reported an effect). For the remaining five only ~ 20 – 40 % of teams reported significant results. Now imagine a similar study where RSA was ultimately the method of choice. In such a case, NN_M would introduce yet another degree of freedom into the choice of pipeline where the appropriateness of its application would still remain uncertain.

The preceding hypothetical helps to motivate the second issue, which is whether NN_M is intended to be a method that is supposed to impact what effects are detected. On the one hand, the cautious response may be to say that it is simply intended as a method to improve within-subject reliability. However, if NN_M has no material impact on the explainable variance, then it seems we have little motivation to apply the analysis at all. But if the goal is to increase the explainable variance, then this is itself only of interest because it might make a difference to what effects are observed. In which case, the utility of NN_M is after all because it might change what effects are observed. On the other hand, if the goal of NN_M as an intervention is to possibly influence what effects are found, then the known unknowns about NN_M are again a source of concern. In our analysis we only used a single model RDM to assess the impact of NN_M on RSA effect sizes. However, it is typical to test multiple model RDMs, with models tested against the null hypothesis, against each other, or jointly used to model dissimilarity values using multiple regression or "variance partitioning" variants such as commonality analysis (Groen et al. 2018; Hebart et al. 2018; Newton and Spurell, 1967). Therefore, it is as yet unknown whether carrying out NN_M might boost some effects at the expense of others.

This second issue does not mitigate against the utility of NN_M, but we do believe it is worth highlighting for researchers deliberating on whether to apply the method. A different perspective is to favor carrying out multiple analyses for a single dataset to show that some target result is robust across all of them in order to improve methodological transparency. For example, Steegen et al. (2016) propose a "multiverse" approach where one tests for results based on all datasets that are generated across all possible combinations of data processing choices. In one application of this approach, Moors and Hesselmann (2019) found that only 14% of pipelines revealed apparent effects when analyzing a dataset for evidence of unconscious arithmetic. Applying the same logic to pipelines for RSA, NN_M is again one more degree of freedom for analysis pipelines that one might implement. But whether a desired effect is observed should not depend on whether NN_M is carried out but instead should be robust across many data processing choices. If we attempt to follow this alternative approach, then one can again wonder whether even in principle prescriptions like NN_M are well motivated.

4.3. Distinguishing "modest" vs "ambitious" applications of representational similarity analysis

Despite being relatively critical of both the prescriptions to use NN_M and its underlying motivation, we are nonetheless reluctant to recommend *not* using NN_M. In their discussion, Walther et al. suggest that the choice of dissimilarity metric may depend on the question one is asking: is the goal to determine how discriminable patterns are or their dissimilarity regardless of its shape? In the former case, it may be more desirable to use a distance metric like Mal, in the latter, Cor. Similarly, we believe that the choice of whether or not one should use NN_M may depend on the research question and the precise way in which RSA is being employed.

NN_M has been introduced as a salve for improving signal estimates and thus the reliability of neural RDMs. As we have seen, however, its effectiveness is plausibly restricted by whether we have prior reason to believe that the noise contained in the residuals of a GLM is spatially structured. However, whether or not this is the case is more akin to an issue of measurement than of reliability. As pointed out by Bodadilla-Saurez et al. (2019), when it comes to RSA, and our choice of dissimilarity metric, the issue of measurement is not the same as that of RDM reliability. Different dissimilarity metrics are not simply different in terms of how reliable they are in their estimates, but also in terms of what kinds of relationships they estimate. Depending on the sort of similarity structure one is hypothesizing may be latent in a brain region, different metrics will be appropriate (Ramirez, 2017). This consideration points to a more fundamental distinction, which is that theories are not built upon the backs of data; rather, our choice of analyzes of data are revealing of phenomenon that we wish to explain (Bogen and Woodward, 1988; Woodward, 2011). This distinction, emphasized in the philosophy of science, is also germane to neuroimaging methods like RSA (Carlson et al. 2018). NN_M is not simply method for improving reliability but depends on assumptions about the form of the signal being detected and its spatial extent. So, whether NN_M is applicable will, as with choice of metric, depend on prior assumptions about the signal being measured and the hypotheses being tested.

Here we believe a distinction between two types of applications of RSA may be helpful in researchers deciding whether to utilize NN_M. As we emphasized at the outset, RSA has both methodological and theoretical virtues. Many uses of RSA clearly aim to take advantage of the former: comparing neural RDMs from multiple brain regions to those derived from computational models or behavior. In such cases we suspect that researchers are content to use Cor as a measure since whether dissimilarity values solely reflect pattern discriminability, or partially depend on mean signal amplitude, may not be of interest. Instead, what is of chief importance is that RSA allows for direct comparison of many different data types where condition rich designs are used. In such "modest" applications of RSA, where the topographic structure of a large number of ROIs is likely unknown, we believe NN_M may do more harm than good, and GLMdenoise may potentially be more appropriate for boosting the estimate of signal estimates. insofar as we suspect that most studies using RSA aim for such modesty, NN_M may therefore have a limited scope of application. However, in other cases the use of RSA is driven more by its theoretical benefits. For example, studies that use RDMs to construct a low-dimensional representation of activity patterns in order to directly compare models of behavior may depend on very particular assumptions about both dissimilarity relations, but also the structure of the neural population code in a region (Davis, Love, and Preston, 2012; Op de Beeck, Wagemans, and Vogels, 2001; Ramírez, 2018; Ritchie and Op de Beeck, 2019). In such "ambitious" applications of RSA, NN_M may indeed be useful insofar as the design of such studies are attentive not just to the reliability of neural RDMs, but also precise hypotheses about the structure of the activation space that is being measured. Another case where NN_M may be useful are studies where within-subject reliability of particular importance. For example, several studies have found that individual differences in the structure of neural RDMs are reliable and meaningful (Charest et al. 2014; Feilong et al. 2018). For such applications NN_M may be particularly well-suited, depending on the experimental conditions and ROIs.

In summary, the broad prescriptions supporting the use of NN_M suggest it is appropriate for modest uses of RSA. We believe our results show that such a recommendation is not yet supported. However, more positively we believe that it may be applicable when the underlying signal, and accompanying noise, have a known structure and this fact is being leveraged in more ambitious uses of RSA.

4.4. Conclusion and future directions

NN_M has been proposed as method for boosting the reliability of neural RDMs when carrying out RSA. As we have seen, this method in fact produces mixed results based on differences between datasets that are related to stimuli and choices of regions. What factors account for these differing results remains unclear even though the datasets and analyses that we performed already provide some exploration of obvious candidates. Further research may clarify when NN is an appropriate analysis step to take. Empirically we believe promising directions for such research might include larger numbers of datasets, and comparisons of randomly selected regions of cortex as could be achieved with searchlight methods (Etzel et al. 2013). Alternatively, important insights might be gained from further simulation studies based on brain areas where the spatiotemporal properties of the BOLD signal (and noise) are well understood and can be compared to neural data. At the same time, such modeling approaches can only take us so far given that the spatiotemporal profile of the BOLD signal is likely heterogenous throughout the cortex. Were such profiles well understood, it would likely obviate the need for a method like NN_M in the first place. Furthermore, as we have argued, the very motivation for NN_M and its prescriptive nature can also be questioned.

Data and code availability statement

The (raw) fMRI data of all three datasets is available via the Open Science Framework (D1: https://osf.io/b9svy/; D2: https://osf.io/qp54f/; D3: https://osf.io/hpwjx/), as well as code for running multivariate noise normalization on a token subject from D1 (https://osf.io/hnq6m/). Further code for running the analyses reported here will be made available upon request by the first author.

Acknowledgements

J.B.R. received funding from the FWO (Fonds Wetenschappelijk Onderzoek) and European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 665501, via an FWO [PEGASUS][2] Marie Skłodowska-Curie fellowship (12T9217N). H.L.M received funding from the Marguerite-Marie Delacroix Foundation (GV/B-336). S.B. received support from an FWO Postdoctoral Fellowship Program (12S1317N) and Research Grant (1505518N). H.P.O. were funded by a grant from the KU Leuven Research Council (C14/16/031) and Excellence of Science Grant (G0E8718N).

Bibliography

Allefeld, C., Haynes, J.D., 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage 89, 345–357.

Attneave, F., 1950. Dimensions of similarity. Am. J. Psychol. 63 (4), 516–556.

Benson, N.C., Butt, O.H., Datta, R., Radoeva, P.D., Brainard, D.H., Aguirre, G.K., 2012. The retinotopic organization of striate cortex is well predicted by surface topology. Curr. Biol. 22 (21), 2081–2085.

Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., Love, B.C., 2019. Measures of neural similarity. Computational Brain & Behavior 1–15.

Bogen, J., Woodward, J., 1988. Saving the phenomena. The Philosophical Review 97 (3), 303–352.

... & Botvinik-Nezer, R., Holzmeister, F., Camerer, C.F., Dreber, A., Huber, J., Johannesson, M., Avesani, P., 2020. Variability in the analysis of a single neuroimaging dataset by many teams. Nature 1–7.

Bracci, S., de Beeck, H.O., 2016. Dissociations and associations between shape and category representations in the two visual pathways. J. Neurosci. 36 (2), 432–444.

Carlson, T., Goddard, E., Kaplan, D.M., Klein, C., Ritchie, J.B., 2018. Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. Neuroimage 180, 88–100.

Charest, I., Kievit, R.A., Schmitz, T.W., Deca, D., Kriegeskorte, N., 2014. Unique semantic space in the brain of each beholder predicts perceived similarity. Proc. Natl. Acad. Sci. 111 (40), 14565–14570.

Charest, I., Kriegeskorte, N., Kay, K.N., 2018. GLMdenoise improves multivariate pattern analysis of fMRI data. Neuroimage 183, 606–616.

Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O., 2010. Shrinkage algorithms for MMSE covariance estimation. IEEE Trans. Signal Process. 58 (10), 5016–5029.

Cai, M.B., Schuck, N.W., Pillow, J.W., Niv, Y., 2019. Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. PLoS Comput. Biol. 15 (5), e1006299.

Cronbach, L.J., 1949. Essentials of Psychological Testing. Harper, New York.

Davis, T., Love, B.C., Preston, A.R., 2012. Learning the exception to the rule: Model-based fMRI reveals specialized representations for surprising category members. Cereb. Cortex 22 (2), 260–273.

Diedrichsen, J., Provost, S., Zareamoghaddam, H., 2016. On the distribution of cross-validated Mahalanobis distances. arXiv preprint arXiv:1607.01371.

... & Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Oya, H., 2019. fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat. Methods 16 (1), 111–116.

Etzel, J.A., Zacks, J.M., Braver, T.S., 2013. Searchlight analysis: promise, pitfalls, and potential. Neuroimage 78, 261–269.

Feilong, M., Nastase, S.A., Guntupalli, J.S., Haxby, J.V., 2018. Reliable individual differences in fine-grained cortical functional architecture. Neuroimage 183, 375–386.

Friston, K.J., Jezzard, P., Turner, R., 1994. Analysis of functional MRI time-series. Hum. Brain Mapp. 1 (2), 153–171.

Groen, I.I., Greene, M.R., Baldassano, C., Fei-Fei, L., Beck, D.M., Baker, C.I., 2018. Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. Elife 7, e32962.

Guggenmos, M., Sterzer, P., Cichy, R.M., 2018. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. Neuroimage 173, 434–447.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293 (5539), 2425–2430.

Hebart, M.N., Görgen, K., Haynes, J.D., 2015. The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. Frontiers in neuroinformatics 8, 88.

Hebart, M.N., Bankson, B.B., Harel, A., Baker, C.I., Cichy, R.M., 2018. The representational dynamics of task and object processing in humans. Elife 7, e32816.

Hendriks, M.H., Daniels, N., Pegado, F., Op de Beeck, H.P., 2017. The effect of spatial smoothing on representational similarity in a simple motor paradigm. Fronti. Neurol. 8, 222.

Kay, K., Rokem, A., Winawer, J., Dougherty, R., Wandell, B., 2013. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Fronti. Neurosci. 7, 247.

Kietzmann, T.C., Spoerer, C.J., Sörensen, L.K., Cichy, R.M., Hauk, O., Kriegeskorte, N., 2019. Recurrence is required to capture the representational dynamics of the human visual system. Proc. Natl. Acad. Sci. 116 (43), 21854–21863.

Kriegeskorte, N., Diedrichsen, J., 2019. Peeling the onion of brain representations. Annu. Rev. Neurosci. 42, 407–432.

Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17 (8), 401–412.

Kriegeskorte, N., Mur, M., Bandettini, P.A., 2008a. Representational similarity analysis–connecting the branches of systems neuroscience. Frontiers in syst. Neurosci. 2, 4.

… & Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Bandettini, P.A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60 (6), 1126–1141.

Ledoit, O., Wolf, M., 2004. A well-conditioned estimator for large-dimensional covariance matrices. J. Multivariate Anal. 88 (2), 365–411.

Lee Masson, H., Op de Beeck, H., 2018. Socio-affective touch expression database. PLoS One 13 (1), e0190921.

Lee Masson, H., Van De Plas, S., Daniels, N., Op de Beeck, H., 2018. The multidimensional representational space of observed socio-affective touch experiences. Neuroimage 175, 297–314.

Liu, P., Chrysidou, A., Doehler, J., Hebart, M., Wolbers, T., Kuehn, E., 2021. The organizational principles of de-differentiated topographic maps in somatosensory cortex. Elife 10, e60090.

Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53 (1), 103–118.

Moors, P., Hesselmann, G., 2019. Unconscious arithmetic: Assessing the robustness of the results reported by Karpinski, Briggs, and Yale (2018). Conscious. Cogn. 68, 97–106.

National commission for the protection of human subjects of biomedical and behavioral research, 1979. The Belmont report: Ethical principles and guidelines for the protection of human subjects of research 45.

Newton, R.G., Spurrell, D.J., 1967. Examples of the use of elements for clarifying regression analyses. J. Royal Statist. Soc.: Series C (Applied Statistics) 16 (2), 165–172.

Nieuwenhuis, S., Forstmann, B.U., Wagenmakers, E.J., 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. Nat. Neurosci. 14 (9), 1105–1107.

Nili, H., Walther, A., Alink, A., Kriegeskorte, N., 2020. Inferring exemplar discriminability in brain representations. PLoS One 15 (6), e0232551.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A toolbox for representational similarity analysis. PLoS Comput. Biol. 10 (4), e1003553.

Oosterhof, N.N., Connolly, A.C., Haxby, J.V., 2016. CoSMoMVPA: multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. Frontiers in neuroinformatics 10, 27.

Op de Beeck, H.P., 2010. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage 49 (3), 1943–1948.

Op de Beeck, H.P., Deutsch, J.A., Vanduffel, W., Kanwisher, N.G., DiCarlo, J.J., 2008. A stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex. Cereb. Cortex 18 (7), 1676–1694.

Op de Beeck, H., Wagemans, J., Vogels, R., 2001. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nat. Neurosci. 4 (12) 1244-12.

Ramírez, Fernando M., 2017. Representational confusion: the plausible consequence of demeaning your data. bioRxiv, 195271.

Ramírez, F.M., 2018. Orientation encoding and viewpoint invariance in face recognition: inferring neural properties from large-scale signals. Neuroscientist 24 (6), 582–608.

Ritchie, J.B., Bracci, S., de Beeck, H.O., 2017. Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. NeuroImage 148, 197–200.

Ritchie, J.B., Op de Beeck, H., 2019. A varying role for abstraction in models of category learning constructed from neural representations in early visual cortex. J. Cogn. Neurosci. 31 (1), 155–173.

Ritchie, J.B., Zeman, A.A., Bosmans, J., Sun, S., Verhaegen, K., de Beeck, H.P.O., 2020. Untangling the animacy organization of occipitotemporal cortex. bioRxiv.

Saxe, R., Kanwisher, N., 2003. People thinking about thinking people: the role of the temporo-parietal junction in “theory of mind. Neuroimage 19 (4), 1835–1842.

Shepard, R.N., 1964. Attention and the metric structure of the stimulus space. J. Math. Psych. 1 (1), 54–87.

Soch, J., Allefeld, C., 2018. MACS–a new SPM toolbox for model assessment, comparison and selection. J. Neurosci. Methods 306, 19–31.

Steegen, S., Tuerlinckx, F., Gelman, A., Vanpaemel, W., 2016. Increasing transparency through a multiverse analysis. Perspectives on Psychol. Sci. 11 (5), 702–712.

van Bergen, R.S., Jehee, J.F., 2018. Modeling correlated noise is necessary to decode uncertainty. Neuroimage 180, 78–87.

van Meel, C., de Beeck, H.P.O., 2020. An investigation of the effect of temporal contiguity training on size-tolerant representations in object-selective cortex. Neuroimage, 116881.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., Diedrichsen, J., 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage 137, 188–200.

Welvaert, M., Rosseel, Y., 2013. On the definition of signal-to-noise ratio and contrast–to-noise ratio for FMRI data. PLoS One 8 (11), e77089.

Woodward, J.F., 2011. Data and phenomena: a restatement and defense. Synthese 182 (1), 165–179.