NeuroImage 180 (2018) 41–67

Contents lists available at ScienceDirect

NeuroImage

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

Interpreting the dimensions of neural feature representations revealed by dimensionality reduction

Erin Goddard [a,b,c,*], Colin Klein [c,d], Samuel G. Solomon [e], Hinze Hogendoorn [b,f], Thomas A. Carlson [b,c]

[a] McGill Vision Research, Dept of Ophthalmology, McGill University, Montreal, QC, H3G 1A4, Canada
[b] School of Psychology, University of Sydney, Sydney, NSW, 2006, Australia
[c] ARC Centre of Excellence in Cognition and Its Disorders (CCD), Macquarie University, Sydney, NSW, 2109, Australia
[d] Department of Philosophy, Macquarie University, Sydney, NSW, 2109, Australia
[e] Department of Experimental Psychology, University College London, Gower Street, London, WC1E 6BT, United Kingdom
[f] Helmholtz Institute, Neuroscience & Cognition Utrecht, Experimental Psychology Division, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Keywords:
Multivariate pattern analysis
Multi-dimensional scaling (MDS)
Principal component analysis (PCA)
Exploratory analysis

ABSTRACT

Recent progress in understanding the structure of neural representations in the cerebral cortex has centred around the application of multivariate classification analyses to measurements of brain activity. These analyses have proved a sensitive test of whether given brain regions provide information about specific perceptual or cognitive processes. An exciting extension of this approach is to infer the structure of this information, thereby drawing conclusions about the underlying neural representational space. These approaches rely on exploratory data-driven dimensionality reduction to extract the natural dimensions of neural spaces, including natural visual object and scene representations, semantic and conceptual knowledge, and working memory. However, the efficacy of these exploratory methods is unknown, because they have only been applied to representations in brain areas for which we have little or no secondary knowledge. One of the best-understood areas of the cerebral cortex is area MT of primate visual cortex, which is known to be important in motion analysis. To assess the effectiveness of dimensionality reduction for recovering neural representational space we applied several dimensionality reduction methods to multielectrode measurements of spiking activity obtained from area MT of marmoset monkeys, made while systematically varying the motion direction and speed of moving stimuli. Despite robust tuning at individual electrodes, and high classifier performance, dimensionality reduction rarely revealed dimensions for direction and speed. We use this example to illustrate important limitations of these analyses, and suggest a framework for how to best apply such methods to data where the structure of the neural representation is unknown.

1. Introduction

Neuroimaging and multielectrode recordings enable simultaneous measurement from neuronal populations. Collecting such measurements for a large stimulus set produces large, multidimensional data sets. To effectively extract meaningful information about the brain from these rich data sets one must find ways to summarize the information, and do so without obscuring the rich relationships in the data that these methods are designed to reveal. One family of approaches to summarizing complex data sets is dimensionality reduction methods, which re-represent multi-dimensional data in a space defined by fewer dimensions than the original data. Common examples of dimensionality reduction methods include principal component analysis (PCA), multi-dimensional scaling (MDS), and cluster analyses.

For large data sets, dimensionality reduction can be an effective way of summarizing and visualizing population neural activity (for example, Mazor and Laurent, 2005; Stokes et al., 2013). This allows for quick sanity checks of the data, and can increase statistical power compared with simple averaging across trials or electrodes/voxels (Cunningham and Yu, 2014). Dimensionality reduction can also be helpful for navigating intractably large stimulus spaces, and for generating models of such spaces (Adolphs et al., 2016). These uses exemplify the strengths of dimensionality reduction for summarizing data in a more accessible format.

* Corresponding author. McGill Vision Research, Dept of Opthalmology, McGill University, Montreal, QC, H3G 1A4, Canada.
E-mail address: erin.goddard@mcgill.ca (E. Goddard).

http://dx.doi.org/10.1016/j.neuroimage.2017.06.068
Received 23 May 2017; Accepted 23 June 2017
Available online 27 June 2017

Table 1
Summary of the dimensionality reduction methods from the Matlab function *rotatefactors*, the FastICA package, and the Matlab Toolbox for Dimensionality Reduction. For methods with free parameters, the values selected are shown in the column 'Parameter Values'. In every case the default parameter values were used. The setting 'Normalize' = 'on' indicates that the rows of the PCA components were normalized to have a unit Euclidean norm prior to rotation, then unnormalized after rotation. The variable k indicates the number of nearest neighbors in a neighborhood graph. The variable sigma indicates the variance of a Gaussian kernel. For descriptions of the remaining parameters, see the Matlab Toolbox for Dimensionality Reduction.

Model name abbreviation	Full model name	Parameter values
PCA	Principal Component Analysis	N/A
MDS	Multi-Dimensional Scaling	N/A
Varimax	Varimax rotation on PCA components	'Normalize' = 'on'
Quartimax	Quartimax rotation on PCA components	'Normalize' = 'on'
Parsimax	Parsimax rotation on PCA components	'Normalize' = 'on'
FastICA	Fast fixed-point algorithm for Independent Component Analysis	'type' = 'kurtosis'
MaxKurtosisICA	Kurtosis-maximizing Independent Component Analysis	N/A
Isomap	Isomap	k = 12
LLE	Locally Linear Embedding	k = 12
LDA	Linear Discriminant Analysis	N/A
ProbPCA	Probabilistic Principal Component Analysis	max_iterations = 200
FactorAnalysis	Factor Analysis	N/A
GPLVM	Gaussian Process Latent Variable Model	sigma = 1
Sammon	Sammon mapping	N/A
LandmarkIsomap	Landmark Isomap	k = 12; percentage = 0.2
Laplacian	Laplacian Eigenmaps	k = 12; sigma = 1
HessianLLE	Hessian Locally Linear Embedding	k = 12
LTSA	Local Tangent Space Alignment	k = 12
DiffusionMaps	Diffusion maps	t = 1; sigma = 1
KernelPCA	Kernel Principal Component Analysis	kernel = 'gauss'
SNE	Stochastic Neighbor	perplexity = 30
SymSNE	Symmetric Stochastic Neighbor Embedding	perplexity = 30
tSNE	t-Distributed Stochastic Neighbor Embedding	perplexity = 30; initial_dims = 30
LPP	Locality Preserving Projection	k = 12; sigma = 1
NPE	Neighborhood Preserving Embedding	k = 12
LLTSA	Linear Local Tangent Space Alignment	k = 12
Autoencoder	Deep autoencoders	lambda = 0
NCA	Neighborhood Components Analysis	lambda = 0
MCML	Maximally Collapsing Metric Learning	N/A
LMNN	Large Margin Nearest Neighbor metric learning	k = 3

A further, and more contentious, use of dimensionality reduction is to infer something about the how the brain itself represents the world. Before we proceed, it is important to distinguish three related but importantly different concepts: *features, feature spaces,* and *representational spaces* in the brain. *Features* are properties of stimuli. Features can be physical properties of stimuli, e.g. color, spatial frequency, motion direction. They also can also be psychological constructs based on theory and behavior, e.g. the constructs of valence and arousal in emotion perception. A *feature space* is a multidimensional model in which feature values correspond to coordinates in the space. Where a feature space has defined dimensions, any novel stimulus may be assigned a location (or locations) within the space based on its features; and for each point in the space a stimulus with those feature values could be constructed. Feature spaces can vary in how succinctly and intuitively they organize stimuli, but there will often be multiple equally parsimonious feature spaces that provide a good account of the stimulus set. For example, colors that are discriminable to human observers can be captured in one of many different three-dimensional features spaces: for example, the RGB space of a computer display, or HSL (hue, saturation, lightness) space. These color spaces may be suitable or not for a particular task, but are equally valid as feature spaces. Importantly, feature models describe what is being represented, but this may or may not bear any resemblance to the way the brain actually represents information.

A key challenge for cognitive neuroscience is to understand how the brain represents this information, and so to infer the structure of *representational spaces* in the brain. A *representational space* is the feature space that corresponds to how a brain region is representing a given set of stimuli under specific task conditions, where neural activity varies in predictable ways along the dimensions of the space. If a feature space defines *what* an organism is representing, the representational space defines *how* the brain is representing this information. If we were to accurately map the representational space of a given brain region we should be able to predict the response of the region to practically limitless variation in stimulus. Hypothesis-driven investigation of representational spaces choses a small set of feature dimensions and uses them to construct a set of stimuli, with the aim of characterizing how the brain represents stimuli along these dimensions. This leads to the concern that hypothesis-driven approaches are only ever testing a small subset of any possible feature space. Further, the way the *brain* carves up stimuli may differ to how we find it natural to do so, and so large portions of feature space may go unexplored.

This has motivated the use of 'data-driven' approaches for defining the feature dimensions that are of relevance to the brain. In this context, dimensionality reduction approaches have been employed to 'discover' the brain's representational space. This is an attractive concept since it opens the possibility of circumventing the need to define stimulus dimensions a priori, and allows the generation of data that are not tied to a particular model of the feature space. These data-driven approaches have greatest potential for higher-order brain regions, where the natural dimensions of the feature space are unknown.

One field of research where such approaches have gained popularity is that of visual object recognition. Kriegeskorte et al. (2008) applied multidimensional scaling (MDS) and cluster analysis to inferotemporal (IT) cortex responses in human and monkey, and presented their results as "reveal[ing] the properties that dominate the representation of our stimuli in the population code without any prior hypotheses". They further used this data to argue that animacy is a dominant categorical feature in the representational space of IT. Similarly, Connolly et al. (2012) employed cluster analysis to infer the presence of categorical structure within the representation of different animate object classes. Sha et al. (2015), again using similar methods, argue against animacy as a categorical dimension in the representational space of ventral visual cortex.

In this search for the 'true' dimensions of objects representations in ventral visual cortex, dimensionality reduction is treated as giving more direct access to the underlying representational structure than can be gained using hypothesis-driven methods. For example, Caspari et al. (2014) applied a cluster analysis to data from occipito-temporal cortex in order "to view the structure of the [data] … without a bias for a-priori defined stimulus classes". Vul et al. (2012) applied a cluster analysis and found clusters for face, place and body responses in ventral visual cortex (an organisation hypothesized previously), and concluded that their "discovery suggests that the observed dominance of these response profiles in the ventral visual pathways has not been due to the biases present in the way the hypothesis space has been sampled in the past but to inherent properties of the ventral visual pathway" (see also Lashkari et al., 2010). In these ways, the consequence of treating dimensionality reduction as 'data-driven' and 'hypothesis-neutral' is that the results can be conferred a special status as being untainted by the experimenter's preconceptions.

Dimensionality reduction has not only being applied in this way in the field of visual object perception, it is also being applied to other fields of research where the representational space of the brain is largely unknown. These include understanding the representational structure of face perception (e.g. Nestor et al., 2016), of prefrontal cortex during working memory (Machens et al., 2010), of sensorimotor cortex during speech production (Bouchard et al., 2013), and of conceptual semantic representations (Zinszer et al., 2016; Huth et al., 2016a). A common theme motivating such work is the hope that by using 'data-driven' methods we might discover previously unconsidered features of the

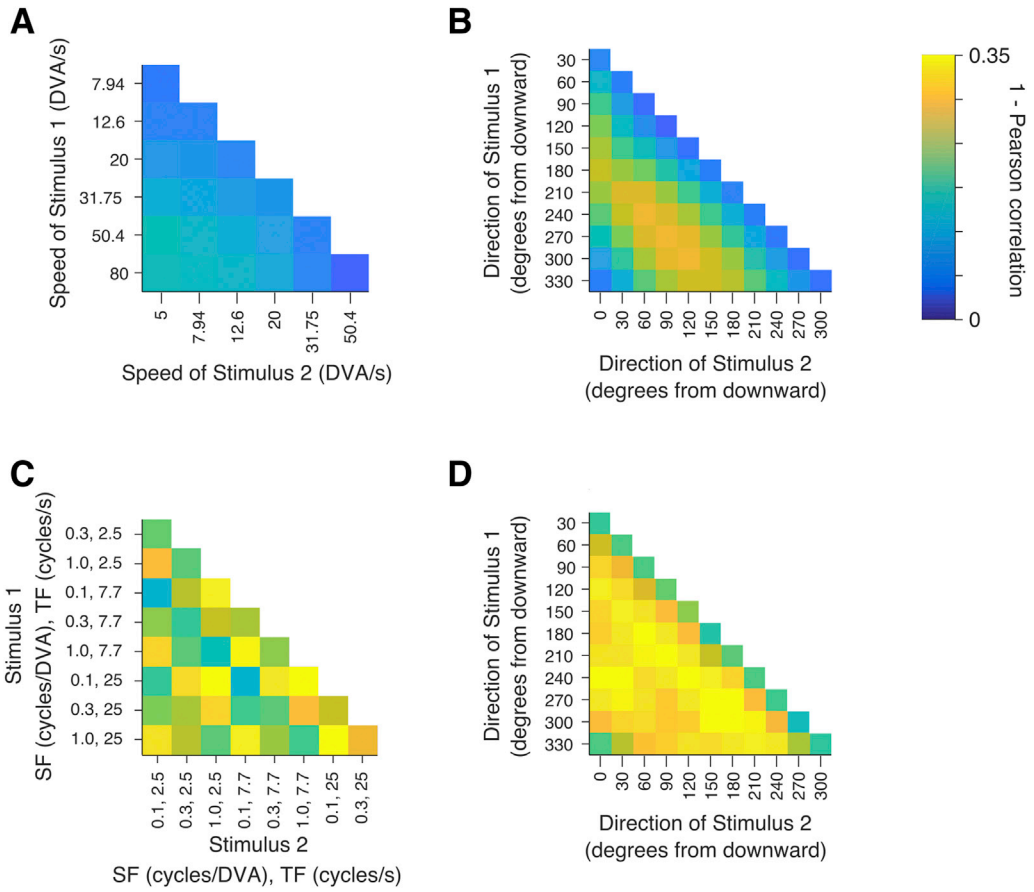


Fig. 1. Summary of spike rate correlation for moving dot fields (**A–B**) and moving gratings (**C–D**). In each case, spike-rate dissimilarity values (1 – r) were calculated for a single pair of stimuli, and then averaged according to the stimulus features labelled above. Spike-rate dissimilarity values were calculated within each data set, by correlating the pattern of spike rates across electrodes during the stimulus-induced response (66–564 ms after stimulus onset). In **A** and **B** the average spike-rate dissimilarity values for moving dot field stimuli are shown as a function of dot field speed (**A**) and direction (**B**). In **C** and **D** the average spike-rate dissimilarity values for moving grating stimuli are shown averaged across spatial and temporal frequency (**C**) and direction (**D**).

brain's representational space, and that we can arrive at these findings in a more timely manner than if we rely on a series of hypothesis-driven experiments that test predefined dimensions. Kanwisher (2010) summarises this viewpoint by noting "if we proceed by testing only the categories that seem plausible to us, then we risk getting trapped within the confines of our own preconceptions." Her suggested solution is to use dimensionality reduction and other approaches which "circumvent these biases by searching for structure in the functional responses of the ventral visual cortex in a hypothesis-neutral fashion".

As dimensionality reduction is gaining traction as a method for analyzing higher-order representational spaces, we believe it is timely and important to consider the strengths and limitations of this approach. In this paper we seek to improve the usefulness of dimensionality reduction by sharpening the conceptual definition of 'data-driven' versus 'exploratory' as applied to this context. We use an empirical example to illustrate a number of practical challenges for interpreting the output of dimensionality reduction. Finally, we outline a framework for how best to employ dimensionality reduction for understanding neural representational spaces.

First, we outline some conceptual considerations. Since these methods are unsupervised, the results of dimensionality reduction analyses are often interpreted as being a measure of the neural representational space that is 'hypothesis-neutral' (Kanwisher, 2010; Kriegeskorte et al., 2008) and 'bias-free' (Caspari et al., 2014). However, we argue here that such an approach is hypothesis-neutral and bias-free only if both (1) the methods are unsupervised and (2) the stimulus set adequately samples the relevant feature space.

Furthermore, even operating under the assumption that the stimulus set is unbiased, there are issues concerning the interpretation of data from dimensionality reduction. Regardless of the input, dimensionality reduction methods provide a solution – whether it is sensible or not. The interpretation of the extracted dimensions is not necessarily straightforward (Adolphs et al., 2016), and it requires the experimenter to recognize sensible structure in the output, which introduces further possibilities for bias. The choice of method also has embedded issues and assumptions that are often not given full consideration. For example, many dimensionality reduction methods (including PCA and MDS) suffer from rotational indeterminacy, i.e. the solutions obtained can be arbitrarily rotated. The criteria used for selecting a solution could also affect a researcher's capacity to "discover" structure in the data. The choice of method also imposes assumptions of knowledge about the structure of information in the representation. Cluster analyses, for example, assume categorical structure in the representation, while other methods assume a continuous feature representation.

The theoretical issues above weaken the claim that dimensionality reduction is an "unbiased"/'hypothesis-neutral' approach for revealing representational spaces. There is also a practical limitation for evaluating their efficacy: as previously noted, these methods have mainly been applied in cases where the underlying structure of the brain's representational space is unknown, meaning that it is impossible to evaluate how successful these methods have been at extracting the representational space of neural responses. Here we sought to fill that gap. We reasoned that if dimensionality reduction is useful for revealing the structure of neural representational spaces for complex, multidimensional stimulus spaces, they should also be able to extract known feature dimensions in a simpler case, where the stimuli systematically sampled a small number of

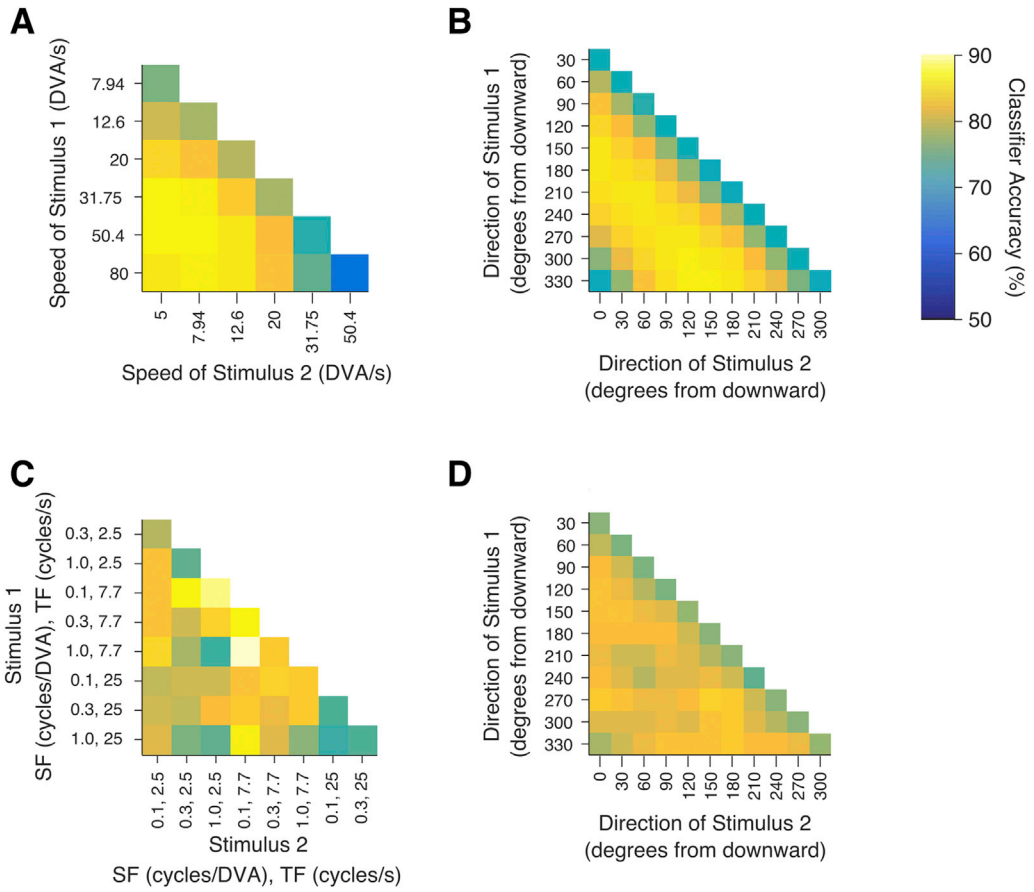


Fig. 2. Summary of classifier performance for moving dot fields (**A–B**) and moving gratings (**C–D**). In each case, classifiers were trained to discriminate a single pair of stimuli, so chance performance is always 50% correct (darkest blue). Classifiers were trained on multiunit spike rates from a single animal within short (2 ms) time bins, and classifier performance is averaged across the duration of the stimulus-induced response (66–564 ms after stimulus onset) and across data sets. In **A** and **B** average discriminability of moving dot field stimuli is shown as a function of dot field speed (**A**) and direction (**B**). In **C** and **D** average discriminability of moving grating stimuli are shown averaged across spatial and temporal frequency (**C**) and direction (**D**).

feature dimensions.

Here we evaluated the effectiveness of dimensionality reduction methods for 'discovering' the dimensions of a representational space where we had clear predictions for the expected dimensions. We applied a range of dimensionality reduction methods to analyze multi-electrode recordings from middle-temporal area (MT) in anesthetized marmosets who were presented with a range of simple moving stimuli that varied systematically in motion direction, speed, spatial frequency and temporal frequency. Area MT has been extensively studied, and contains a high proportion of cells that are selective for motion direction and speed (Maunsell and van Essen, 1983; Albright, 1984; Movshon et al., 1985), the activity of which correlates with perception of motion (Newsome et al., 1989; Salzman et al., 1990; Britten et al., 1996). We use the results of these analyses to illustrate potential challenges for interpreting the results of dimensionality reduction, and integrate these considerations with a conceptual framework for using dimensionality reduction as an exploratory and/or data-driven method for understanding neural representational spaces.

2. Materials and methods

We analyzed spiking activity in multielectrode recordings from area MT of 6 sufentanil-anesthetized marmoset monkeys, collected using protocols that have been described previously (McDonald et al., 2014; Solomon et al., 2015; Goddard et al., 2017). The same animals and data sets used here have also been analyzed in previous work (Goddard et al., 2017), without the use of dimensionality reduction methods. All data sets (raw data spike counts and classifier performance) used here are freely available for download from a Dryad database (http://dx.doi.org/10.5061/dryad.6f8f0).

2.1. Experimental preparation

We obtained six adult marmosets (*Callithrix jacchus*; 5 males; weight 290–400 g) from the Australian National Health and Medical Research Council (NHMRC) combined breeding facility. Procedures took place at the University of Sydney and were approved by Institutional (University of Sydney) Animal Ethics Committee and conform to the Society for Neuroscience and NHMRC policies on the use of animals in neuroscience research.

2.2. Electrophysiological recordings

In each animal, a craniotomy was made over area MT, a large durotomy was made and extracellular recordings were obtained using a 10 × 10 grid of parylene-coated platinum iridium microelectrodes (1.5 mm in length, spacing 0.4 mm; Blackrock Microsystems), pneumatically inserted to a depth of approximately 1 mm (Rousche and Normann, 1992). Signals were band-pass filtered (0.3–6 kHz), and sampled by a Tucker Davis Technologies RZ2 at 24 kHz. For all implants, we identified electrodes that were likely to be located within area MT or MTc based on the directional-sensitivity of the multi-unit recordings, and using the trajectory of receptive field positions (Rosa and Elston, 1998), as described in detail by Solomon et al. (2015). Across animals, 59–96 of the possible 96 electrodes were located within area MT or MTc, and were included in the analyses below.

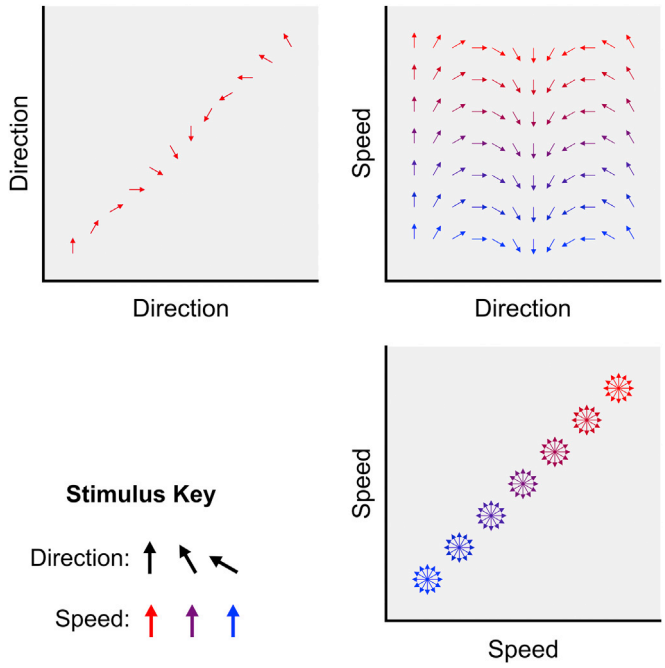


Fig. 3. Illustration of the expected/hypothetical dimensions for multi-unit responses to moving dot fields (84 unique stimuli, of 12 directions and 7 speeds). The hypothetical direction and speed dimensions are plotted individually (top left and bottom right) and against one another (top right). The location of each moving dot field stimulus in these spaces is given by the origin of an arrow, with the dot field direction and speed given by the direction and color of the arrow respectively. In each plot all 84 unique stimuli are plotted, although in the top left plot the arrows of same direction but different speed are overlapping.

2.3. Visual stimuli

Visual stimuli were drawn at 8-bit resolution using commands to OpenGL, by custom software (EXPO; P. Lennie) running on a G5 Power Macintosh computer. Stimuli were displayed on a calibrated cathode ray tube monitor (Sony G520, refresh rate 100 Hz, mean luminance 45–55 cd/m^2, width 40 cm and height 30 cm). The monitor was viewed at a distance of 45 cm. During measurements, one eye, usually the ipsilateral eye, was occluded.

In one set of stimuli, white circular dots (Weber contrast 1.0; diameter 0.4 d.v.a (degrees visual angle)) moved across a quasi-circular area (diameter 48 d.v.a.; cropped at 37 d.v.a. vertically); outside each dot, the monitor was held at the mean luminance. Dots were presented at a density of 0.3 dots/s/d.v.a. and moved with 100% coherence and infinite lifetime. The position of each dot at the beginning of a trial was specified by a random number generator; the same set of positions was used on every trial. Each stimulus was presented for 500 ms; the screen was held for 300 ms at the mean luminance between each trial There were 20 repeats of each stimulus type [7 speeds (5, 7.9, 12.6, 20, 31.8, 50.4 & 80 d.v.a. s^{-1}) $\times$ 12 directions (30° steps)], giving a total of 1680 trials. A total of 6 data sets were collected from 5 animals for the moving dot field stimuli. The animals (using naming conventions from the previous published work) were ma025 (contralateral and ipsilateral eyes), ma026 (contralateral), ma027 (contralateral), my145 (contralateral) and my147 (contralateral).

In the second stimulus set, a large sine-wave grating (Michelson contrast 0.5) drifted within a circular window (diameter 30 d.v.a.) with hard edges; outside the window, the monitor was held at the mean luminance. The spatial frequency was either 0.1, 0.32 or 1 cycles/d.v.a., and temporal frequency was 2.5, 7.69 or 25 Hz. Each stimulus was presented for 500 ms; the screen was held for 50 ms at the mean luminance between each trial. There were 20 repeats of each stimulus type [9 spatiotemporal frequencies x 12 directions (30° steps)], giving a total of 2160 trials. A total of 6 data sets were collected from 4 animals: ma025 (contralateral and ipsilateral eyes), ma026 (contralateral and ipsilateral eyes), ma027 (contralateral) and my147 (contralateral).

Each stimulus set included interleaved 'blank' trials (1/13 of the total number of trials) on which no stimulus was displayed, and the screen remained at the mean luminance. Each set of stimuli, including these blanks, were presented in pseudo-random order.

2.4. Preliminary data analysis

For each of the electrodes identified as being within area MT, we used the Matlab function *findpeaks* to identify candidate waveforms with peak amplitude that exceeded 3 standard deviations of the raw signal on that channel. We did not sort spike waveforms into separate neuronal sources, so spike rates were expressed as the number of spikes per electrode.

2.5. Spike rate correlation analysis

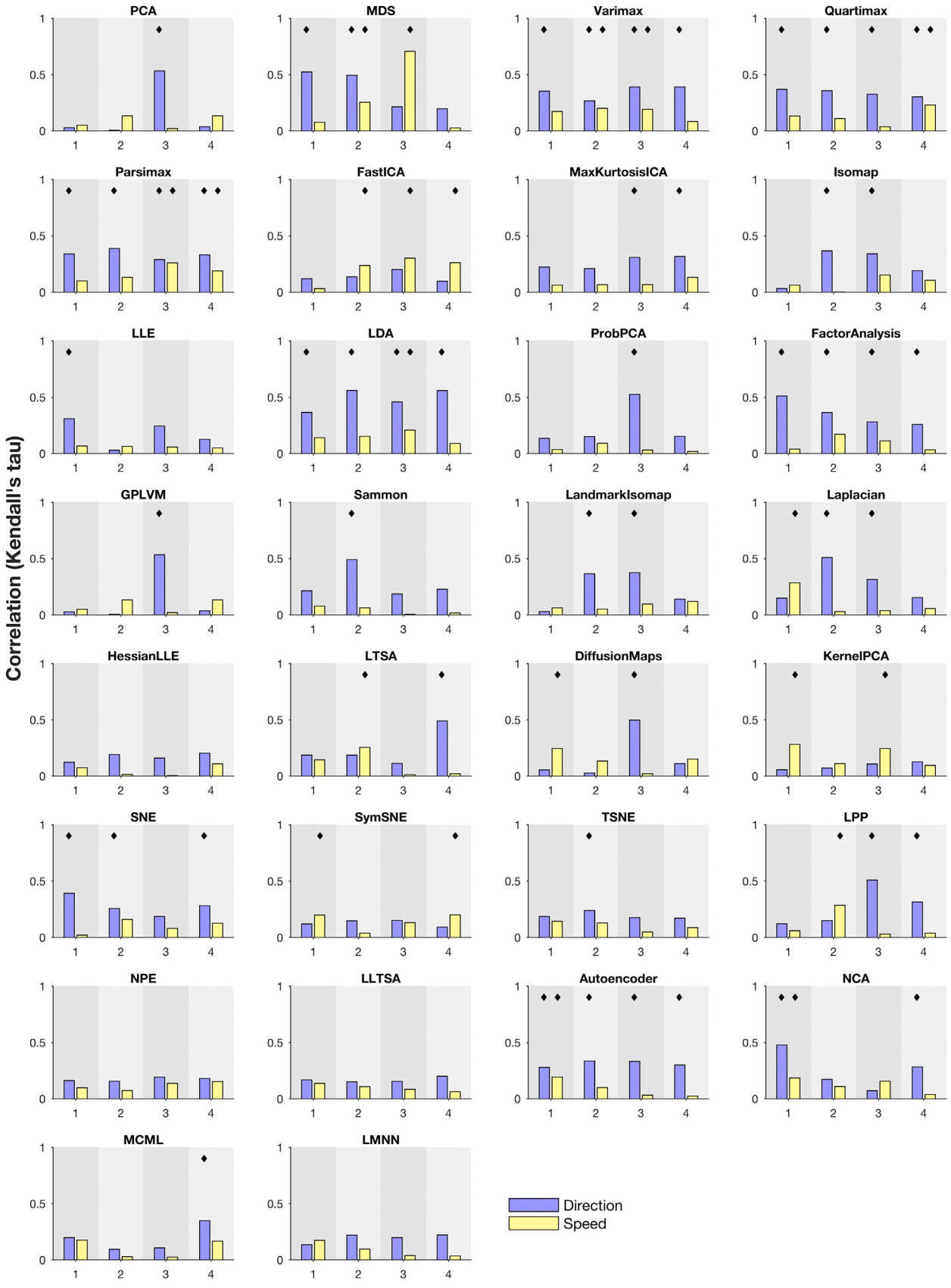
We considered the neural population response in three cases: moving dot fields (6 data sets, $n = 84$ unique stimuli), moving grating stimuli (6 data sets, $n = 108$ unique stimuli) and both dot and grating stimuli (5 data sets from 4 animals, $n = 192$ unique stimuli). For each of dataset, we defined an *nxn* 'dissimilarity matrix' based on 'spike-rate dissimilarity' (1 - Pearson correlation) between the spike-rates across electrodes (total of 59–96 per dataset) to a given pair of stimuli. We correlated the mean pattern of spike-rates across electrodes to different stimuli. Within each data set, we averaged spike-rates across stimulus trials of the same type, and then averaged the spike-rate across the sustained period of the stimulus induced response (66–564 ms), generating a single pattern of spike-rates across electrodes for each stimulus. In previous analyses (Goddard et al., 2017) we found that the population response to stimulus features was relatively stable over this time, after dynamics around the response onset.

Each cell of the dissimilarity matrix was the spike-rate dissimilarity (1 - r) between the response to stimulus A and stimulus B, where stimulus A varied from 1 to n with column and stimulus B varied from 1 to n with row. The dissimilarity matrices were by definition symmetric about the diagonal, with a diagonal of zeros (i.e. 1 - r when r = 1). After measuring the spike-rate dissimilarity separately for each dataset, we averaged these values to obtain a group average dissimilarity matrix. These dissimilarity matrices are akin to those used in previous studies (for example, Kriegeskorte et al., 2008).

2.6. Classification analysis

As an alternative to the correlation-based dissimilarity matrices, we also constructed dissimilarity matrices where the dissimilarity of the neural responses was defined by classifier accuracy, which we have used previously (Goddard et al., 2017). In our previous work we first down-sampled the multi-unit activity of each channel to 500 Hz, extracting the number of spike waveforms on that channel in each 2 ms time bin, and repeated the classification process (described below) at each time point (every 2 ms) in the 600 ms window in order to measure how classification accuracy evolved over time. Here we retained the fine temporal resolution of the original analysis in order to avoid ceiling effects in classifier performance, but averaged classifier performance across the same portion of the stimulus induced response as used in the spike-rate dissimilarity analysis (66–564 ms).

To perform the classification analysis we first reduced the dataset by applying principal component analysis to the entire dataset for each animal, comprising the entire 600 ms of neuronal response following stimulus onset, for each of 2160 (gratings) or 1680 (dots) trials and up to 96 channels. Data from the first n components that accounted for 99% of the variance were retained; data from remaining components were discarded. Across animals and stimulus type (gratings and dots) n ranged from 54 to 93. Note that the application of PCA to the raw data here was


PCA
MDS
Varimax
Quartimax
Parsimax
FastICA
MaxKurtosisICA
Isomap
LLE
LDA
ProbPCA
FactorAnalysis
GPLVM
Sammon
LandmarkIsomap
Laplacian
HessianLLE
LTSA
DiffusionMaps
KernelPCA
SNE
SymSNE
TSNE
LPP
NPE
LLTSA
Autoencoder
NCA
MCML
LMNN
Correlation (Kendall's tau)
Direction
Speed
Data-driven dimension

for data reduction, and is unrelated to the dimensionality reduction described below.

We trained and tested classifiers (linear discriminant analysis, LDA) to discriminate population responses on trials of different stimulus conditions. We also repeated the entire analysis using a linear support vector machine (SVM) classifier and obtained very similar results (data not shown). For each possible pair of the 84 unique dot field stimuli we trained the classifier to discriminate between two stimulus conditions then measured classifier accuracy using 10-fold cross-validation. The classification rule was learned using 90% of trials (18 trials of each type), and the accuracy of this rule was tested on the remaining 10% of trials (2 trials of each type). This process was repeated for each of 10 partitions of the data, such that all data were included in the test set once, and no data were ever used in both the training and test set (leave-one-out train-and-test).

Similarly, we trained classifiers to discriminate each pairing of the 108 unique grating stimuli, and again tested the classifier accuracy using 10-fold cross-validation. Finally, for those animals where both the moving grating stimuli and moving dot stimuli were presented to the same eye (ipsilateral or contralateral to the recorded MT), we repeated the PCA and classification analysis for a single data set of 3840 trials (with 192 unique stimuli).

In every case the entire classification analysis was performed separately for each animal, and the average classification accuracy was obtained by averaging classifier performance (in units of d') across animals.

2.7. Dimensionality reduction

From the spike-rate dissimilarity analysis and the classification analysis we obtained two alternative dissimilarity matrices for each of the 3 cases: moving dot fields (average of 6 data sets, $n = 84$ unique stimuli), moving grating stimuli (average of 6 data sets, $n = 108$ unique stimuli) and both dot and grating stimuli (average of 5 data sets, $n = 192$ unique stimuli). We applied a range of dimensionality reduction techniques to each of these six dissimilarity matrices.

For each dissimilarity matrix, we treated the matrix as a space with n dimensions, each with n observations, and applied the 30 dimensionality reduction methods listed in Table 1 to reduce the data from an n dimensional space to a m dimensional space, where m varied from $m = 1$ to $m = n$-1. To implement multidimensional scaling (MDS) of our dissimilarity matrices we used the Matlab function *mdscale*. For independent component analysis (ICA) we used the 'FastICA' package (downloaded from MathWorks, version: 14 January, 2017, see http://research.ics.aalto.fi/ica/fastica/), from which we used the fast, fixed-point algorithm for Independent Component Analysis, as well as the kurtosis maximization ICA. ICA is conceptually similar to PCA in that it seeks orthogonal dimensions that explain variance while reducing the dimensionality of the dataset. The difference is that ICA iteratively maximizes the absolute value of kurtosis, rather than explained variance. In principle, this should make it better able than PCA to identify meaningful dimensions in a dataset, particularly if the values on those dimensions are not normally distributed. We applied the Matlab function *rotatefactors* to the results of our Principal Component Analysis (PCA) to obtain the 'Varimax', 'Quartimax' and 'Parsimax' rotations. Each of these factor analytic rotations aim to discover any latent sources of variance whose signals are likely to be mixed in the components extracted by PCA. For example, 'Varimax' maximizes the variance of the squared loadings of each factor on all the dimensions, which aims for a solution where each factor has either a large or a small loading on each data dimension, ideally yielding optimal separability and interpretability. The remaining 24 dimensionality reduction methods we implemented using 'The Matlab Toolbox for Dimensionality Reduction' (version 0.8.1b, March 21, 2013, see van der Maaten and Hinton, 2008).

To evaluate how well the extracted dimensions correlate with the expected dimensions of each representational space, we generated hypothetical dimensions for each case and correlated these hypothetical dimensions with the observed dimensions using Kendall's *tau* (a rank correlation). In each case, we generated hypothetical dimensions for each of the stimulus dimensions that varied in the stimulus set. We assumed that for an extracted dimension to correspond to this feature dimension there should be ordering of the stimuli along the extracted dimension so that stimuli of greater feature difference are further apart than those of smaller feature difference. For the moving dot field stimuli, we generated 'direction' and 'speed' dimensions, for moving grating stimuli we generated 'direction', 'spatial frequency (SF)', 'temporal frequency (TF)' and 'speed (that is, TF/SF)' dimensions. For the case with both moving dot field and moving gratings we included 'direction', 'SF', 'TF', 'speed' and 'category', where the 'category' dimension was a binary classification of the stimuli into dot fields and gratings. Examples of these hypothetical dimensions are plotted in Figs. 2 and 6. For each of these dimensions, we were interested in whether the extracted dimensions would show the same ordering of stimuli according to the feature values. Since there is no a priori reason for the order to be ascending or descending (for example, for speed to increase or decrease along an extracted 'speed' dimension), we specified a pair of hypothetical dimensions for each of these stimulus features, where one hypothetical dimension was a mirror reversal of the other.

Similarly, since direction is a circular dimension, and there were 12 directions in our stimulus set, we generated 12 alternatives for the hypothetical direction dimensions, where each started with a different direction. We then mirror reversed each of these possibilities, to give a total of 24 hypothetical direction dimensions.

Finally, when we included both moving dot field and moving grating stimuli in a single data set, we needed to predict how the dot fields, which are broadband in SF and TF, should be ordered relative to the gratings of a single SF and TF. To avoid assuming a single correct solution, we created 4 alternatives for the hypothetical SF and TF dimensions, where the dot field stimuli were either the first or last along the dimension, intermediate to the first and second SF/TF, or intermediate to the second and third SF/TF. We mirror reversed each of these 4 possibilities to create 8 hypothetical SF and TF dimensions in the case where dot fields and gratings were considered as a single data set.

To measure how well the extracted dimensions corresponded to these hypothetical dimensions, we rank correlated the extracted dimensions with each of the 2, 8 or 24 alternatives, and used the maximum correlation value across the alternatives as the measure of the extracted dimension's correlation with that feature dimension. When assessing the significance of the correlations, we used a Bonferroni correction to adjust the p values of these tests for the number of alternatives tested in each case. For the moving dot field data we calculated these correlation values for the first 4 dimensions from each method, and for the moving grating and combined cases we considered the first 6 dimensions. By using a rank correlation between the hypothetical dimensions and the observed dimensions we were testing simply for an ordering of the stimuli according to their feature values, rather than testing how the feature values were spaced along the dimension. We chose to use rank correlations to avoid making assumptions about whether (for example) a neural representation of stimulus speed should be mapped onto physical speed in a linear, logarithmic, or other monotonic relationship.

For the combined moving dot field and grating data we also executed

Fig. 4. Correlation between data-defined dimensions and the known stimulus dimensions of dot field direction and speed. Here we correlated the known stimulus dimensions with the dimensions extracted when dimensionality reduction was applied to the dissimilarity matrix based on spike-rate dissimilarity values. For each dimensionality reduction method tested, we considered only the top 4 dimensions that were extracted, and correlated these with each of the hypothetical direction and speed dimensions. Filled diamonds show correlation values that were significantly ($p < 0.05$) above zero, with Bonferroni correction for the multiple hypothetical stimulus dimensions that were compared with each data-driven dimension (see Materials and Methods for details).

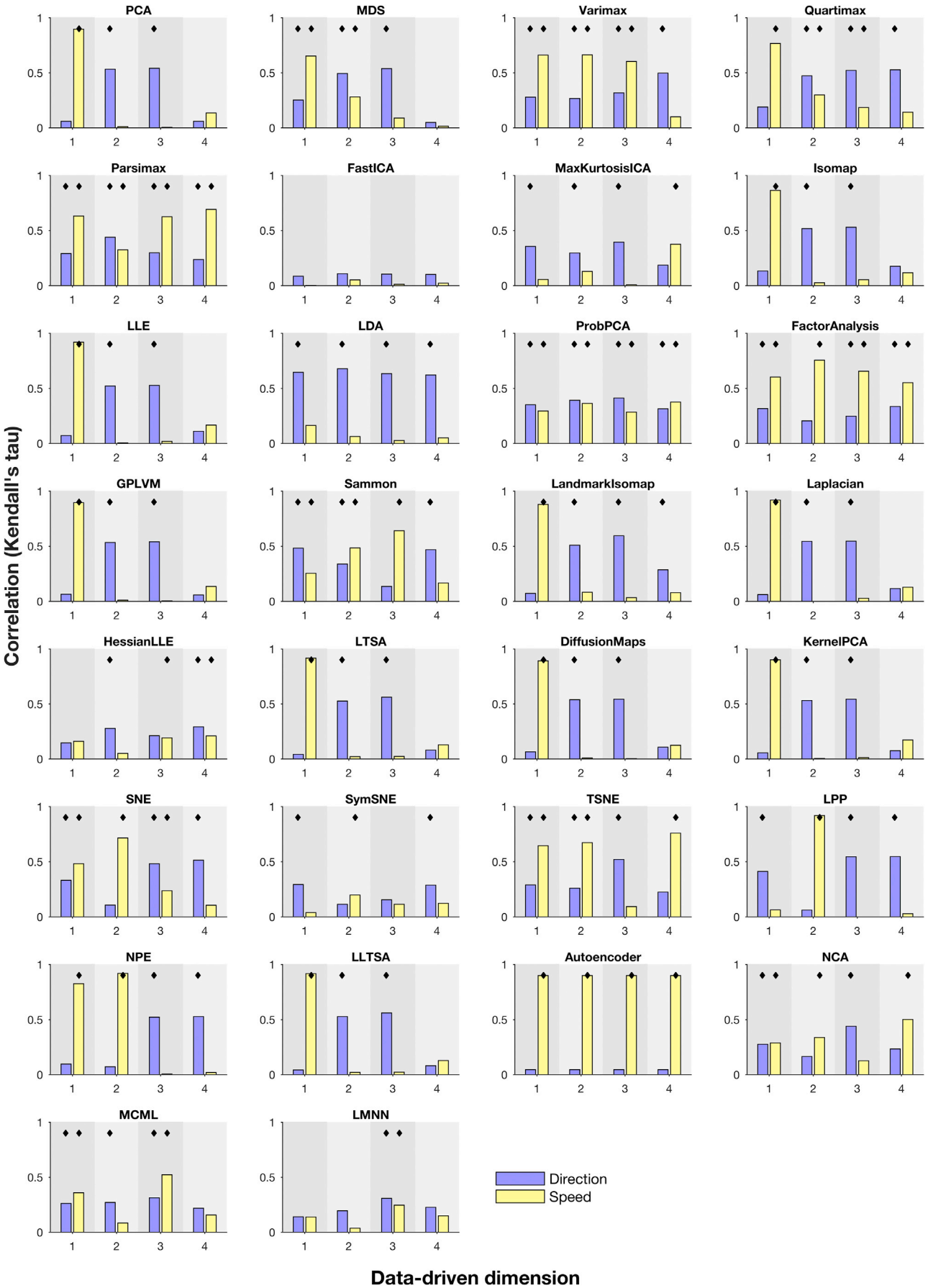


Fig. 5. Correlation between known stimulus dimensions and the data-defined dimensions extracted from the dissimilarity matrix based on classifier performance. Plotting conventions as in Fig. 4.

a clustering algorithm (the Matlab function *linkage* with the default nearest distance method) to generate a hierarchical cluster tree of the dissimilarity matrix.

3. Results

To evaluate the effectiveness of dimensionality reduction for recovering the dimensions of neural feature representations, we applied these methods to a brain region where the feature dimensions should be well defined: area MT responses to simple moving stimuli. We analyzed multi-unit activity in multielectrode recordings from area MT of 6 sufentanil-anesthetized marmoset monkeys. Each animal was shown stimuli from one or both of two stimulus sets: a set a moving dot fields of varying direction and speed, and a set of moving gratings of varying direction, spatial and temporal frequency. For every pair of unique stimuli, we used a correlation-based measure of the dissimilarity in the pattern of spike-rates across electrodes (spike-rate dissimilarity = 1-r, see Fig. 1) and we measured the discriminability of the population responses within each animal using multivariate pattern classification analysis (see Fig. 2). To avoid ceiling effects in classifier performance, we performed the analysis on spike counts from 2 ms bins, then averaged classifier performance across the stimulus-induced response (66–564 ms after stimulus onset). Even with short time bins, classification performance was high (see Fig. 2).

3.1. Response dissimilarity increased with increasing stimulus feature difference

As expected, both measures of response dissimilarity tended to increase with increasing stimulus feature difference. Spike-rate dissimilarity tended to increase when the spike-rates were in response to stimuli that were more different along one or more dimensions. Similarly, classifier accuracy tended to increase when the stimuli it was trained to discriminate were more different along one or more feature dimensions. For moving dot fields, spike-rate dissimilarity and classifier performance were lowest when the stimuli were separated by only a single step in direction and/or speed (the blue/green diagonals in the matrices in Figs. 1 and 2). For speed (Figs. 1A and 2A), spike-rate dissimilarity and classifier performance generally increased as the speed difference increased, and were lowest for stimuli of highest speed. For dot field direction (Figs. 1B and 2B), spike-rate dissimilarity and classifier accuracy were greatest when the dot fields were 180° apart (moving in opposite directions). Note that since direction is a circular variable, when the stimuli were 30° and 330° from downward they are only 60° apart. This pattern of results is consistent with the existing literature on area MT, namely that it encodes both the speed and direction of moving patterns.

For moving grating stimuli, classifier accuracy tended to increase when the spatial and/or temporal frequency difference increased (Fig. 2C), and was lowest for stimuli of high spatial frequency and/or high temporal frequency. This pattern was less clear in the correlation results (Fig. 1C). Compared with the moving dot fields, the average spike-rate dissimilarity and classifier accuracy for moving gratings was more uniform across direction differences, although it was still lowest for stimuli of the same or smallest (30°) direction difference (Figs. 1D and 2D).

In previous work (Goddard et al., 2017) we used the classification accuracy data to demonstrate the dependence of the population response on direction, speed, spatial and temporal frequency, and how the encoding of these features evolves over time. Our analyses confirmed what can also be seen by inspecting Fig. 2, that there is information about each of these stimulus features in the population response, consistent with a population response that varies systematically with each of these stimulus feature dimensions. We next asked whether standard dimensionality reduction methods could independently recover the feature dimensions that were systematically varied in the stimulus set, and which appear to be encoded systematically in the population response.

3.2. Dimensionality reduction applied to neural responses to moving dot stimuli

To evaluate the effectiveness of dimensionality reduction for 'discovering' the dimensions of a neural representational space we first considered the classifier responses to moving dot fields. This is the simplest case in our data, where the stimuli varied along only 2 dimensions (direction and speed), and there was a robust increase in classifier performance as the stimuli differed along either of these dimensions.

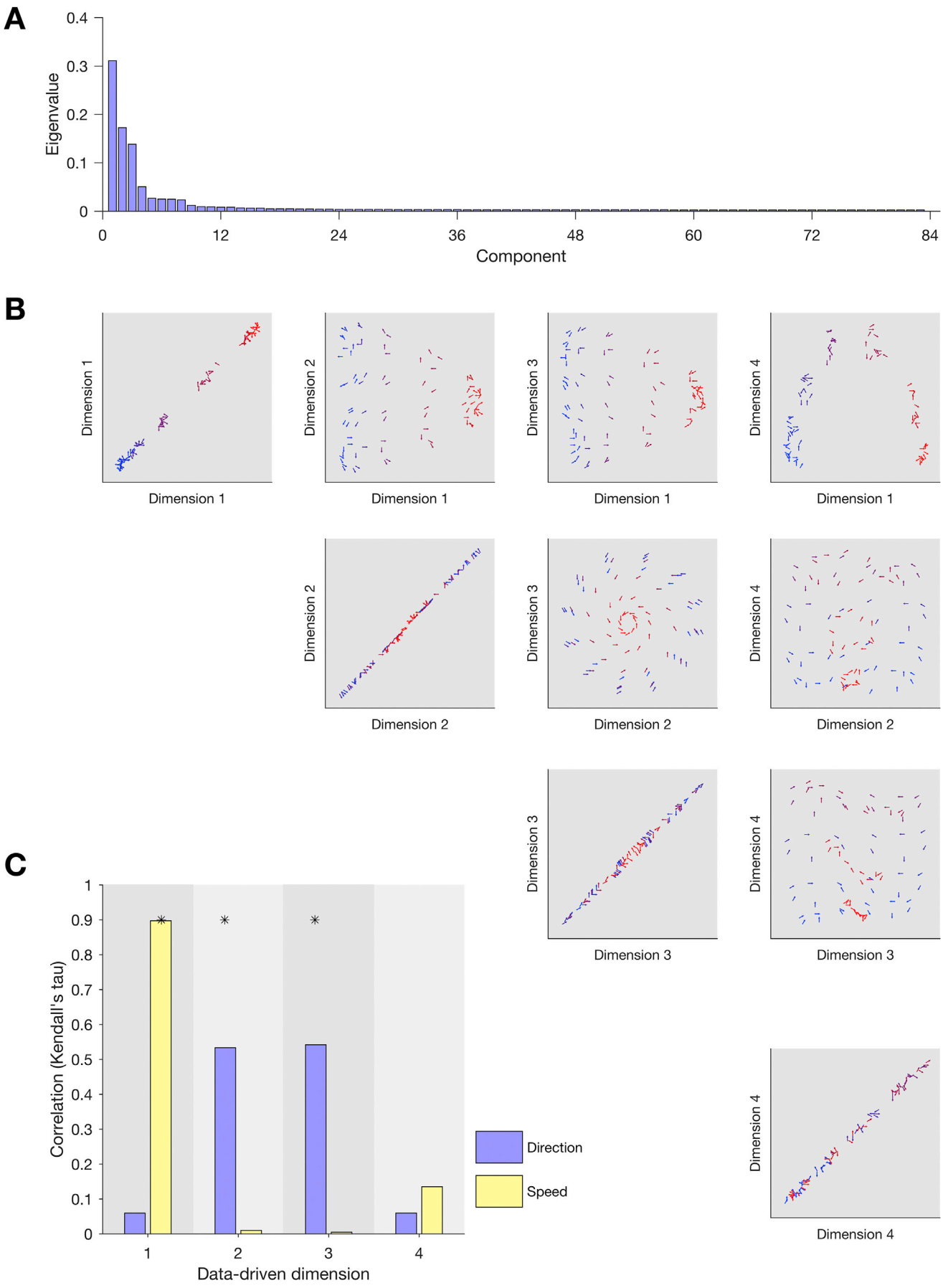
First we generated the simplest possible hypothetical dimensions that would lead to the interpretation that the representational space of MT neurons is defined by speed and direction. These hypothetical dimensions are plotted in Fig. 3. We arrived at these dimensions by assuming that direction and speed should be extracted as orthogonal dimensions, and that there should be ordering within along both these dimensions so that stimuli of greater feature difference are further apart along the relevant dimension. Since direction is a circular dimension, the solution in Fig. 3 is one of 12 equally correct solutions, in which the leftmost direction is a different direction in each case. Similarly, for both the speed dimension and each of the 12 correct direction dimensions, a left-right flipping of the order along the dimension is equivalently correct. This gave us 2 speed dimensions and 24 direction dimensions that we treated as correct solutions.

We tested a large range of dimensionality reduction approaches (see Figs. 4 and 5, details in Table 1), including the commonly used Principal Component Analysis (PCA) and Multi-Dimensional Scaling (MDS). For each of these methods we rank correlated each of the hypothetical direction and speed dimensions with each of the first 4 dimensions from the dimensionality reduction. The maximum correlation between each of these data-defined dimensions and any one of the direction and speed dimensions is plotted in Fig. 4 (for dimensions extracted from spike-rate dissimilarity data) and Fig. 5 (for dimensions extracted from classifier accuracy data). Since we used a rank correlation measure (Kendall's tau) it is possible for the 'data-defined' dimensions to reach a correlation of 1 with a stimulus-defined dimension by ordering the stimuli according to either their direction or speed.

There was considerable variation across dimensionality reduction methods in how well they extracted the direction and speed as relevant feature dimensions (Figs. 4 and 5). Some methods found dimensions that correlated well for direction but did not isolate a speed dimension, or vice versa. The maximum correlation with the speed dimension achieved by any method was $tau = 0.92$ (LLE method, 1st dimension and NPE method, 2nd dimension, for classifier accuracy data), and for direction this was lower at $tau = 0.68$ (LDA method, 2nd dimension, for classifier accuracy data).

Overall, the dimensions extracted from the classifier accuracy data (Fig. 5) tended to correlate with the stimulus dimensions to a greater extent than those extracted from the spike-rate dissimilarity data (Fig. 4). This was also the case for the moving grating stimuli and combined dot and grating stimuli datasets considered below. For this reason, for the remainder of this paper we focus exclusively on the dimensions extracted from the classifier accuracy data, although we include corresponding results for the spike-rate dissimilarity data in our supplementary material, part 1.

PCA and MDS, the two methods that have been most widely used in previous work on object representations in IT, were both among the best performing methods: for both PCA and MDS, the first 4 dimensions extracted included dimensions that correlated reasonably well with direction and speed. We consider these feature dimensions extracted by PCA and MDS in greater detail in Figs. 6 and 8 respectively. We also consider the results of the PCA combined with Varimax rotation in greater detail (Fig. 7), since unlike PCA and MDS, the PCA with Varimax rotation is specifically designed to isolate separate factors contributing to


A
Eigenvalue
0.4
0.3
0.2
0.1
0
0
12
24
36
48
60
72
84
Component
B
Dimension 1
Dimension 2
Dimension 3
Dimension 4
C
Correlation (Kendall's tau)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
Data-driven dimension
Direction
Speed

variance in the data. For the remaining dimensionality reduction methods, we include similar plots of the extracted dimensions in our supplementary material (parts 2 and 5, for data based on spike-rate dissimilarity and classifier accuracy respectively).

Visual inspection of the plots in Fig. 6B reveals how the first few dimensions from the PCA relate to stimulus speed and direction. The first dimension maps onto stimulus speed, ordering the speeds from low to high (blue to red). The second and third dimensions also have structure related to speed, but this is non-monotonic (speeds tend to go from low to high then back to low), so despite this structure both these dimensions have a low rank correlation with speed which was not significantly above zero. The second and third dimensions order the stimuli according to direction, in a manner that is clearest when these dimensions are plotted against one another. However, despite what may appear to be obvious structure when dimensions 2 and 3 are plotted against one another, the critical issue to consider here is whether it would be possible to 'recognize' that direction was a relevant dimension for this neural representational space if it were not already known. Considered separately, neither the second nor third dimensions appear to be natural dimensions of the MT representational space. The dimensionality reduction has failed to extract a circular dimension as a single dimension, which is not necessarily a failure of the dimensionality reduction, but it does limit the interpretability of the result when used for exploring neural representational spaces. If higher order areas such as IT also have what are best explained as circular dimensions, these could be missed by such methods, or instead extracted as multiple linear dimensions, that are close to meaningless for understanding the true neural representational space.

When the components extracted by the PCA are rotated using Varimax (Fig. 7), there was more, rather than less, conflation of the speed and direction dimensions. After the rotation the first, second, and third dimensions were all significantly correlated with both direction and speed, and no single dimension reaches the same correlation with direction as the second and third dimension from unrotated PCA components.

In the results of the MDS (Fig. 8) the extracted dimensions are further removed from the known stimulus dimensions of direction and speed. As for the results from PCA, there is clearly structure in the result that is related to the speed and direction dimensions. But in this case there is less separation of the speed and direction dimensions, so that each of the first three dimensions include systematic variation with speed and direction. For a naive observer, there would be even less chance of recognizing speed and direction as relevant dimensions in this representational space.

3.3. Dimensionality reduction applied to neural responses to moving grating stimuli

Next we considered the results of dimensionality reduction applied to neural responses to the moving grating stimuli. This stimulus set is slightly more complex, since the stimuli varied along three dimensions rather than two: grating motion direction, spatial frequency (SF) and temporal frequency (TF). Since speed is a function of SF and TF (speed = TF/SF) we compared each extracted dimension with hypothetical dimensions based on either direction, SF, TF or speed, as shown in Fig. 9. We plot the correlations between these hypothetical dimensions and the top six dimensions that were extracted from the classifier accuracy data by each dimensionality reduction method (Fig. 10).

As for the responses to moving dot fields, for the responses to moving gratings there was considerable variation in the extracted dimensions across dimensionality reduction methods. The maximum correlation with the direction dimension achieved by any method was *tau* = 0.67 (PCA method, 4th dimension), for SF the maximum *tau* = 0.76 (Autoencoder method, 5th dimension), for TF the maximum *tau* = 0.64 (SNE method, 4th dimension), and for speed the maximum *tau* = 0.85 (LTSA and LLTSA methods, 1st dimension).

As before, PCA and MDS were among the best performing methods in terms of extracting dimensions that had a relatively high correlation with each of the stimulus dimensions, and in terms of identifying separate dimensions for orthogonal stimulus dimensions. We plot the results of PCA, PCA with Varimax rotation, and MDS in greater detail in Figs. 11–13 respectively. In feature space, while speed correlates with both SF and TF (and so overlap between speed and SF and speed and TF is expected), the feature dimensions of direction, SF and TF are each orthogonal to one another. From Fig. 10, MDS was one of the best methods for isolating dimensions that selectively correlated with direction, SF and TF. However, the dimension isolating TF was the 5th dimension extracted, which raises the issue of whether a naive observer would be likely to consider that there might be signal in the 5th dimension or whether they would only consider the first 2 or 3 dimensions. This issue is also illustrated by considering the plots in part **B** of Figs. 11–13. Even though there is again clearly some structure in the arrangement of the arrows in the different plots, it is difficult to discern by eye what are the dominant organizing principles along the different dimensions. Without the bar plot in Fig. 13C, it would be hard to notice that the 5th dimension extracted by the MDS is ordering the stimuli by TF to a greater extent than the other dimensions, even though the arrows are color coded by TF. And yet for this method to reveal a previously unknown dimension of the representational space, we are relying on an experimenter recognizing the ordering of stimuli by TF along the 5th dimension without having prior knowledge that TF is a potentially relevant feature.

As for the moving dot stimuli, there is again the issue that the dimensionality reduction methods tend not to extract direction as a single dimension, but across two or more dimensions that correlate with direction. Once again, applying a Varimax rotation to the PCA components resulted in further conflation of stimulus dimensions rather than further separation of these factors. Clearly, although Varimax rotation is designed in principle to isolate theoretically independent dimensions, this was not successful in this dataset.

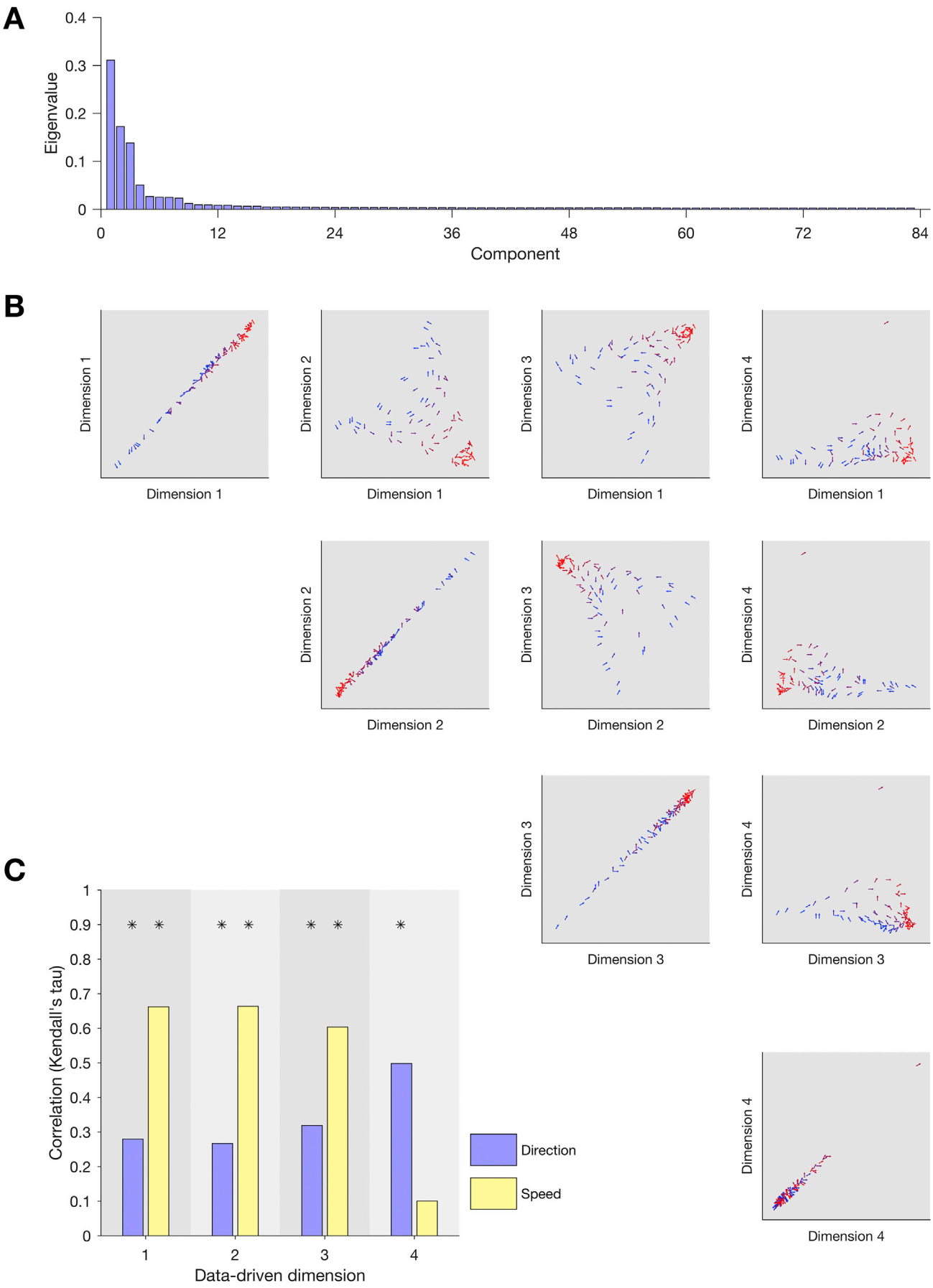
Overall, the plots in part **B** of Figs. 11–13 illustrate how even for what is still a relatively low dimension (3 dimensional) stimulus space, the results of dimensionality reduction are complex and difficult to interpret. This is also illustrated in the corresponding figures for the remaining dimensionality reduction methods, included in our supplementary material (parts 3 and 6, for data based on spike-rate dissimilarity and classifier accuracy respectively).

3.4. Dimensionality reduction applied to neural responses to combined moving dot and grating stimuli

As a final illustrative example, we applied the dimensionality reduction methods to the neural responses for an entire set of stimuli, including moving dot fields and moving gratings. Although the previous results suggest that introducing more complexity is unlikely to make the results clearer, we wanted to include in our analysis a dataset with a categorical stimulus dimension (here, dots versus gratings) since much of the previous work with dimensionality reduction methods has included categorical variables such as animacy in area IT, (although even for the animacy category there is now evidence that animacy might be better conceived as a continuum variable in object representations, see Sha et al., 2015).

As before, we correlated the extracted dimensions with a series of

Fig. 6. Summary of the representational space resulting from dimensionality reduction by principal components analysis (PCA) of the dissimilarity matrix based on classifier performance, for responses to moving dot fields (84 unique stimuli, of 12 directions and 7 speeds). **A**: Eigenvalues of the 84 components, showing that the first few components capture most of the variance in the data. **B**: Data for the 84 unique stimuli projected into spaces defined either by a single dimension or a pair of dimensions. Each moving dot field stimulus is defined by an arrow, where the direction and speed of the stimulus are given by the direction and color of the arrow respectively (blue = slowest speed, red = fastest speed). **C**: The correlation between the 'data-defined' and stimulus-defined dimensions is replotted (from Fig. 5) for PCA.


A
Eigenvalue
0.4
0.3
0.2
0.1
0
0
12
24
36
48
60
72
84
Component
B
Dimension 1
Dimension 2
Dimension 3
Dimension 4
C
Correlation (Kendall's tau)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
Data-driven dimension
Direction
Speed

hypothetical dimensions, this time including the categorical dimension 'DotVsGrat' (dots versus gratings). Since dot fields are broadband for both SF and TF, the categorical variable 'DotVsGrat' is orthogonal with direction, but covaries with SF, TF and Speed. As shown in Fig. 14, there was again considerable variation across the dimensionality reduction methods. The dimensionality reduction methods were not noticeably better at extracting the categorical DotVsGrat dimension, nor did this dimension tend to be extracted within the top few dimensions for any method. The correlation with this hypothetical dimension tended to be lower than the other hypothetical dimensions. The maximum correlation with the DotVsGrat dimension achieved by any method was no higher than for other dimensions, with maximum *tau* = 0.66 (LLTSA method, 3rd dimension, although this dimension correlated more strongly with TF than DotVsGrat), while for direction the maximum *tau* = 0.65 (LDA method, 2nd dimension), for SF the maximum *tau* = 0.72 (LTSA method, 3rd dimension), for TF the maximum *tau* = 0.69 (LLTSA method, 3rd dimension), and for speed the maximum *tau* = 0.67 (Sammon method, 2nd dimension).

As before, we show the results for PCA, PCA with Varimax rotation and MDS in more detail (in Figs. 15–17 respectively). For each of these methods, the second and third dimensions correlated well with direction, and when plotted against one another they show circular structure, as in the results for moving dot fields alone. Corresponding figures for the remaining dimensionality reduction methods are included in our supplementary material (parts 4 and 7, for data based on spike-rate dissimilarity and classifier accuracy respectively).

For the combined responses to moving dot fields and moving gratings, we also show the result of a cluster analysis of classifier performance (Fig. 18). Due to the hierarchical nature of the cluster analysis, the results for dot fields or gratings alone (not shown) are similar to a cluster tree with the branches of the other stimulus set removed. Similar to the results for dimensionality reduction, for this cluster analysis there was evidence that the analysis was extracting structure from the data set, but the end result did not reveal the underlying hierarchical structure in the data in a way that would be easily interpretable for a naive observer. At first glance the cluster analysis appears to have separated the data according to category (dot fields versus gratings) at a fairly high level (Fig. 18A), although closer inspection reveals that while the top and second level branches separated over half the moving grating responses from the dot field responses, the remaining grating stimuli responses are not separated from the dot field responses until the 11th branch. Together, the cluster analysis and the dimensionality reduction results for the entire data set demonstrate that even when the data set include a clear categorical variable this categorical structure will not necessarily be revealed by a data-driven approach.

4. Discussion

Dimensionality reduction approaches such as MDS and PCA are useful illustrative tools for visualizing complex data sets, but there is an emerging trend of treating the results of such methods as revealing new information about the structure of the brain's representational space, particularly when the feature dimensions are unclear (for example, in the representation of natural objects: Kriegeskorte et al., 2008; Sha et al., 2015; Caspari et al., 2014; Cunningham and Yu, 2014).

We argue that enthusiasm for such methods ought to be tempered. Dimensionality reduction can give information *about* representational spaces, but this is often much weaker information than researchers suppose. The constraints on a truly data-driven, hypothesis-free analysis are quite strict, and (we think) rarely met. Even when these constraints are met there are theoretical considerations when interpreting such representational spaces (as outlined above, and by Ritchie et al., In Press; de Wit et al., 2016; Carlson et al., 2018). However, the current results show that there are also practical limitations on the potential for dimensionality reduction to uncover novel representational spaces.

4.1. Feature spaces and representational spaces

As outlined above, feature spaces are mathematical descriptions of a given stimulus set, and a given stimulus set can usually be captured by many equally parsimonious alternatives. Our stimuli give a simple example: movement and speed of a dot in a 2D plane can be represented either in polar or Cartesian coordinates. In the former, the dimensions correspond to speed and direction; in the latter, they correspond the length of an x-vector and a y-vector that can composed to show speed and direction. Both feature spaces are entirely adequate to express the parameters of stimulus speed and distance (as they must be: there is a simple mathematical translation between the two).

Representational spaces, by comparison, encapsulate a theory of how a given neural population encodes stimuli. As we do not know the representational space (or even the most appropriate feature space) of most brain regions, it has been difficult to validate dimensionality reduction as a means of uncovering representational spaces. Here we tested how effective these approaches were at 'uncovering' the representational space of area MT, which is known to encode direction, speed and other features of motion for simple stimuli. We found that dimensionality reduction methods and cluster analysis were poor at extracting and separating the known stimulus feature dimensions, even though there were robust neural responses to these features and even though the stimulus set included systematic variation along the known dimensions. We conclude by discussing why this is, and what might be done about it.

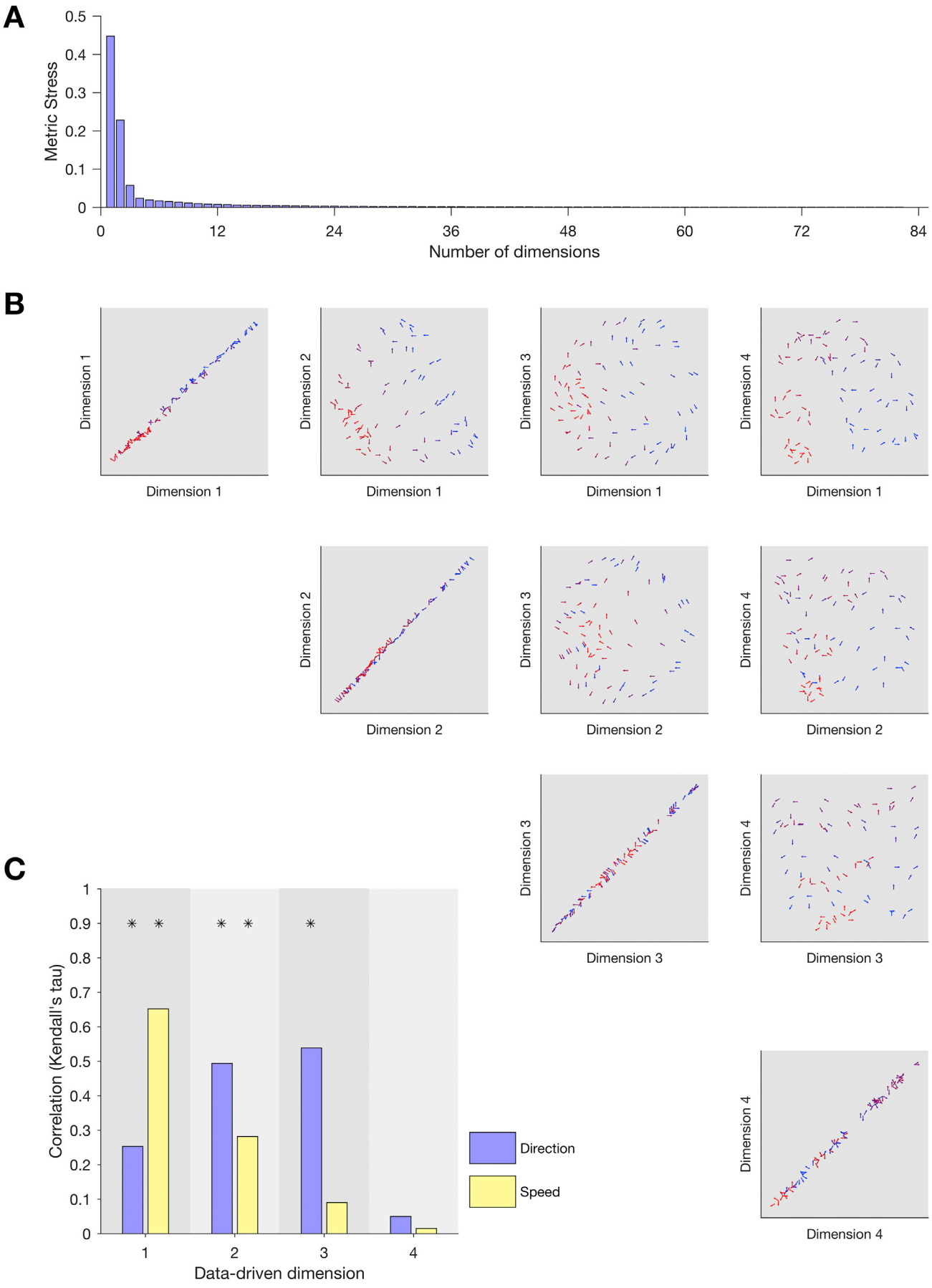
4.2. Interpretive Mania: the difficult position of the naive observer

For dimensionality reduction to be an effective way of uncovering information about representational spaces, it must reveal structure that can be readily interpreted by a naive observer who does not already know the dimensions of the feature space. Our dimensionality reductions revealed clear structure. But it is difficult to see how any of this could be used to discover that MT is systematically responsive to direction, speed, spatial and temporal frequency, if these were not already known sources of variation in feature space. We assume it is uncontroversial that MT represents direction, speed, and other features of visual motion: MT not only responds robustly to motion stimulus dimensions (Maunsell and van Essen, 1983; Albright, 1984; Movshon et al., 1985) but its activity has been associated with the perception of these motion features (Newsome et al., 1989; Salzman et al., 1990; Britten et al., 1996).

The problem is therefore that the extracted spaces bear some complex and hard-to-determine relationship to the actual representational space of MT, rather than being a simple readout of the underlying representational space. It may be that dimensionality reduction methods have failed to extract this structure because of nonlinearities in the responses of MT neurons, or because of multiplexed codes for different stimulus features within the same population (Goddard et al., 2017). Whatever the reason, these effects will only be compounded for higher-order areas such as IT. Any higher order area is likely to show greater nonlinearity in responsiveness, and is more likely to have a high dimensional neural representation (Rigotti et al., 2013; Lehky et al., 2014). Both of these factors suggest that for higher-order areas, dimensionality reduction would likely result in even less interpretable output.

The problem is not, note, that the derived spaces are inadequate to account for the variation in our stimulus. Indeed, if we did not know *that*

Fig. 7. Summary of the representational space resulting from dimensionality reduction by principal components analysis (PCA) with Varimax rotation applied to the dissimilarity matrix based on classifier performance, for responses to moving dot fields, with plotting conventions as in Fig. 6. **A**: Eigenvalues of the 84 components extracted by PCA. **B**: Data for the 84 unique stimuli projected into spaces defined by one or two of the top 4 dimensions resulting from the PCA with Varimax rotation. **C**: The correlation between the 'data-defined' and stimulus-defined dimensions is replotted (from Fig. 5) for Varimax.


A
Metric Stress
0.5
0.4
0.3
0.2
0.1
0
0
12
24
36
48
60
72
84
Number of dimensions
B
Dimension 1
Dimension 2
Dimension 3
Dimension 4
C
Correlation (Kendall's tau)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
Data-driven dimension
Direction
Speed

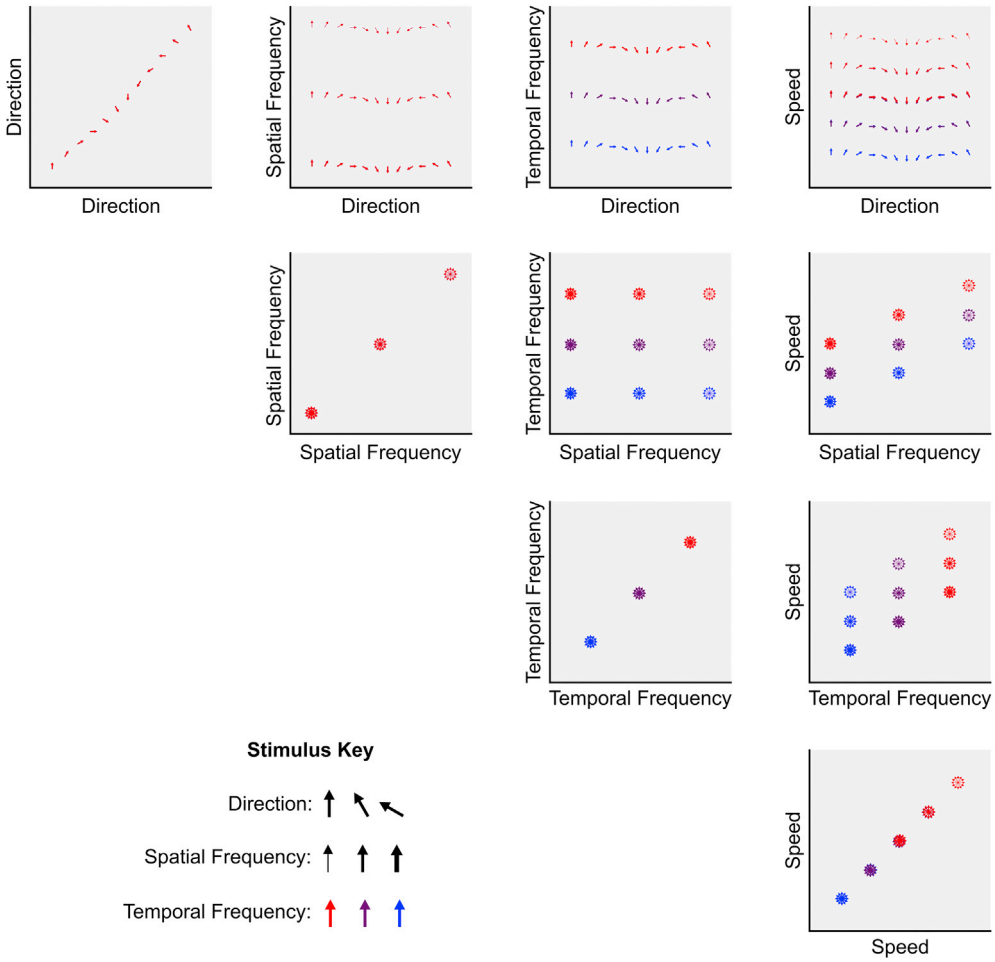


Fig. 9. Illustration of the expected data-driven dimensions for responses to moving gratings, based on the stimulus dimensions. Hypothetical dimensions based on direction, SF, TF and speed are plotted individually and against one another. The origin of each arrow indicates the location of the neural response to a single stimulus in the specified space. In each plot arrows for each of the 108 stimuli are plotted, but in many cases they are overlapping. The direction of each arrow indicates the direction of the stimulus, while its width indicates the SF, and its color indicates TF.

MT represented moving stimuli, the fact that we can extract feature spaces in which moving stimuli can be systematically situated would be decent evidence that MT does represent moving stimuli. Nor is the problem (yet) that we do not know how to choose the one that corresponds to the actual representational space. Instead, the problem is that it is very hard to interpret what has come out, and to link that back to anything intelligible about either the stimulus or to the brain.

The attractive feature of data-driven methods was supposed to be that they gave 'objective' results, without bias from experimenters. Yet even if we accept that these data may contain new insights about the representational space of area MT, the experimenter has to work hard to extract anything meaningful—and it is therefore unclear whether anything extracted is really objective. This is trivially apparent from the fact that different dimensionality reduction methods showed considerable variation in the dimensions they extracted. It is unclear how the true solution could be selected from these possible candidates, or on what basis one could decide to reject all these solutions.

Similarly, there remains the problem of how to evaluate which methods are extracting true signal and which are best interpreted as noise. Even within a single method, there is the issue of how one should decide which dimensions are most informative for interpreting the representational space. For example, MDS was one of the better methods for isolating temporal frequency from other dimensions in the moving gratings data (Fig. 10), but it isolated temporal frequency on the 5th dimension. It is unclear whether a naive observer could realize that this dimension was capturing important information about the representational space when most of the variance appears to be explained by fewer dimensions (the metric stress for the MDS solution is approaching an asymptote by the 5th dimension, as seen in Fig. 13A).

It may be tempting to conclude that close enough is good enough in these results, for example, when looking at the results of PCA for moving dots data (Fig. 6). However, aside from the problems of choosing which method has revealed useful structure, and which of the extracted dimensions are the most relevant, there is a further problem. Even for this very simple case where the moving dot stimuli varied along only 2 dimensions, the PCA requires at least 3 of the extracted dimensions to account for the data, and both dimension 2 and 3 are misleading unless they are combined into a circular dimension in a Cartesian plane. This is unsurprising: by definition PCA will extract orthogonal linear dimensions. Yet when it comes to interpretation, a naive observer could not know whether the feature space is best captured by linear, circular, or other dimensions. When considering a stimulus set with 3 or 4

Fig. 8. Summary of the representational space resulting from dimensionality reduction by multi-dimensional scaling (MDS) of the dissimilarity matrix based on classifier performance, for responses to moving dot fields, with plotting conventions as in Fig. 6. **A**: The metric stress of the MDS solution where increasing numbers of dimensions were allowed. **B**: Data for the 84 unique stimuli projected into spaces defined by one or two of the dimensions from a space where the MDS solution was restricted to 4 dimensions. **C**: The correlation between the 'data-defined' and stimulus-defined dimensions is replotted (from Fig. 5) for MDS.

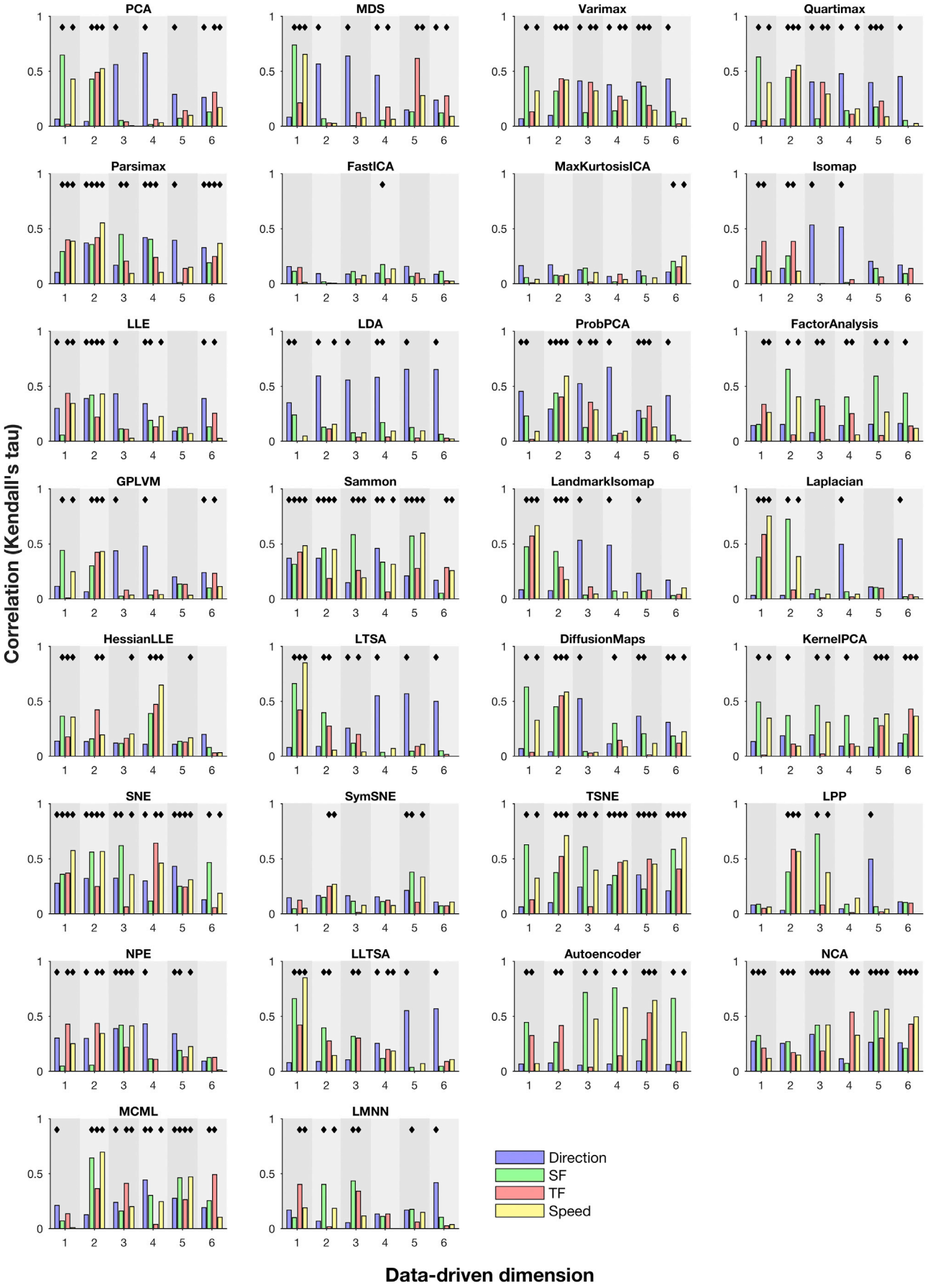


Fig. 10. Correlation between data-defined dimensions extracted from the dissimilarity matrix based on classifier performance and the known stimulus dimensions of moving grating direction, spatial frequency (SF), temporal frequency (TF), and speed (TF/SF). For each dimensionality reduction method tested, we considered only the top 6 dimensions that were extracted, and correlated these with each of the hypothetical direction and speed dimensions (see Materials and Methods for details). Plotting conventions as in Fig. 4.

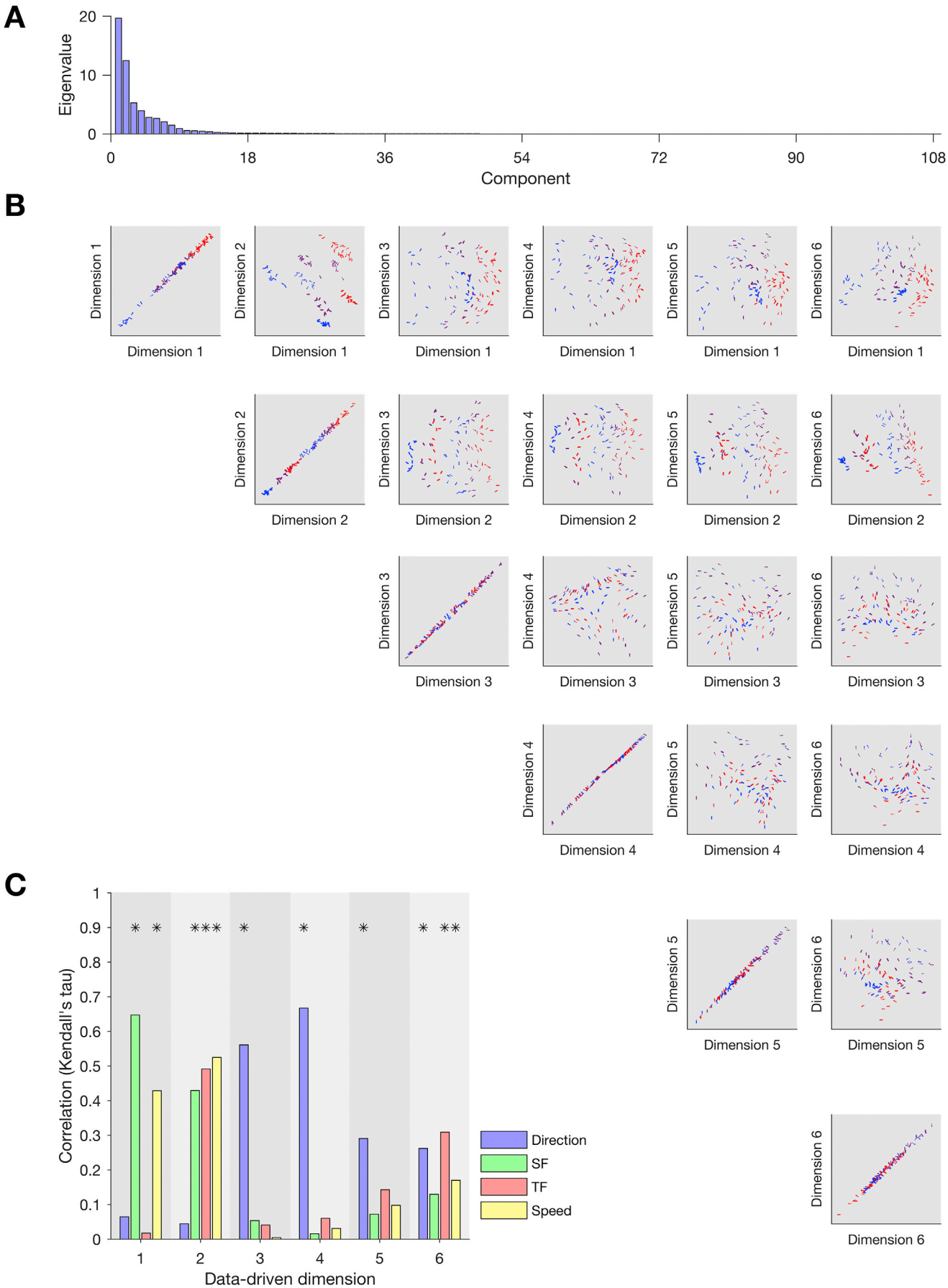


Fig. 11. Summary of the representational space resulting from dimensionality reduction by principal component analysis (PCA) of the dissimilarity matrix based on classifier performance, for responses to moving gratings, with plotting conventions as in Fig. 6, except that in **B** the direction of each arrow indicates the direction of the stimulus, while its width indicates the SF (thick = low, thin = high), and its color indicates TF (red = low, blue = high).

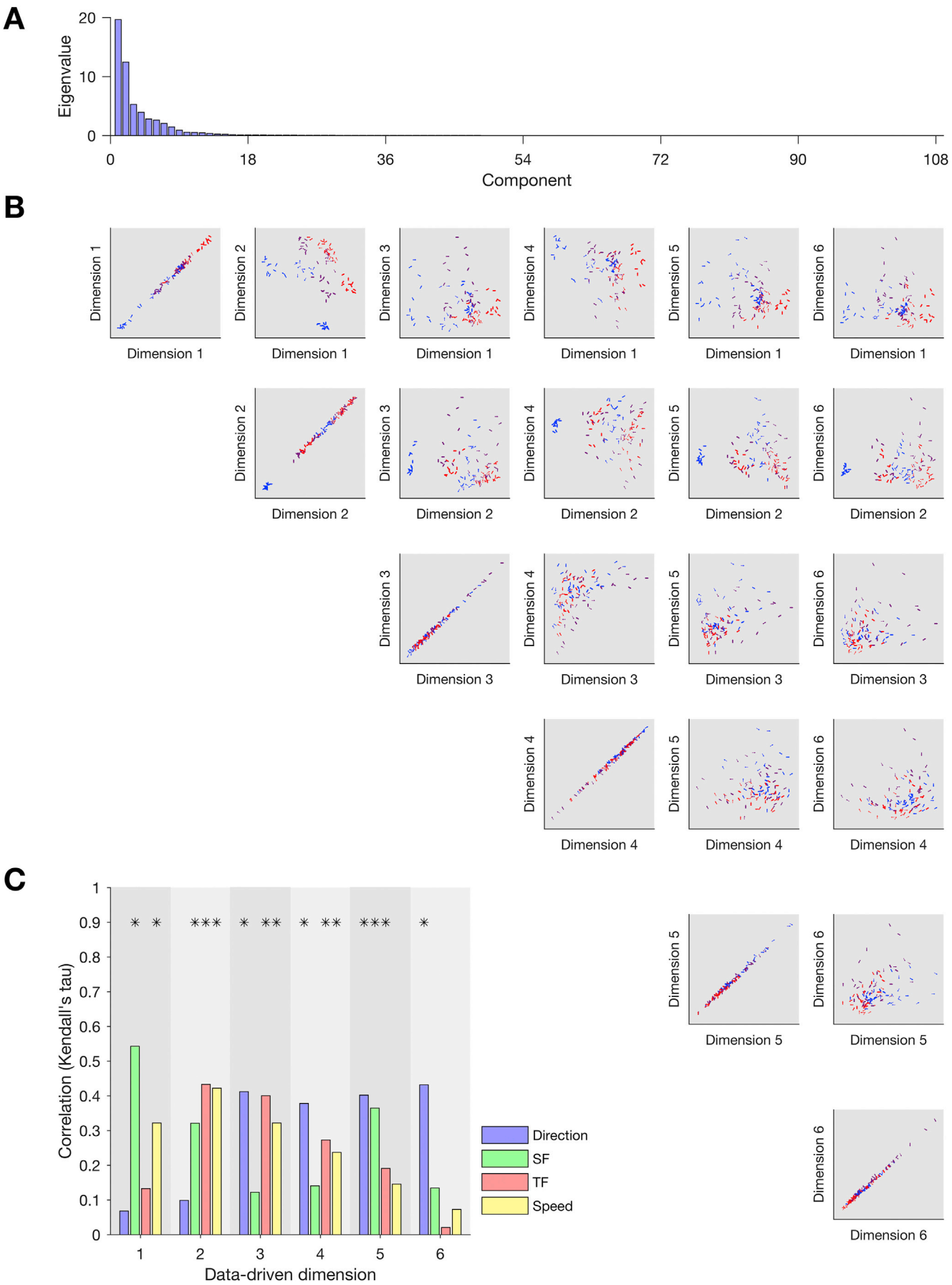


Fig. 12. Summary of the representational space resulting from dimensionality reduction by principal components analysis (PCA) with Varimax rotation applied to the dissimilarity matrix based on classifier performance, for responses to moving gratings, with plotting conventions as in Fig. 11.

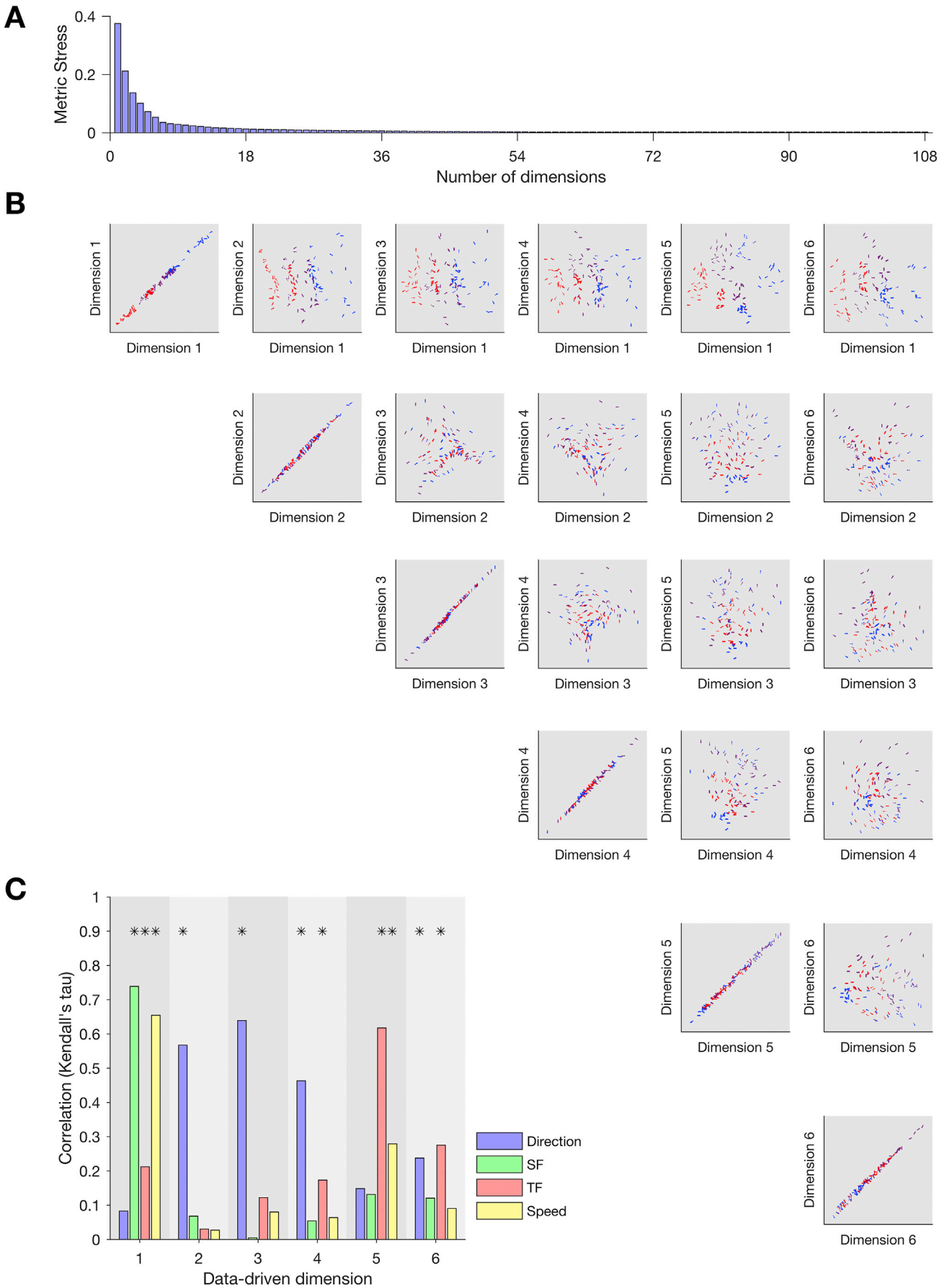
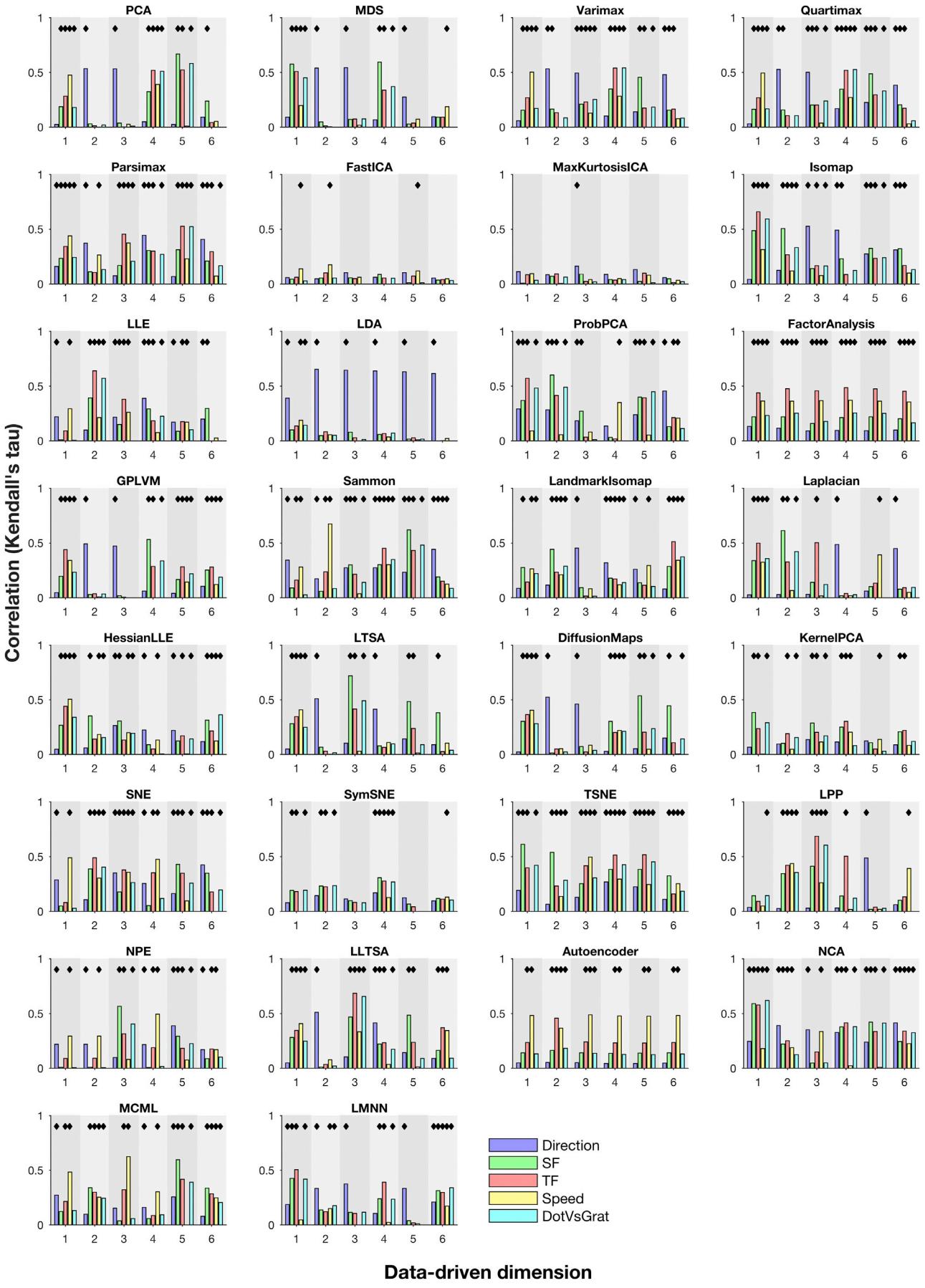


Fig. 13. Summary of the representational space resulting from dimensionality reduction by multi-dimensional scaling (MDS) of the dissimilarity matrix based on classifier performance, for responses to moving gratings, with plotting conventions as in Fig. 8, and definition of the arrows as in Fig. 11.


PCA
MDS
Varimax
Quartimax
Parsimax
FastICA
MaxKurtosisICA
Isomap
LLE
LDA
ProbPCA
FactorAnalysis
GPLVM
Sammon
LandmarkIsomap
Laplacian
HessianLLE
LTSA
DiffusionMaps
KernelPCA
SNE
SymSNE
TSNE
LPP
NPE
LLTSA
Autoencoder
NCA
MCML
LMNN
Correlation (Kendall's tau)
Data-driven dimension
Direction
SF
TF
Speed
DotVsGrat

dimensions (in the gratings data, and the combined data) the result is even less clear.

4.3. Stimulus selection: a chicken and egg problem

A distinct but related worry involves the choice of stimuli. As we knew (some of) the parameters of the representational space of MT, we could take care to vary the stimuli parametrically across these dimensions, and to 'tile' this feature space so that each combination of features was included. An alternative but valid approach (perhaps more appropriate when the underlying space is less clear) would be to randomly sample across combinations of levels, so as to approach adequate coverage (see Judd et al., 2012). Within the time constraints of a standard experiment, even a random sampling will most likely be a random sampling of a very small feature space (e.g. Nestor et al., 2016), or else a very sparse sampling of a larger space. Thus even with a random sampling, the experimenter makes choices about feature sampling that will likely alter the extracted representational spaces.

A related issue for stimulus selection is that the true dimensions of the representational space may covary with a spurious stimulus dimension that is mistakenly interpreted as the true dimension. For example, if we covaried direction and color of the moving dots, we might mistakenly conclude that area MT systematically responds to color. While such a stimulus set is easily rejected as flawed, since direction and color are recognizable stimulus dimensions, this becomes a non-trivial problem where the stimulus dimensions are unknown.

These two desiderata—adequately sampling a feature space on the one hand, and not inadvertently mixing up features on the other—seem like they should be basic preconditions for an adequate data-driven analysis. We were able to meet them because, again, we knew what we were looking for. But the goal of a data-driven analysis is to discover such dimensions precisely when the dimensions are unknown and these preconditions are difficult to meet. In this way, a data-driven analysis on an unknown feature space faces a chicken-and-egg problem. Without knowing the feature space, it is difficult to know whether one has appropriately sampled the feature space, but one cannot determine an unknown feature space without appropriate sampling.

Although the stimulus sets are often large, they are usually not large enough or random enough for warrant the conclusion that the results are 'data-driven'. The stimulus sets are typically selected to sample a stimulus range that covers the dimensions of interest for the experimenter. Stimulus selection is difficult, for example when taking into account the fact that there are low level visual similarities between objects of the same category when investigating object perception (Groen et al., 2012, 2013). Further, in many cases, experimenters choose the stimuli because they appear to vary in ways that are perceptually salient to us. But then are we really extracting the dimensions of the underlying feature space—that is, what the brain actually represents—or could we be using brain data to recapture and summarize features of the stimulus? Or to summarize feature dimensions that were perceptually salient to the experimenter?

When the dimensions and stimulus categories of interest are already defined in the stimuli, this makes it impossible to judge whether the dimensions that are recovered are simply describe the stimulus set, or whether they are true dimensions that the underlying neural populations represent. This may appear to be a subtle distinction, but it has significant consequences for how the results of these analyses should be interpreted.

Data-driven methods promise a way to extract feature spaces without the experimenter introducing any direct hypotheses about the feature space. Yet without an independent method to guide stimulus selection, there is always the possibility that the choice of stimulus set constitutes an indirect hypothesis about the feature space. In some cases the indirect hypothesis is clear, for example in Caspari et al. (2014) when the stimulus set was designed to contain equal numbers of a small number of different categories. In such cases we believe it is inappropriate to treat this as bias-free or hypothesis-neutral confirmation: at best, it shows that dimensionality reduction is able to extract structure that we already assumed was there. Conversely, one might reject a proper tiling of feature space as uninterpretable. Consider again our results: if one did not have prior knowledge of the feature, one might be tempted to blame the stimulus set for the lack of clear representational structure or (worse) tweak it until more intelligible results were found.

We reiterate that both of these problems have arisen for a straightforward case, where we have a good sense of what the underlying brain region represents, and we selected our stimuli accordingly. It is even less probable that these methods could uncover meaningful novel structure when they are applied to brain areas with unknown dimensions of interest, using stimulus sets which may vary along hundreds of dimensions. In many ways the data here are a best case scenario for the dimensionality reduction methods, with a robust neural signal measured for a systematic stimulus set. The issues found here will likely only get worse for data where neural signals are weaker, or the stimuli have not been systematically varied along behaviorally relevant orthogonal dimensions.

4.4. What next? The merit of exploratory analyses and 3 recommendations for data-driven approaches

We do not aim to paint a completely bleak picture. Nor do we want to reject existing literature where dimensionality reduction has been used to reveal compelling and reasonable structure in the brain's representational space (for example, Kriegeskorte et al., 2008; Vul et al., 2012; Caspari et al., 2014). Instead, we suggest, most previous work should be interpreted as exploratory rather than data-driven. Dimensionality reduction analyses lay on a scale from the exploratory to the truly data-driven. Even exploratory analyses play an important role: danger arises when mistaking an analysis for one that is stronger than it actually is.

An exploratory analysis, in the simplest form, gives you evidence that, for the stimuli, parameters, and contexts that were examined, there is a feature space that can capture the variation in those parameters. That feature space may not generalize to other stimuli, it may be a distorted projection of a high-dimensional feature space, and it may fail in different contexts (e.g. a move from dots to naturalistic stimuli, or if there were a different sampling of the same feature space). As per above, the feature space that is extracted may also be one of many ways to capture that variation. The conclusion is therefore quite weak. However, exploratory analyses play an important role because they suggest further hypotheses regarding representational spaces and can be used to generate testable models (for example, Machens et al., 2010), which is a nontrivial advantage.

We think that many previous results that were described as data-driven are likely better considered to be exploratory analyses, meaning that the extracted structure may reflect the design of the stimulus set rather than purely reflecting the feature dimensions that are of most importance to the neural population (Kriegeskorte et al., 2008; Vul et al., 2012; Caspari et al., 2014). This does not mean that the results are failing to reveal real structure, but it does mean that they should not be considered hypothesis-neutral and therefore given undue weight when evaluating evidence for the functionality of a brain region. As exploratory analyses, the analyses do not carry any weight as evidence for a given representational structure, but the revealed structures can be followed up with more traditional hypothesis tests which seek to disambiguate the

Fig. 14. Correlation between dimensions extracted from the dissimilarity matrix based on classifier performance and the known stimulus dimensions in the data. We considered dimensions based on grating stimulus direction, spatial frequency (SF), temporal frequency (TF), speed (TF/SF) and form category (DotVsGrat: moving dot field or moving grating). For each dimensionality reduction method tested, we considered only the top 6 dimensions that were extracted, and correlated these with each of the hypothetical direction and speed dimensions (see Materials and Methods for details). Plotting conventions as in Fig. 4.

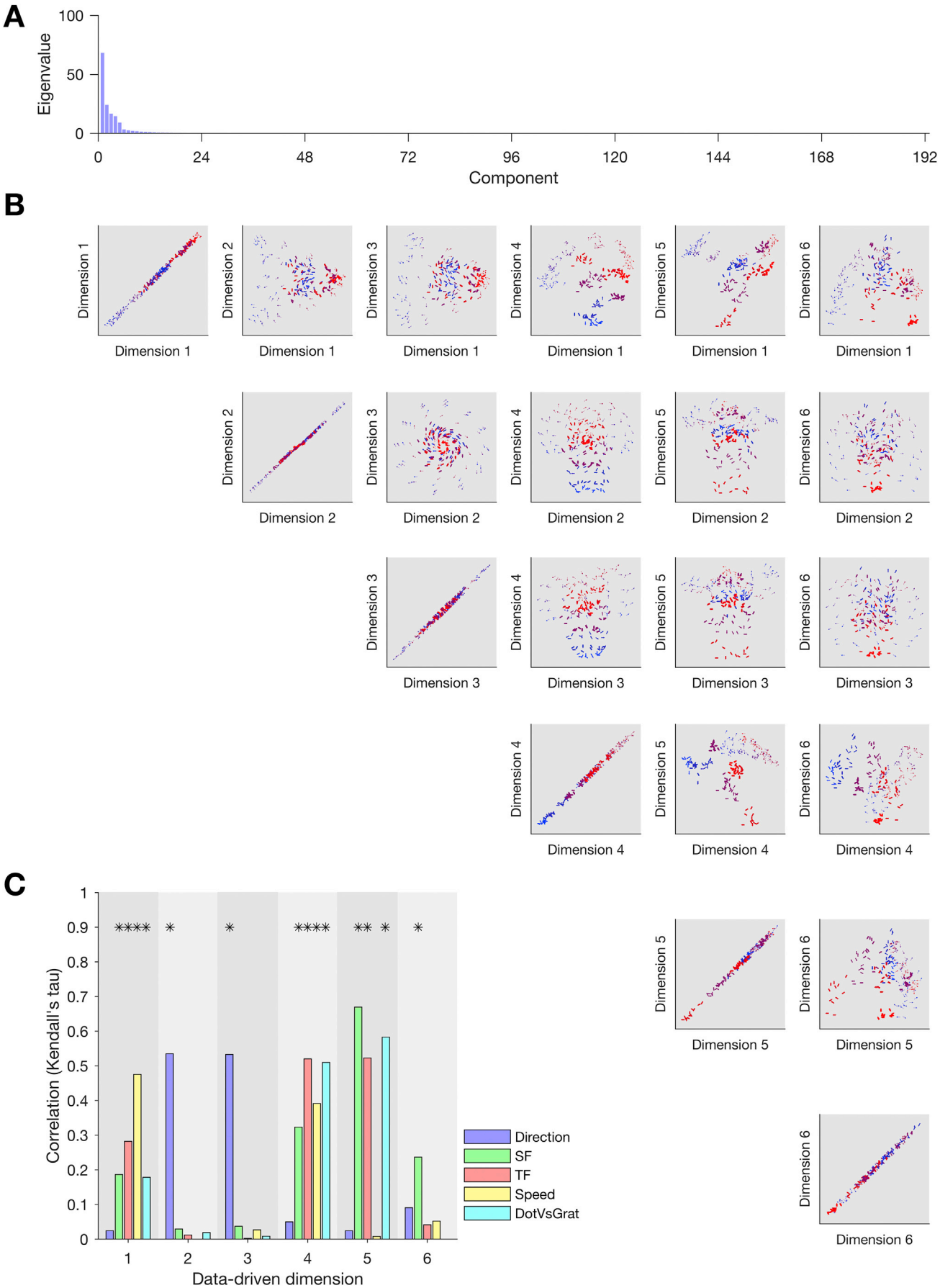


Fig. 15. Summary of the representational space resulting from dimensionality reduction by principal component analysis (PCA) of the dissimilarity matrix based on classifier performance, for responses to the total stimulus set, with plotting conventions as in Fig. 6. The definition of the arrows is as in Fig. 9, with the addition that dotted arrows are dot field stimuli, and solid arrows are grating stimuli.

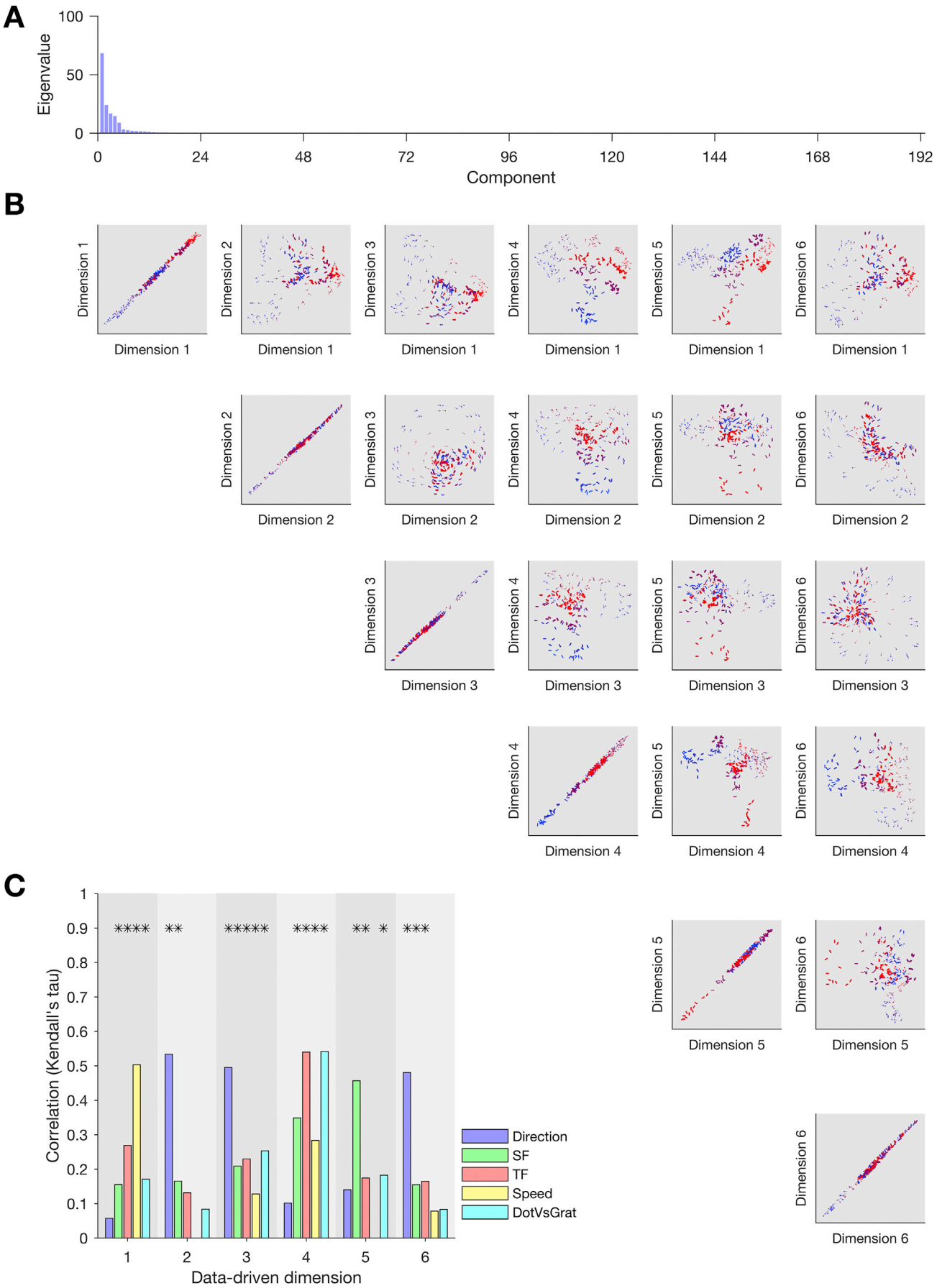


Fig. 16. Summary of the representational space resulting from dimensionality reduction by principal components analysis (PCA) with Varimax rotation applied to the dissimilarity matrix based on classifier performance, for responses to the total stimulus set, with plotting conventions as in Fig. 15.

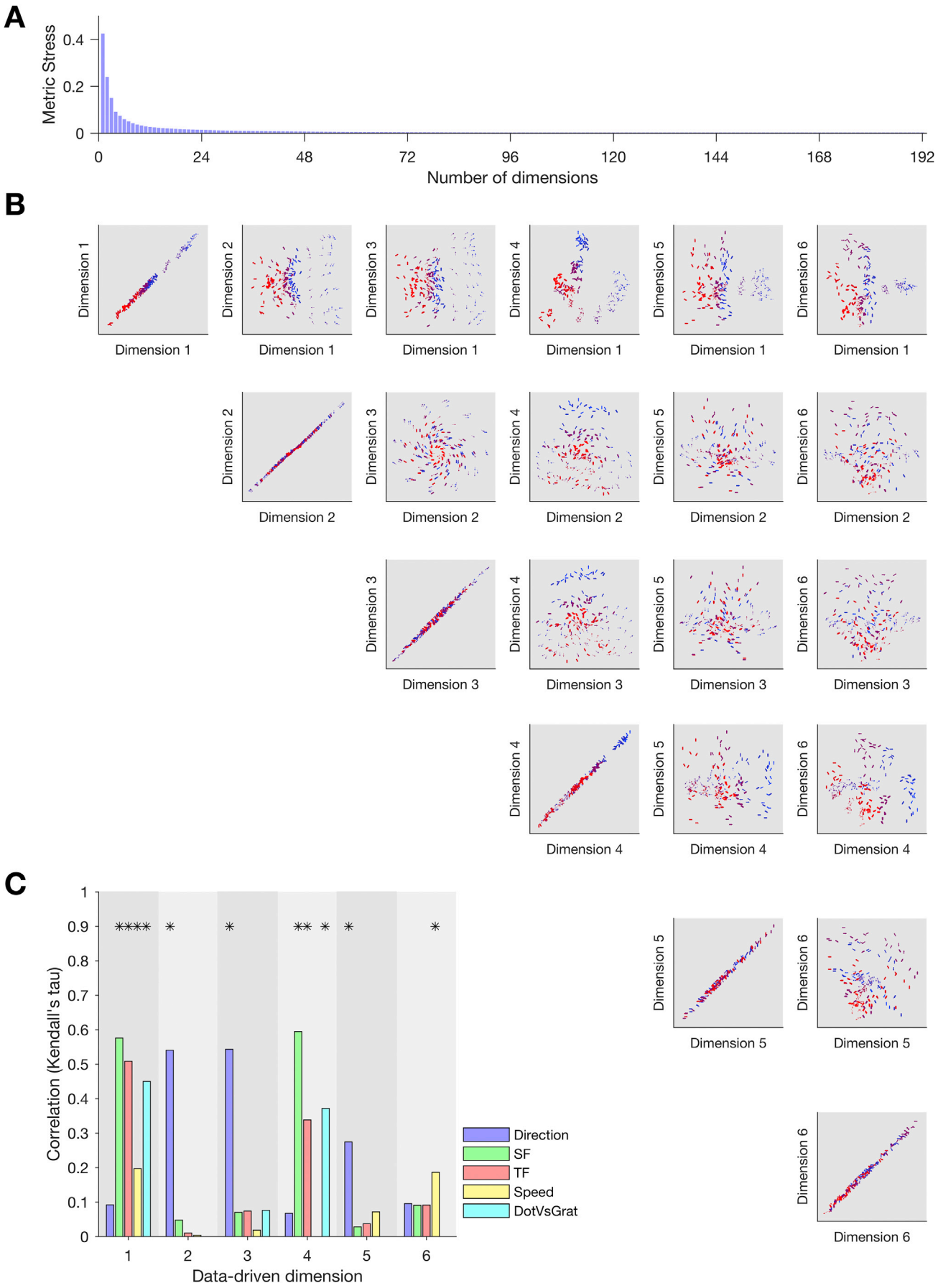


Fig. 17. Summary of the representational space resulting from dimensionality reduction by multi-dimensional scaling (MDS) of the dissimilarity matrix based on classifier performance, for responses to the total stimulus set, with plotting conventions as in Fig. 15.

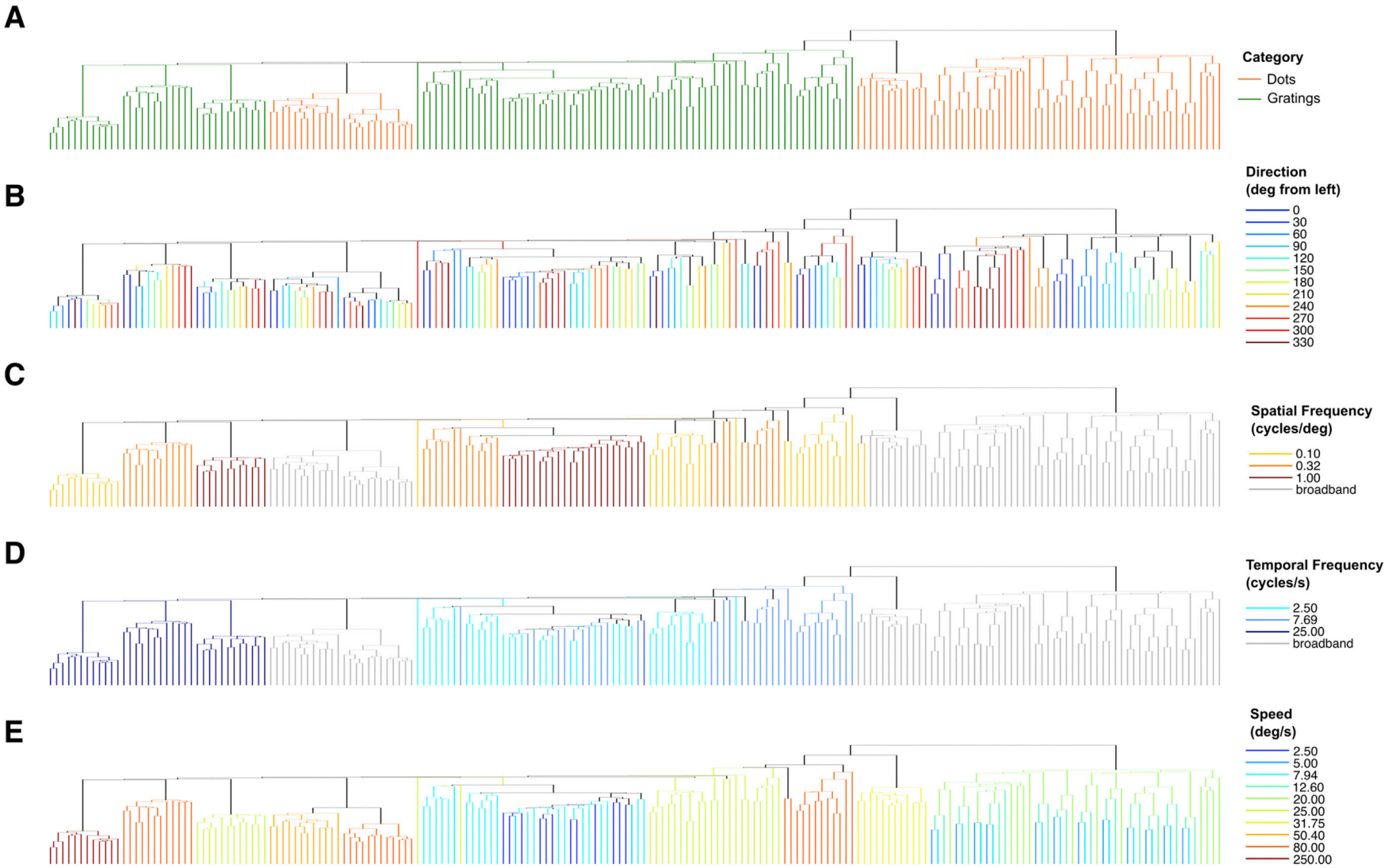


Fig. 18. Result of the cluster analysis of the dissimilarity matrix based on classifier performance, for responses to the total stimulus set. In **A–E** the same hierarchical tree is plotted, where each vertical line at the base of the tree indicates the response to a single stimulus. In **A–E** the lines are shaded according to stimulus category (dot field versus grating), direction, SF, TF and speed, respectively.

hypothesized model from alternative accounts.

Especially when used as an exploratory approach, the interpretability of the extracted dimensions will likely be enhanced by relating them to expected dimensions of the neural representation, defined either by stimulus dimensions, measures of behavior, or other independently defined measurements (Nestor et al., 2016; Mante et al., 2013; Cohen and Maunsell, 2010). For example, (Mante et al., 2013) used dimensionality reduction as an intermediate step in relating their population neuronal data to the task-defined space. Similarly, Cohen and Maunsell (2010) related an extracted dimension to behavioral performance and used this dimension as an index of attentional state in their task. Churchland and Cunningham (2014) use 'hypothesis-guided' dimensionality reduction as a means of distinguishing between alternate models for motor cortex responses during reaching. However, even when comparing extracted dimensions with independently defined dimensions, the interpretability is limited by the impossibility of distinguishing between a deficient model of the brain region's representational space and a failure of the dimensionality reduction methods to uncover the true dimensions.

What can be done to move towards a full data-driven analysis, especially when the underlying feature space is unknown? In addition to the stimulus selection considerations outlined above, we think there are several options, and that work on dimensionality reduction ought to concentrate on producing more.

First, our results show that a single dimensionality reduction is of questionable value. More useful might be the application of a range of dimensionality reduction approaches to check how robust the findings are across different methods. Evaluating the suitability of different dimensionality reduction methods is beyond the scope of this paper, but it is likely more important to use a range of possible methods than to identify the single most appropriate one. Our results show that the relationship between different approaches is unlikely to be straightforward agreement. If that is the case, then the experimenter ought to justify why one particular approach is the best—or, more likely, talk about the relationships between the dimensions extracted by different techniques and what they might mean for the underlying representational structure.

Second, for each of these methods, it would be instructive to consider a range of the extracted dimensions, rather than only the first few. More generally, it would be good to develop principled ways to determine the number of dimensions that are considered, and for excluding some from the search. A running theme of the above has been that it is easy to see structure where there is none, and easy to dismiss as noise or failed technique what is actually unexpected structure. Similarly, it is appropriate to plot each of the dimensions that are explaining considerable variance, since although some may not have readily interpretable structure, there may be structure that is seen by others, or that can be interpreted in light of future findings. Also consider that there may be latent dimensions in the representational space that are circular, or comprise some other interaction between 2 or more of the extracted dimensions.

Third and finally, if the extracted dimensions capture true features of the underlying space, they should be replicable across a range of different data sets. Extracting a feature space that captures variation within the stimulus set is useful. But unless this is tested with other stimuli, and in other contexts, it is unclear whether the dimensions are meaningful for understanding the representational space, and predicting neural responses to novel stimuli.

The MT data presented here demonstrate that even when the stimulus set clearly varies parametrically along feature dimensions, the methods do not necessarily extract these dimensions in any straightforward way. This highlights the importance of ensuring that any extracted dimensions are reliable, stable, and robust across a range of factors including dimensionality reduction method and stimulus set. Determining stability and robustness across stimulus set is an extension of cross-validation techniques. At a minimum, if the recovered structure is robust it should be replicable when the stimulus set is divided in half and tested separately, or when applied to an entirely new data set (for example, see Huth et al., 2016b). As discussed above, when aiming to be as 'data-driven' as possible, a good stimulus set should be as large, diverse and randomly structured as feasible.

In summary, dimensionality reduction is a potentially useful tool for understanding the structure of neural representations, particularly suited to exploratory analyses. Such exploratory analyses are especially useful for identifying the most promising avenues in which to invest future efforts. To maximize the usefulness of dimensionality reduction, researchers should interpret results from these approaches in accordance with the extent to which their design is exploratory or data-driven. For data-driven designs, the aim should be to reveal representational structures that randomly or evenly sample a large stimulus space, and that are reliable, stable and robust to methodological and stimulus variations.

Acknowledgments

This project was funded under Australian Research Council Future Fellowships to C.K. and T.A.C. (FT140100422, FT120100816), an ARC Discovery Project to T.A.C. (DP160101300), and a National Health and Medical Research Council of Australia Project Grant to S.G.S. (APP1005427). We thank S.S. Solomon, S.K. Cheong, S.C. Chen and A.S. Pietersen for assistance with electrophysiological data collection.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.neuroimage.2017.06.068.

References

Adolphs, R., Nummenmaa, L., Todorov, A., Haxby, J.V., 2016. Data-driven approaches in the investigation of social perception. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. 371.

Albright, T.D., 1984. Direction and orientation selectivity of neurons in visual area MT of the macaque. J. Neurophysiol. 52, 1106–1130.

Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332.

Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S., Movshon, J.A., 1996. A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis. Neurosci. 13, 87–100.

Carlson, T.A., Goddard, E., Kaplan, D.M., Klein, C., Ritchie, J.B., 2017. Ghosts in machine learning for cognitive neuroscience. Neuroimage 180 (Part A), 88–100.

Caspari, N., Popivanov, I.D., De Mazière, P.A., Vanduffel, W., Vogels, R., Orban, G.A., Jastorff, J., 2014. Fine-grained stimulus representations in body selective areas of human occipito-temporal cortex. Neuroimage 102, 484–497. Pt 2.

Churchland, M.M., Cunningham, J.P., 2014. A dynamical basis set for generating reaches. Cold Spring Harb. Symp. Quant. Biol. 79, 67–80.

Cohen, M.R., Maunsell, J.H.R., 2010. A neuronal population measure of attention predicts behavioral performance on individual trials. J. Neurosci. 30, 15241–15253.

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618.

Cunningham, J.P., Yu, B.M., 2014. Dimensionality reduction for large-scale neural recordings. Nat. Neurosci. 17, 1500–1509.

de Wit, L., Alexander, D., Ekroll, V., Wagemans, J., 2016. Is neuroimaging measuring information in the brain? Psychon. Bull. Rev. 23, 1415–1428.

Goddard, E., Solomon, S., Carlson, T., 2017. Dynamic population codes of multiplexed stimulus features in primate area MT. J. Neurophysiol. http://dx.doi.org/10.1152/jn.00954.2016. http://jn.physiology.org/content/early/2017/03/31/jn.00954.2016.

Groen, I.I.A., Ghebreab, S., Lamme, V.A.F., Scholte, H.S., 2012. Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. PLoS Comput. Biol. 8 e1002726.

Groen, I.I.A., Ghebreab, S., Prins, H., Lamme, V.A.F., Scholte, H.S., 2013. From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. J. Neurosci. 33, 18814–18824.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016a. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.

Huth, A.G., Lee, T., Nishimoto, S., Bilenko, N.Y., Vu, A.T., Gallant, J.L., 2016b. Decoding the semantic content of natural movies from human brain activity. Front. Syst. Neurosci. 10, 81.

Judd, C.M., Westfall, J., Kenny, D.A., 2012. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. J. Personal. Soc. Psychol. 103, 54–69.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proc. Natl. Acad. Sci. U. S. A. 107, 11163–11170.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141.
Lashkari, D., Vul, E., Kanwisher, N., Golland, P., 2010. Discovering structure in the space of activation profiles in fMRI. Neuroimage 50, 1085–1098.
Lehky, S.R., Kiani, R., Esteky, H., Tanaka, K., 2014. Dimensionality of object representations in monkey inferotemporal cortex. Neural Comput. 26, 2135–2162.
Machens, C.K., Romo, R., Brody, C.D., 2010. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. J. Neurosci. 30, 350–360.
Mante, V., Sussillo, D., Shenoy, K.V., Newsome, W.T., 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84.
Maunsell, J.H., van Essen, D.C., 1983. Functional properties of neurons in middle temporal visual area of the macaque monkey. i. selectivity for stimulus direction, speed, and orientation. J. Neurophysiol. 49, 1127–1147.
Mazor, O., Laurent, G., 2005. Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. Neuron 48, 661–673.
McDonald, J.S., Clifford, C.W.G., Solomon, S.S., Chen, S.C., Solomon, S.G., 2014. Integration and segregation of multiple motion signals by neurons in area MT of primate. J. Neurophysiol. 111, 369–378.
Movshon, J.A., Adelson, E.H., Gizzi, M., Newsome, W.T., 1985. The Analysis of Moving Visual Patterns (chapter 54). Vatican Press, pp. 117–151.
Nestor, A., Plaut, D.C., Behrmann, M., 2016. Feature-based face representations and image reconstruction from behavioral and neural data. Proc. Natl. Acad. Sci. U. S. A. 113, 416–421.
Newsome, W.T., Britten, K.H., Movshon, J.A., 1989. Neuronal correlates of a perceptual decision. Nature 341, 52–54.
Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., Fusi, S., 2013. The importance of mixed selectivity in complex cognitive tasks. Nature 497, 585–590.
Ritchie, J.B., Kaplan, D.M., Klein, C., 2017. Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. Br. J. Philos. Sci. http://dx.doi.org/10.1101/127233 (in press).
Rosa, M.G., Elston, G.N., 1998. Visuotopic organisation and neuronal response selectivity for direction of motion in visual areas of the caudal temporal lobe of the marmoset monkey (callithrix jacchus): middle temporal area, middle temporal crescent, and surrounding cortex. J. Comp. Neurol. 393, 505–527.
Rousche, P.J., Normann, R.A., 1992. A method for pneumatically inserting an array of penetrating electrodes into cortical tissue. Ann. Biomed. Eng. 20, 413–422.
Salzman, C.D., Britten, K.H., Newsome, W.T., 1990. Cortical microstimulation influences perceptual judgements of motion direction. Nature 346, 174–177.
Sha, L., Haxby, J.V., Abdi, H., Guntupalli, J.S., Oosterhof, N.N., Halchenko, Y.O., Connolly, A.C., 2015. The animacy continuum in the human ventral vision pathway. J. Cognit. Neurosci. 27, 665–678.
Solomon, S.S., Chen, S.C., Morley, J.W., Solomon, S.G., 2015. Local and global correlations between neurons in the middle temporal area of primate visual cortex. Cereb. Cortex 25, 3182–3196.
Stokes, M.G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., Duncan, J., 2013. Dynamic coding for cognitive control in prefrontal cortex. Neuron 78, 364–375.
van der Maaten, L., Hinton, G., 2008. Visualizing data using t-sne. J. Mach. Learn. Res. 9, 2579–2605.
Vul, E., Lashkari, D., Hsieh, P.J., Golland, P., Kanwisher, N., 2012. Data-driven functional clustering reveals dominance of face, place, and body selectivity in the ventral visual pathway. J. Neurophysiol. 108, 2306–2322.
Zinszer, B.D., Anderson, A.J., Kang, O., Wheatley, T., Raizada, R.D.S., 2016. Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. J. Cognit. Neurosci. 1–11.