e-Neuroforum 2010 · 1:89–98
DOI 10.1007/s13295-010-0011-0

D. Durstewitz · E. Balaguer-Ballester
Bernstein Center for Computational Neuroscience Heidelberg-Mannheim,
Central Institute of Mental Health & Heidelberg University, Mannheim

Statistical approaches for reconstructing neuro-cognitive dynamics from high-dimensional neural recordings

Introduction

The life sciences in general, and neurosciences in particular, produce data sets of ever-increasing size and complexity. Simultaneous recordings from tens to hundreds of neurons using optical imaging techniques or bunches of tetrodes are now becoming routine. Likewise, in non-invasive neuroimaging approaches such as functional magnetic resonance imaging (fMRI), the activity of up to tens of thousands of units may be obtained as a function of time. In statistical terms, all these data sets comprise examples of multivariate time series (**Fig. 1**) where a vector (or a matrix preserving spatial neighborhoods) of activity values is observed as a function of time. More and more often several of these techniques are combined, like fMRI and electroencephalography (EEG), yielding data sets which are multimodal in addition to being multivariate. Generally, these multivariate neural recordings are also not made in isolation, but are to be related to a wealth of simultaneously obtained behavioral, genetic, molecular, or other information.

These often huge and complex data sets pose tough challenges for data analysis and statistical inference, by which we mean here the methods of extracting interesting information from the multivariate measurements and assigning probabilities to our observations (thus inferring properties of the population from a sample). Of course, more "traditional" approaches, such as considering stimulus-related firing rate changes of single units, or computing (pairwise) cross-correlations between recorded units, will still work. From this perspective the advances in neural recording and imaging techniques are basically just means to generate more data in shorter periods of time, which are then analyzed like traditional recordings from just one or a few units. However, the simultaneous measurement of tens to thousands of neural elements offers completely novel and exciting opportunities which may require more advanced mathematical techniques for their full exploitation, many of which may have already been around in statistics for quite some time. Here we will summarize some of these approaches with a focus on multiple single-unit recordings from behaving animals (although the methods discussed are much more widely applicable).

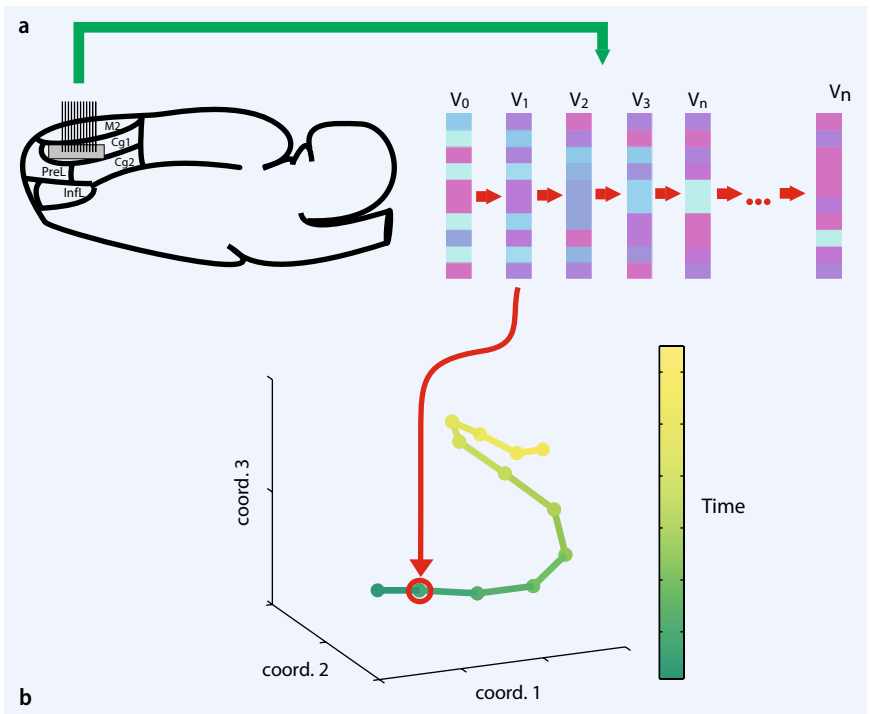


Fig. 1 ▲ A multivariate (vector) time series obtained from multiple single-unit recordings (**a**) and its projection into a three-dimensional space (**b**) for purposes of data reduction and visualization (see also **Fig. 2**). (Part **a**, *left*, reproduced with kind permission from [9], copyright 2008 by the National Academy of Sciences of the USA)

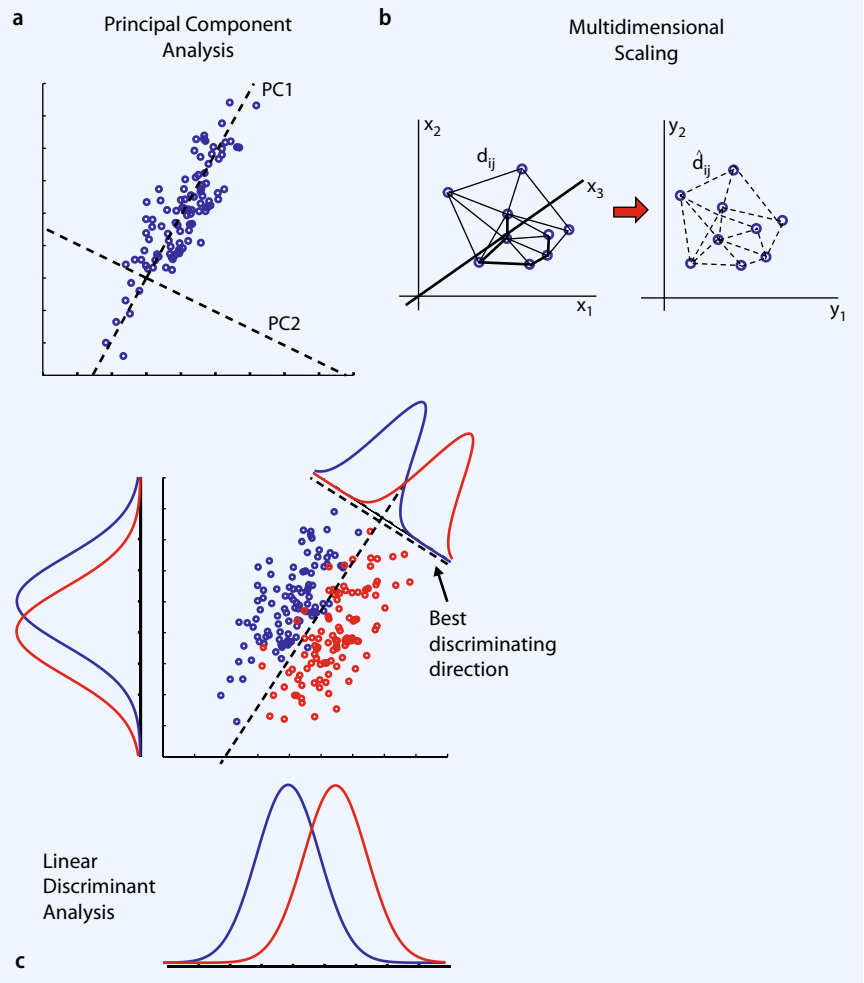


Fig. 2 ▲ Methods for dimensionality reduction. **a** Principal component (*PC*) analysis rotates axes of the original space such that they align with directions in the data cloud along which the variance is largest, and such that they are orthogonal to each other (that is, as illustrated, the first PC captures most of the data variance, the second PC captures the second largest proportion of data variance spread perpendicularly to the first PC, and so on). The hope is that just a few dimensions suffice to represent most of the variation in the data. For instance, in the example shown, PC2 may be dropped as data points spread out mainly along PC1. **b** Multidimensional scaling directly attempts to find a lower dimensional embedding of the data that preserves distances (indicated by the *connecting lines*) in the original space to the largest degree possible. **c** Fisher's discriminant analysis (FDA) finds a set of new axes (also called canonical variates) which bring out differences between predefined groups (*red* and *blue colored dots*) most clearly, i.e., along which groups can be most clearly differentiated. For a total of *g* groups, a maximum of *g*-1 such axes will result, which, however, are not necessarily orthogonal to each other. Points in this example were drawn from two-variate Gaussians. The distribution graphs indicate that the two classes of points largely overlap along the original x- and y-axis, but are quite well separated along the best discriminating direction according to the FDA criterion

Visualizing multivariate data sets

Dimensionality reduction and neural state spaces

The challenge of analyzing multivariate neural time series may be seen as one of discovering spatio-temporal structure and patterns in the data, which are then to be interpreted in the cognitive or behavioral context imposed by the experimental setting. A first step in this endeavor may be visualizing the multivariate activity to get a feel for how the neural dynamic evolves in time (**Fig. 1**), or how it depends on different task or behavioral conditions (**Fig. 3**). One common procedure here is to reduce the high dimensionality of the original data set to just two or three dimensions ready for visualization by exploiting redundancies in the data (**Fig. 2**). Other methods try to summarize the data in terms of prototypes or classes (e.g., different types of cluster analysis). The statistical literature abounds with different mathematical techniques to achieve these goals (e.g., [5, 8]), some dating back more than a 100 years. **Fig. 2** illustrates some of the most prominent of these techniques which have been used to visualize multiple single-unit recordings, and their principles of operation: Probably the most commonly used method is principal component analysis (PCA, **Fig. 2a**) which is a linear technique for projecting the data into a lower-dimensional space that conserves most of the original data variance. More precisely, PCA works by rotating the axes of the original space in such a way that they: (a) align with the directions of highest data variance in the original space, and (b) are orthogonal to each other, i.e., uncorrelated if the mean of each variable were subtracted off beforehand, such that each rotated dimension reflects a 'novel aspect' of the data roughly 'independent' of the others. The hope is that only a few dimensions may explain most of the variation in the data, and hence dimensions which capture only little data scatter can be discarded without much loss of information, as illustrated in **Fig. 2a**. A related technique that has been particularly popular in psychology for a long time is factor analysis (FA). At first glance FA is similar to, and therefore often confused with, PCA in the sense that it tries to account for the data by a linear combination of a few uncorrelated factors by exploiting redundancies. However, there is a crucial and important difference: While PCA tries to align its dimensions with directions of maximum variance taking the data as they are, FA includes an explicit noise model (i.e., the observed data are assumed to result from a set of uncorrelated latent factors plus noise), and tries to find dimensions such that they capture most of the *correlations* within the set of original variables. In consequence, the solutions found by PCA and FA can be very different (see [15] for a nice example). Both PCA and FA have been used to represent multiple single-unit recordings in just a few dimensions (e.g., [10, 15]).

A quite different approach to dimensionality reduction is a set of related techniques called multi-dimensional scaling (MDS). These techniques aim to project the data into a much lower-dimension-

al space while preserving the *distances* or *dissimilarities* between the original data points (or just their ordinal relationships in the case of non-metric MDS) to the highest degree possible (**Fig. 2b**). This should work well if the data mostly lie in an approximately linear subspace or—in the case of extensions like Isomap (see below)—are confined to some lower-dimensional manifold. With regards to the visualization of neural activity flows, this set of techniques has a number of distinct advantages as compared to PCA: Distances between points in PCA space can be highly distorted as the data are projected onto the few most-variance-capturing dimensions, ignoring vector components perpendicular to these directions. Hence with PCA it is not so clear how geometric structures in the original space relate to a reduced PC space, while MDS preserves much more of the original geometry of the data cloud by matching original distances as closely as possible. Moreover, PCA may give rise to somewhat 'awkward' representations if the data variance spreads very unevenly among axes, while with MDS each new dimension is treated 'equal' in a sense. This comes at the price that MDS is an iterative optimization technique with potentially high memory and computation time requirements, and where only suboptimal ('locally optimal') solutions may be found, unlike PCA where the new set of axes can be computed fast and explicitly by solving a simple eigenvalue problem (note, however, that there is a variant of metric MDS, called 'classical MDS', which gives solutions equivalent to PCA if distances are Euclidean). As indicated above, there are also 'non-metric' versions of MDS which aim to preserve only the rank order of the original dissimilarities in the lower-dimensional space. Recently, an MDS-based approach called 'Isomap' [14] was developed to recover the putative lower-dimensional (nonlinear) manifold on which the data are lying by defining distances among points as those on the manifold (geodesic/shortest-path distances). Locally-linear embedding (LLE) [12] is another recently proposed dimensionality reduction technique following a similar objective.

Like PCA and FA, MDS [9], Isomap [3], and LLE [2] have all been applied to visualize features of neural activity. **Fig. 3** gives an example from our own work where simultaneous recordings from 10–40 neurons (which define what we refer to as the multiple single-unit activity, MSUA, space) have been embedded in a three-dimensional space by metric MDS, such that each point in this space represents the simultaneous activity of a population (the population state) of a set of recorded neurons. These recordings were obtained while rats were situated in the working memory and decision-making task illustrated in **Fig. 3a**: During a training phase rats had access to four out of eight baited arms in a radial maze, while the remaining four randomly chosen arms were blocked by physical barriers. After retrieving food from all four open arms, the rat was then confined to the last arm visited for 1 min or longer (the delay phase), after which all barriers were removed such that the rat could now access the four arms still baited in the test phase. Hence, to efficiently solve this task, the animal has to maintain during the delay trial-unique information about the arms already visited or still to visit, and has to utilize this information in the test phase at each arm entry to make a choice about whether to enter the arm or not. Optimal performance on this task is highly dependent on various subdivisions of the prefrontal cortex. In terms of cognitive demands, this task may be divided into various periods associated with 'choice points' (**Fig. 3a**) at which the animals decide whether to enter an arm or not, reward phases where the animals retrieve a food pellet, baseline training and test phases where the animals move from one arm to another, and the delay period. Since memory requirements are different during training and test choices and rewards, one may furthermore speculate that the neural dynamics evolve differently for choices and rewards during training and test. To illustrate the relationship of neural population activity as represented in MDS space to these different cognitively defined task phases, points in MDS space were color-coded according to the task period where they stem from (**Fig. 3b**). As shown in **Fig. 3b**, the different task phases defined above seem to segregate in MDS space, i.e., different task phases are associated with dif-

Abstract

e-Neuroforum 2010 · 1:89–98
DOI 10.1007/s13295-010-0011-0

D. Durstewitz · E. Balaguer-Ballester

Statistical approaches for reconstructing neuro-cognitive dynamics from high-dimensional neural recordings

Abstract
Recent advances in multiple single-unit recording and optical imaging techniques now routinely enable observation of the activity from tens to hundreds of neurons simultaneously. The result is high-dimensional multivariate time series which offer an unprecedented range of possibilities for gaining insight into the detailed spatio-temporal neural dynamics underlying cognition. For instance, they may pave the way for reliable single-trial analyses, for investigating the role of higher-order correlations in neural coding, the mechanisms of neural ensemble formation, or more generally of transitions among attractor states accompanying cognitive processes. At the same time, exploiting the information in these multivariate time series may require more sophisticated statistical methods beyond the commonly employed repertoire. Here we review, using specific experimental examples, some of these methods for visualizing structure in high-dimensional data sets, for statistical inference about the apparent structure, for single-trial analysis of neural time series, and for reconstructing some of the dynamical properties of neural systems that can only be inferred from simultaneous recordings.

Keywords
Neural dynamics · Statistics · Machine learning · Prefrontal cortex · Multiple single-unit recordings

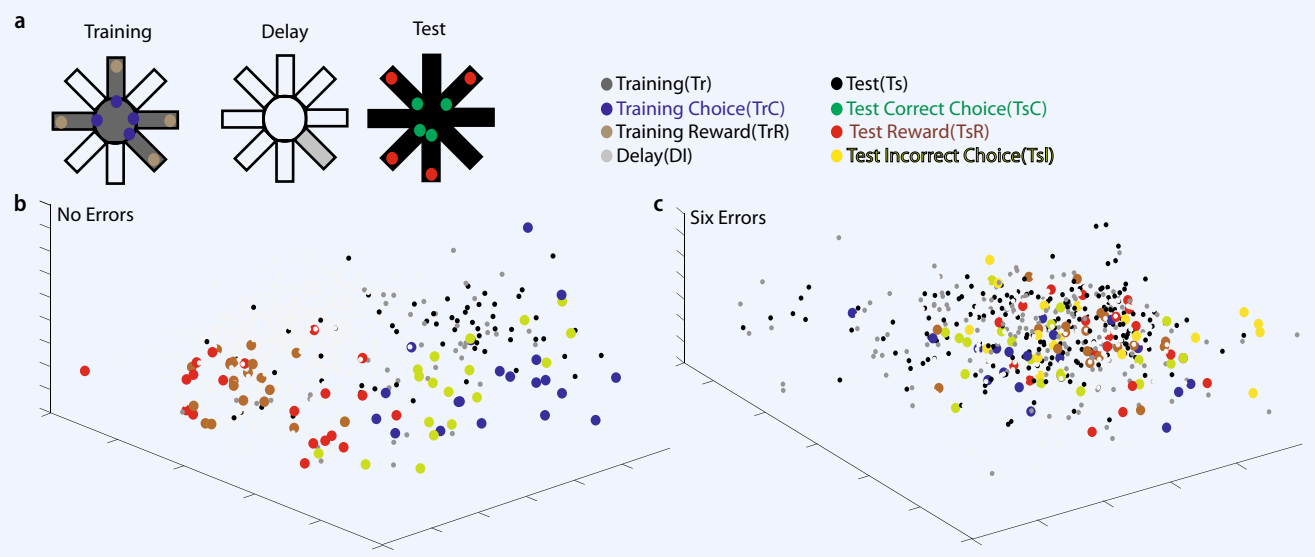


Fig. 3 ▲ Population encoding of different cognitively defined task phases in a working memory and decision making task. **a** Illustration of the delayed win-shift radial arm maze where different task epochs (going in hand with different cognitive processing demands) are color-coded according to the legend. **b,c** Multiple single-unit activity space representation achieved by means of multidimensional scaling (see **Fig. 2b**) for an animal performing with high behavioral accuracy (**b**) and when the same animal committed multiple errors (wrong arm entries) (**c**). Each *color-coded dot* represents the state of the recorded network (i.e., the population vector of instantaneous firing rates) in a 200-ms bin at some point during the task. Points were color-coded according to the task epochs defined in **a**. Population activity distinguished between task epochs when the animal performed well (**b**) but not during bad performance (**c**). (Reproduced with kind permission from [9], copyright 2008 by the National Academy of Sciences of the USA)

ferent patterns of population activity and hence formed different separated clusters in MDS space. This apparent organization of population activity in MDS space, however, breaks down as the animals start to commit a lot of errors defined by revisits to previously exploited arms (**Fig. 3c**). Thus, the segregation of population activity according to task phases in MDS space appears to be functionally relevant to the performance of the task.

Fig. 2c exemplifies another technique for the low-dimensional representation of population activity from MSU recordings which can be used if the focus is on visualization of the *differences* between clouds of points associated with well-defined task events. Fisher's discriminant analysis (FDA) is similar in spirit to PCA, except that it rotates axes such that differences between (the means of) sets of points belonging to different classes are maximized, while at the same time intraclass variances are minimized (**Fig. 2c**, see also **Fig. 5b**). In other words, FDA like PCA is a linear transformation of the original data space, but it tries to find directions in the original space along which two or more a-priori-defined classes are optimally separated instead of maximizing the variance along these directions. In **Fig. 5b** this approach has been used to visualize the differentiation in terms of neural activity between two behavioral rules and two different spatial cues in a rule-switching paradigm explained in further detail below.

Statistical inference about structure in neural state spaces

The techniques summarized above were mainly introduced as tools for data visualization or reduction, although there are also some more genuine data analysis applications. But how do we know whether, for instance, the segregation of population activity in MDS space as shown in **Fig. 3b**, or other features of the visualized population dynamics, are statistically meaningful, i.e., represent significant, beyond chance aspects of population dynamics? Traditional statistical theory offers exact and asymptotic tests, the latter usually founded on the central limit theorem, i.e., the fact that sums of random variables converge to the normal distribution as the number of observations (the sample size) goes to infinity. Many of the traditional statistical tests assume that observations under the null hypothesis are identically (according to the same distribution) and independently distributed. However, in all of the examples with which we started off we are dealing with *time series* generated by biological (or biophysical) systems which by their very nature will exhibit auto-correlations, i.e., observations which were taken close in time tend to be much more similar than observations taken at very different time points, thus violating the independence assumption. Asymptotic statistical tests may still be available provided the time series satisfies certain conditions, but when one has very little knowledge about the true distribution of the test statistic in question, there are also powerful alternatives called the parametric and nonparametric bootstrap. In general, these are relatively easy to use and universally applicable, but also computationally much more intense.

Fig. 4 illustrates how it works for the case of task phase segregation discussed above (see **Fig. 3**). To test whether the observed segregation in the MSUA space is statistically meaningful, we defined a

linear separation error as follows: Given any two task epochs, a hyperplane optimally separating these two epochs was fitted into the original (full-dimensional) MSUA space (**Fig. 4a**), where 'optimal' in this case was defined according to Fisher's linear discriminant criterion introduced above. Hence, the separating hyperplane lies perpendicular to the direction along which the data points are maximally separated and intersects it at a point determined from normal distribution assumptions (**Fig. 4a**). There are other ways to define 'optimality' here, e.g., by fitting the hyperplane such that the margins to the data points closest to it are maximized (leading to maximum margin classifiers, a criterion employed in the so-called support vector machines, SVM). Or of course one may seek nonlinear (e.g., quadratic) separating hyper-surfaces that optimally distinguish different classes of data points. This line of thoughts leads into the broad area of (supervised) classification techniques of which many (like the above mentioned SVM) have recently been employed for predicting a subject's mental state or decision from BOLD signals obtained through fMRI (e.g., [6]). In the present case we are just using the linear classifier to define a test statistic that is sensitive to differences in MSUA space separation (i.e., we are not really interested in 'optimal separation' or prediction), and for this purpose many definitions may do. A linear classifier based on Fisher's discriminant criterion has the particular advantages of being straightforward to compute, simple to understand, and requires no user-dependent parameter settings or tuning.

The specific test statistic was now defined as the relative number of data points classified incorrectly by this approach, i.e., the proportion of data points falling on the wrong side of the separating hyperplane as defined above (**Fig. 4a**). The distribution of this test statistic under the null hypothesis of no significant separation between two task epochs can now be obtained by bootstrapping from the data (i.e., the observed data are taken to define an empirical distribution function from which we resample in certain ways): In the simplest case, if the data were identically and independently distributed, one could just randomly reassign the original data points to the two classes considered (**Fig. 4a**, right; maintaining their relative numbers) and recalculate the test statistic. However, since we are dealing with time series, clustering in MSUA space may also be induced simply by the temporal relations among data points (rather than their cognitive-class membership), and hence the bootstrap data sets should preserve the original auto-correlations to some degree, e.g., up to the length of the stretches of task epoch to be compared. This can be achieved by not just randomly and independently reassigning class labels to individual data points, but by permuting whole blocks of temporally consecutive class labels with block lengths on the order of the duration of the task epochs compared (**Fig. 4b**). There are many variations on this scheme as well as more advanced bootstrapping methods which attempt to retain both the distribution of the original data points as well the power spectrum of the time series (and hence all auto-correlations up to the limits imposed by the finite length of the time series, or, equivalently, the slowest non-zero frequency represented in the power spectrum). In any case, one now generates N (say 1000) replicates (bootstrap data sets), and re-computes the test statistic for each one of them. Following conventional statistical logic, a simple nonparametric test will now denote the original result as being significant at the p=0.05 level if only 5% or less of the bootstrap separation errors rank lower than the one obtained empirically, i.e., for the original data set. In the present example, almost all pairs of cognitive epochs could indeed be significantly separated (after correcting significance levels for multiple testing using the so-called Holm-Bonferroni pro-

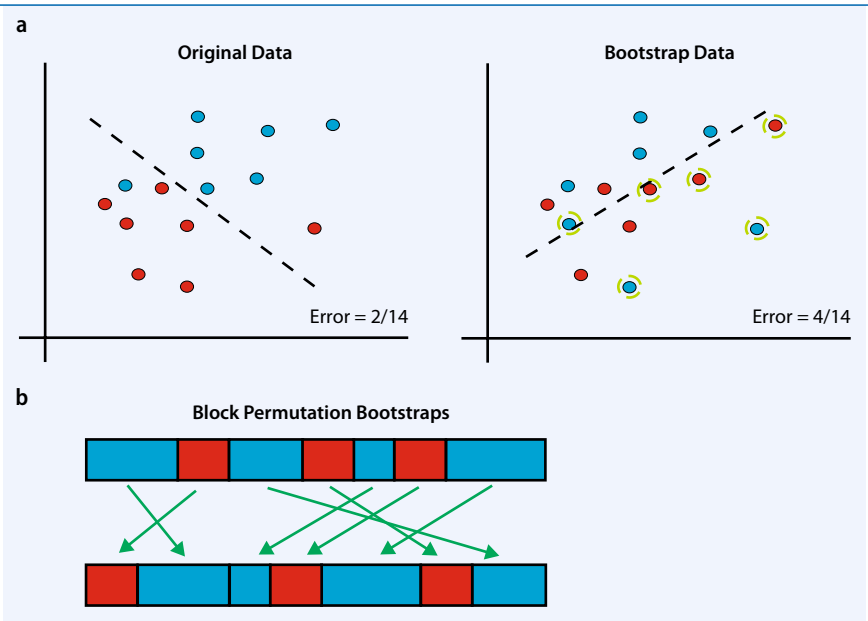
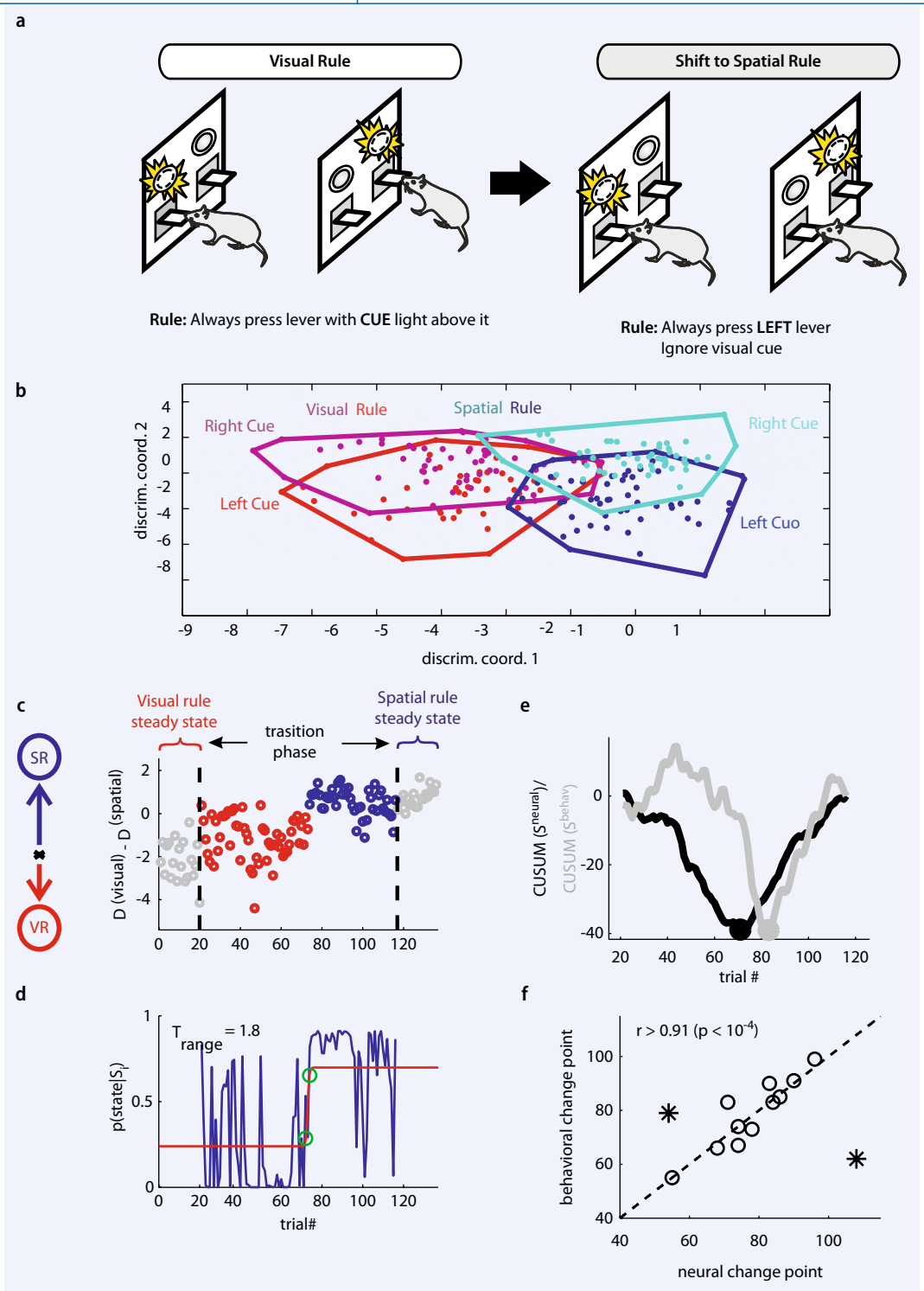


Fig. 4 ▲ Statistics and bootstrap methods for evaluating the significance of the observed task-epoch-related clustering of population activity in **Fig. 3b**. **a** *Left* For any pair of task epochs, a separation error (SE) can be defined as the proportion of points incorrectly assigned by a linear classifier. For this purpose, an optimally separating hyperplane (where 'optimality' can be defined in various ways, e.g., based on normal distribution assumptions) is fitted to the data (see **Fig. 2c**). The relative number of data points falling on the wrong side of this hyperplane is an indication of how well two task epochs can be separated in terms of population activity. *Right* Naively, bootstrap data may now be defined by randomly reassigning data points to the two task epochs (keeping their relative proportions; *green circles* indicate reassigned points). Repeating this procedure N times results in N bootstrap estimates SE^* of separation error under the null hypothesis (H0) that any clustering in MUSA space is purely due to chance, thus 'bootstrapping' the H0 distribution from the data. **b** However, complete random assignment does not account for the fact that almost any time series from a physical system will bear some autocorrelations. More strictly, one would like to test against the H0 that the observed clustering in MSUA space was not purely induced by the fact that consecutive values in a time series tend to be correlated and hence close in MSUA space. Thus, bootstrap data need to be devised that preserve autocorrelations up to the length of a task epoch, and this may be achieved by scrambling whole blocks of temporally consecutive class labels that correspond to different epochs from the same task phase



cedure), and furthermore there were significant differences in task epoch segregation between trials on which either only few or many behavioral errors were committed [9].

Single trial analysis of neural dynamics

One big advantage of having simultaneous recordings from many neurons is that these may allow inferences from the data that are simply not possible with just a few recorded units, examples being the contribution of higher order correlations among many units to neural coding [11], or the temporal dynamics of neural ensembles such as transitions among different ensemble states [10]. Another real virtue is that single trial analyses become feasible since the statistical power and noise reduction usually gained by averaging across many trials can at least partly be offset by somehow combining the measurements

from the many simultaneously recorded neurons (the discriminative power of the recorded network as a whole, for instance, will usually be much larger than that of the best single unit; e.g., [4]). This makes it possible to address a whole set of new or long-standing questions which require single trial resolution, e.g., it may make it possible to dig deeper into the cognitive or behavioral basis of neural trial-to-trial variability (e.g., [15]).

In a recent study, we exploited this advantage to track the neural dynamics in rat prefrontal cortex accompanying the acquisition of a new behavioral rule on a trial-by-trial basis [4]. More specifically, the animals were first trained on a simple discrimination rule where they had to press a lever associated with a cue light ('visual rule'). At some point suddenly and unknown to the animal (i.e., for the first time in the animal's experimental life) reward contingencies were then changed to a new rule which demanded responses always to a fixed side, i.e., always pressing the left or the right lever irrespective of the location of the cue light (**Fig. 5a**). This type of rule switch was shown previously to depend on the integrity of the rat prefrontal cortex. Furthermore, these rules were selected such that a reasonable number of trials on the first rule plus acquisition of the second rule could all be accomplished on one day, ensuring that the same population of neurons was monitored throughout the rule switch. As exemplified in **Fig. 5b**, the two rules once acquired filled clearly distinct regions of MSUA space. The aspect of key interest in this study was, however, how the recorded network transitioned from one rule representation to the other after the rule switch. In particular, we wondered whether there would be a gradual transition stretching out along the whole acquisition phase, or whether the transition would instead be sudden. Sudden transitions of behavioral performance during learning in many different classical and operant conditioning paradigms had been suggested recently by detailed statistical analysis of behavioral learning curves in Randy Gallistel's lab at Rutgers University, New Brunswick. This stands in contrast to the popular view that most forms of animal learning are a relatively slow process in which links between stimuli, responses, and environmental feedback are incrementally strengthened as they become synaptically imprinted.

In accordance with the findings of the Gallistel group, we observed the presence of sudden change points in behavioral performance, i.e., the transition from chance to good performance was not smooth, but often appeared to happen within just a few trials. To quantify the type of transition at the neural level, we computed for each trial the distance of the neural trajectory to the two rule steady states shown in **Fig. 5b**. The idea is that during a gradual learning process the neural state should slowly move from one neural rule representation to the other, i.e., the distance to the visual rule state should gradually increase trial by trial while the one to the spatial rule state would gradually become smaller. Distance in neural space was defined here in terms of the 'Mahalanobis distance' which one may think of as an Euclidean distance between group means normalized by the data scatter within the groups [in fact, the Mahalanobis distance is the Euclidean distance after all variables have been standardized (to have unit variance) and decorrelated]. Thus, in contrast to the Euclidean distance, the Mahalanobis distance takes the spatial spread of data points belonging to different classes into account, thereby, in a sense, incorporating the statistical uncertainty about the precise location of the neural state (on top, rule steady states may be geometrically extended objects for purely systems-dynamical reasons, in addition to the noise and empirical uncertainty about their location). The directions that contribute most to the Mahalanobis distance are those where there is a large difference in means while the within-group variation along that direction is small. This is reminiscent of Fisher's discriminant criterion introduced further above, and indeed assigning data points to classes according to Mahalanobis distances (with pooled covariance matrix) is formally equivalent to linear discriminant analysis.

Fig. 5c plots the trial-time-series of the difference S between the Mahalanobis distances D to the two rule steady states (i.e., $S_i=D_i^{visual}-D_i^{spatial}$). As can be appreci-

Fig. 5 ◀ Single-trial analysis of transitions among neural rule-representations during learning. **a** The task design: Initially rats were given 20 trials of a previously-learned visual cue discrimination task, in which a press of the lever indicated by the cue light was rewarded. Subsequently, without any cue, the rule for reward was changed to a spatial rule in which only the left or only the right lever was rewarded, regardless of the cue light location. **b** Population activity differentiation between the two different behavioral rules (*reddish* and *bluish* colors) and two cue lights (*lighter* and *darker* colors) in a rule-shift task. In this case a two-dimensional representation that highlights differences between the groups corresponding to different cues and rules was obtained by FDA (see **Fig. 2c**). The group-enclosing lines are the convex hulls of each of the four sets of data points, i.e., mark the largest extent of data spread for each group. **c** Difference between the Mahalanobis distances of the current neural population state to the two rule steady states as a function of trial number (the visual and spatial rule steady states are indicated by *dashed vertical lines*). A hidden Markov model (HMM) identified two discrete neural states as indicated by the color coding, with a rather sharp transition around trial 74 between these two states. **d** Conditional probability of the final (spatial) rule state (as defined by the HMM) as a function of trial number. To measure the steepness of the transition, a logistic (sigmoid) function was fitted to this probability and the x-axis range (i.e., number of trials) corresponding to the 10%–90% y-axis coverage of this function was taken as a test statistic (T_{range}). **e** Another way to represent this time series is to cumulate the differences to the mean (see text) and plot as a function of trial number (*black line*). The cumulation reduces the variability and furthermore change points (*black dot*) are easy to identify from this graph. The *gray line* shows the same for the behavioral performance curve (drawn to same scale), with the *gray dot* corresponding to the behavioral change point. **f** Across 11 out of the 13 data sets examined in this study there was a remarkably good temporal agreement between change points identified from the neural dynamics and those from the behavioral performance, with only two exceptions (marked by *asterisks*) where the neural time series showed only little variation (thus making CP location somewhat arbitrary) and/or high behavioral error scores. (Reproduced with kind permission from [4], copyright 2010 by Cell Press, Elsevier)

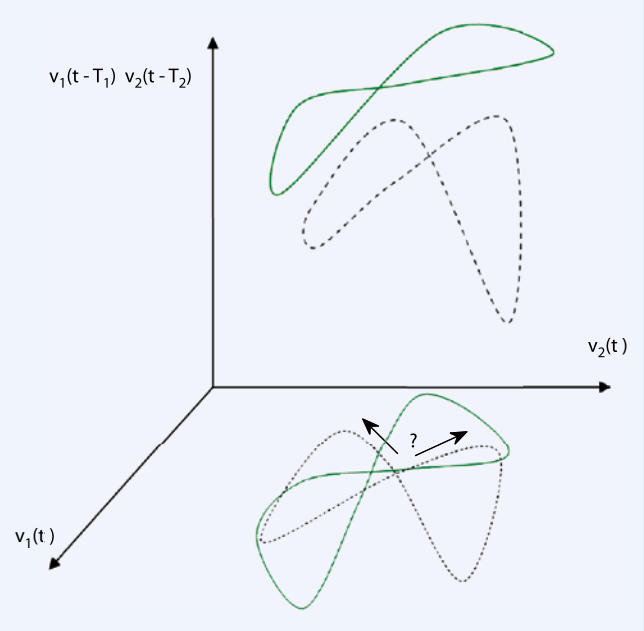


Fig. 6 ▲ Unfolding trajectories of neural ensemble activity via expansion of the MSUA space. The schema shows at the *bottom* two neural trajectories projected into a plane spanned by the instantaneous firing rates of two simultaneously recorded neurons, $v_1(t)$, $v_2(t)$. Since these trajectories may come from a true space of much higher dimensionality, but are forced into the empirically accessed space of just two dimensions, they frequently intersect with themselves and with each other. At each of these intersection points, the flow of the system (the change of activity in time) is not uniquely defined (as indicated by the *arrows* and *question mark*). Therefore, it is not possible to assess for instance a convergent flow of activity as indicative of an attractor state. However, by adding a third axis representing the product of time-lagged versions of the two original variables, $v_1(t-\tau_1)\, v_2(t-\tau_2)$, it is possible to fully disentangle the two trajectories. In this manner, dimensions missing from the original space may be substituted by new axes formed from the measured variables

ated from **Fig. 5c**, there is not a gradual rise of this measure extending across the whole acquisition phase, but a rather steep increase within just a few trials around trial 74. There are different ways to put this statement into more formal and statistical terms. One is based on the idea that there is indeed a (hidden) sequence of underlying neural states generating the observed time series, i.e., at each trial i the neural system is assumed to be in some state K, and each state K could produce the observed values S with some (conditional) probability distribution $p(S|K)$. If the current state K_i is assumed to depend only on the immediately preceding state K_{i-1}, this is called a (first-order) hidden Markov model (HMM). Hence, an HMM is specified by the matrix of transition probabilities among all states K and the conditional probabilities $p(S|K)$. It is usually fitted to the observed data by the Viterbi algorithm, and it has been used by several authors to identify state transitions in multiple single-unit recordings (e.g., [7]). The color-coding in **Fig. 5c** was done according to the states identified this way, and hence changes in color identify change points in the neural dynamics as one moves along the sequence of trials. It turns out that the conditional probability of the final (spatial rule) state $p(K_i = \text{'spatial rule'}|S_i)$ given the observed distances to the two rule steady states (**Fig. 5d**) increases rapidly within just a few trials for at least half of the 13 animals examined. Statistically this is confirmed by the fact that placing the putative change (transition) point randomly along the time series yields significantly shallower slopes for the final-state probability. Thus, the change points identified through the HMM analysis are indeed unique and inconsistent with random fluctuations elsewhere along the time series.

This conclusion is further corroborated by a statistical procedure called change point (CP) analysis: In CP analysis a test statistic is often based on summed (cumulated) differences to the mean, i.e., the quantity

$$CUSUM(S_i) = \sum_{j \le i} (S_j - \langle S \rangle)$$

is plotted as a function of trial number as in **Fig. 5e**. In this representation, change points appear as minima (or maxima) of the curve: For instance, as long as S_i stays below its mean, the curve will continue to decrease, but will steadily rise as long as $S_i > \langle S_i \rangle$, as shown. After detrending the original time series (removing slow consistent drift), this measure should fluctuate around zero if there is no transition in the neural dynamic, while a clear minimum should be present in case S_i hops from a lower to a higher level. Hence, one can take the strongest deflection of the curve from zero,

$$T_{CP} = \max_i \left| \sum_{j \le i} (S_j - \langle S \rangle)/M \right|$$

as an indicator for the presence of change points. To evaluate the significance of the observed T_{CP} and to provide confidence limits for the location of the change point (i.e., its temporal precision), phase-randomized bootstraps can be used: These are bootstrap time series which retain both the distribution of the original values as well as—and importantly—the autocorrelations within the original time series (and thus its power spectrum). Thus these bootstraps will contain all the same linear features and fluctuations as the original time series, but should lack the consistent change in mean that would define the sudden switch in the neural dynamic. In fact, for about half of the cases the original T_{CP} was significantly larger than those for the bootstrap time series, and moreover the CP could be pinpointed to a narrow range of just about six trials with 95% confidence. Finally, we conclude by pointing out that the neural change points were in tight temporal agreement with change

points identified in the behavioral performance (**Fig. 5f**). Hence, these trial-by-trial statistical analyses suggest that learning in this rule-switching task happens rather suddenly, almost as if the animals had a moment of sudden insight, both at the neural and behavioral level.

In summary, multiple single-unit recordings in conjunction with appropriate statistical tools can be exploited to tackle phenomena at the single-trial level that otherwise may have been buried in noise.

Outlook: reconstructing neural trajectories and the flow of activity

Fig. 3b illustrated how neural population activity vectors cluster in MSUA space during a cognitive task. However, in principle, of course there is much more information about the system dynamics one could theoretically extract from these observations of ensemble activity. First, from a dynamical systems point of view it would also be very important to know how activity *moves* between different states, that is, one may want to illustrate and examine the precise neural trajectories connecting different activity patterns. Mazor and Laurent [10] for instance, using such an approach, were able to demonstrate that much more information about a sensory (olfactory) stimulus is encoded in the *transient* neural dynamic than in the stimulus-dependent steady states the system eventually approaches. More specifically, trajectories in the neural state space associated with different stimuli were much more separated (further apart) from each other during the transient response phase than the stimulus-driven fixed points were towards which activity converged after 2–3 s of stimulus presentation.

Second, a frequent assumption in theoretical neuroscience is that computational operations in the nervous system are implemented by moving between attractor states. These are stable configurations or patterns of neural activity towards which the neural dynamic converges in time, at which it tends to persist for a while, and which are resistant to (small) perturbations. For instance, active stimulus or response representations as observed during working memory tasks or spatial representations in the hippocampus have often been claimed to correspond to attracting states of the system dynamics, while cognitive processes (such as the recall of a memory sequence) may unfold by traveling between different states. Simultaneous recordings from many neurons in conjunction with advanced statistical methods for reconstructing state spaces from these recordings may make it possible to visualize and extract such states. There is one conceptual problem that needs to be overcome here. Proving convergence in state space, as a defining condition of an attractor state, requires the neural dynamics to be sufficiently 'unfolded' (**Fig. 6**): Even with multiple implanted tetrodes the number of probed neurons is usually vanishingly small (at least in cortical areas) compared to the total number of neurons. Thus, neural trajectories get squeezed and projected into a space of much smaller dimensionality, the one experimentally accessible, and as a result may intermingle and mix in such a way that a consistent flow of activity in one or the other direction is no longer discernible (**Fig. 6**). Hence, what may be a converging flow of activity in the true high-dimensional neural space may appear as a highly disordered and distorted set of frequent directional changes in low-dimensional projections, preventing us from detecting attractors. A potential solution to this problem is expanding the original MSUA space to much higher dimensionality by forming new variables from the recorded quantities (neural activity values) and adding them as further dimensions to the original MSUA space (**Fig. 6**). In a recent approach ([1], in press), we extended the MSUA space by including time-lagged versions $\nu_i(t\text{-}\tau)$ of the original variables $\nu_i(t)$, as well as higher-order interaction terms such as, e.g., $\nu_i(t\text{-}\tau_i)\times\nu_j(t\text{-}\tau_j)^2$ in its construction. In this manner one may be able to 'unfold' the flow (the directions of neural activity changes) until cognitively relevant attractor states of the full system are completely resolved (**Fig. 6**). A caveat with these methods is that the reconstructed spaces may become so extremely high-dimensional (e.g., several 1000 dimensions) that special algorithms (so-called kernel methods, e.g., [5]) are needed to perform mathematical operations in these spaces. However, such approaches are firmly established in the areas of nonlinear dynamics (so-called embedding theorems; [13]) and statistical learning theory, and they may ultimately offer a much more detailed view of the neural dynamics and hence the neural implementation of computational processes related to cognition.

Corresponding address

Dr. D. Durstewitz
Bernstein Center for Computational Neuroscience Heidelberg-Mannheim, Central Institute of Mental Health & Heidelberg University
J 5, 68159 Mannheim
Germany
daniel.durstewitz@zi-mannheim.de

D. Durstewitz 1989–1994: Studied psychology with focus statistics and mathematics as minor at the Technical University of Berlin. 1994–1998: Doctoral student within the graduate program 'Cognition, brain, & neural networks' at the Ruhr University, Bochum. 1998–2000: Postdoctoral fellow (research associate) at the Salk Institute for Biological Studies (Computational Neurobiology Lab), La Jolla, USA. 2001–2005: Junior group leader within the Emmy Noether program of the DFG (Computational Neuroscience Lab at the Ruhr University, Bochum, Institute for Cognitive Neurosciences). 2005–2008: Reader for Computational Neuroscience at the Centre for Theoretical & Computational Neuroscience, University of Plymouth, UK. Since Oct. 2008: Heisenberg Fellow and group leader at the Central Institute of Mental Health Mannheim, coordinator of the Bernstein Center for Computational Neuroscience Heidelberg-Mannheim.

Dr. E. Balaguer-Ballester
Bernstein Center for Computational Neuroscience Heidelberg-Mannheim, Central Institute of Mental Health & Heidelberg University
J 5, 68159 Mannheim
Germany

E. Balaguer-Ballester 1992-1997: Studied physics (Theoretical Physics) at the University of Valencia, Spain. 1998-2001: Doctoral student at University of Valencia in the departments of Electrical Engineering (Physics) and Vegetal Biology (Faculty of Biology); worked on analysis and prediction of atmospheric pollutants time series. 2001-2005: Research scientist in Artificial Intelligence & Machine Learning Department at TISSAT (a software company based in Valencia, Spain); projects in data mining. 2005-2008: Postdoctoral fellow at the Centre for Theoretical & Computational Neuroscience, University of Plymouth, UK. Since Dec. 2008: Senior postdoctoral fellow at the Central Institute of Mental Health Mannheim and the Bernstein Center for Computational Neuroscience Heidelberg-Mannheim.

Acknowledgments. This work was supported by funding from the Deutsche Forschungsgemeinschaft to DD (Du 354/5-1 & 6-1) and the Bundesministeri-

um für Bildung und Forschung (BMBF 01GQ1003B) via the Bernstein-Center for Computational Neuroscience initiative. We would also like to thank our co-authors Chris Lapish, Nicole Vittoz, Stan Floresco and Jeremy Seamans, who performed the experimental studies underlying the analysis methods discussed here.

References

1. Balaguer-Ballester E, Lapish C, Seamans JK, Durstewitz D (in press) Attracting dynamics of prefrontal cortex ensembles during memory-guided decision-making.
2. Broome BM, Jayaraman V, Laurent G (2006) Encoding and decoding of overlapping odor sequences. Neuron 51:467–482
3. Compte A, Constantinidis C, Tegnér J et al (2003) Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. J Neurophysiol 90:3441–3454
4. Durstewitz D, Vittoz NM, Floresco SB, Seamans JK (2010) Abrupt transitions between prefrontal neural ensemble states accompany behavioral transitions during rule learning. Neuron 66:438–448
5. Hastie T, Tibshirani R, Friedman J (2009) Elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, New York
6. Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. Nat Rev Neurosci 7:523–534
7. Jones LM, Fontanini A, Sadacca BF et al (2007) Natural stimuli evoke dynamic sequences of states in sensory cortical ensembles. Proc Natl Acad Sci U S A 104:18772–18777
8. Krzanowski WJ (2000) Principles of multivariate analysis: a user's perspective. Oxford University Press, Oxford
9. Lapish CL, Durstewitz D, Chandler LJ, Seamans JK (2008) Successful choice behavior is associated with distinct and coherent network states in anterior cingulate cortex. Proc Natl Acad Sci U S A 105:12010–12015
10. Mazor O, Laurent G (2005) Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons. Neuron 48:661–673
11. Ohiorhenuan E, Mechler F, Purpura KP et al (2010) Sparse coding and high-order correlations in fine-scale cortical networks. Nature 466:617–621
12. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326
13. Sauer T, Yorke J, Casdagli M (1991) Embedology. J Stat Phys 65:579–616
14. Tenenbaum JB, Silva V de, Langford JC (2000) Global geometric framework for nonlinear dimensionality reduction. Science 290:2319–2323
15. Yu BM, Cunningham JP, Santhanam G et al (2009) Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. J Neurophysiol 102:614–635